A Supplementary Material

The supplementary materials consist of:

1. Code of AE2.
2. Supplementary video with qualitative examples of AE2.
3. Experimental setup: This includes detailed descriptions of the datasets, the evaluation protocol, and complete implementation details.
4. Further results and visualizations: Due to space constraints in the main paper, we provide a more detailed breakdown of the results reported in the main paper, along with more qualitative examples.

A.1 Supplementary Video

In our supplementary video, we show qualitative examples of one practical application enabled by our learned ego-exo view-invariant representations—synchronous playback of egocentric and exocentric videos. We randomly select ego-exo video pairs from the test set, and use the frozen encoder ϕ to extract frame-wise embeddings. We then match each frame of one video (the reference) to its closest counterpart in the other video using nearest neighbor. As demonstrated across all datasets with several examples, AE2 effectively aligns two videos depicting the same action, overcoming substantial viewpoint and background differences. The supplementary video also includes examples of synchronizing two egocentric or two exocentric videos and explores AE2's failure cases.

A.2 Experimental Setup

A.2.1 Datasets

Since existing video datasets for fine-grained video understanding (*e.g.*, PennAction [82], FineGym [65], IKEA ASM [3]) are solely composed of third-person videos, we curate video clips from five public datasets: CMU-MMAC [13], H2O [32], EPIC-Kitchens [11], HMDB51 [31] and Penn Action [82] and collect a egocentric Tennis Forehand dataset. Our data selection criteria is to find videos that depict the same action from distinct egocentric and exocentric viewpoints. Consequently, TCN pouring [63] is excluded due to its small scale and significant similarity between the egocentric and exocentric views; Assembly101 [62] is not selected since there there are large variations within a single atomic action and the egocentric video is monochromatic.

In all, our dataset compilation from five public data sources, along with our collected egocentric tennis videos, results in four distinct ego-exo datasets, each describing specific atomic actions:

(A) **CMU-MMAC** [13]. The dataset contains 44 subjects cooking five different recipes (brownies, pizza, sandwich, salad, and scrambled eggs), captured from one egocentric and four exocentric views simultaneously. We use the temporal keystep boundaries provided in [2] to extract clips corresponding to the action of breaking eggs from all videos. Among the four exocentric viewpoints, we adopt videos from the right-handed view, as this viewpoint captures the action being performed most clearly. We randomly split the data into training and validation sets across subjects, with 35 subjects (118 videos) for training and 9 subjects (30 videos) for validation and test. There is strict synchronization between egocentric and exocentric video pairs in this dataset.

Table 3: Dataset summary.

Dataset	# Train		# Val		# Test		Fixed	Ego-exo
	ego	exo	ego	exo	ego	exo	exo view?	time-sync?
(A) CMU-MMAC	61	57	5	5	10	10	✓	✓
(B) H2O	29	48	4	8	7	16	✓	✗
(C) Pour Liquid	70	67	10	9	19	18	✗	✗
(D) Tennis Forehand	94	79	25	24	50	50	✗	✗

(B) **H2O** [32]. The dataset features 10 subjects interacting with a milk carton using both hands, captured by one egocentric camera and four static exocentric cameras. We utilize the keystep annotations provided in [33] to extract clips corresponding to the pouring milk action. All four exocentric views are included, as they clearly capture the action. We follow the data split in [33], with 7 subjects (77 videos) in the training set and 3 subjects (35 videos) in the validation and test set. A portion of this dataset (the first 3 subjects) contains synchronized egocentric-exocentric video pairs, while the remaining part does not.

(C) **Pour Liquid**. To evaluate our methods on in-the-wild data, we assemble a Pour Liquid dataset by extracting clips from one egocentric dataset, EPIC-Kitchens [11] and one exocentric dataset, HMDB51 [31]. We utilize clips from the "pour water" class in EPIC-Kitchens and "pour" category in HMDB51. Following the data split in the original datasets, we obtain 137 videos for training and 56 videos for validation and test. It is important to note that the egocentric and exocentric videos are neither synchronized nor collected in the same environment, providing a challenging testbed.

(D) **Tennis Forehand**. To include physical activities in our study, we leverage exocentric video sequences of the tennis forehand action from Penn Action [82] and collect an egocentric dataset featuring the same action performed by 12 subjects using Go Pro HERO8. We adopt the data split from [15] for Penn Action, and divide our egocentric tennis forehand dataset by subject: 8 for training and 4 for validation and testing. This results in a total of 173 clips for training, and 149 clips for validation and testing. The egocentric and exocentric videos, gathered from a range of real-world scenarios, are naturally unpaired.

Table 3 provides a summary of these four datasets. In addition, we recognize that the original datasets lack frame-wise labels and provide dense frame-level annotations to enable a comprehensive evaluation of the learned representations. See Table 4 for a complete list of all the key events we annotate and Fig. 5 for illustrative examples.

Table 4: Number of actions phases and list of key events for each dataset.

Dataset	# phases	List of key events
(A) CMU-MMAC	4	hit egg, visible crack on the eggshell; egg contents released into bowl
(B) H2O	3	liquid starts exiting, pouring complete
(C) Pour Liquid	3	liquid starts exiting, pouring complete
(D) Tennis Forehand	2	racket touches ball

A.2.2 Evaluation

We provide a detailed description of the four downstream tasks below and their corresponding evaluation metrics:

1. **Action Phase Classification**. We train an SVM classifier on top of the embeddings to predict the action phase labels for each frame and report F1 score on test data. Besides the regular setting, we investigate (1) few-shot; and (2) cross-view zero-shot settings.
   (1) Few-shot. We assume that only a limited number of training videos have annotations and can be used for training the SVM classifier.
   (2) Cross-view zero-shot. We assume that per-frame labels of training data are only available on one view for training the SVM classifier, and report the test performance on the other view. We use the terms "exo2ego" to describe the case where we use exocentric data for training the SVM classifier and test its performance on egocentric data, while "ego2exo" represents the reverse case.
2. **Phase Progression**. We train a linear regressor on the frozen embeddings to predict the phase progression values, defined as the difference in time-stamps between any given frame and each key event, normalized by the number of frames in that video [15]. Average R-square measure on test data is reported. This metric evaluates how well the progress of an action is captured by the embeddings, with the maximum value being 1.
3. **Frame Retrieval**. We report the mean average precision (mAP)@K (K=5,10 ,15). For each query, average precision is computed by determining how many frames among the retrieved

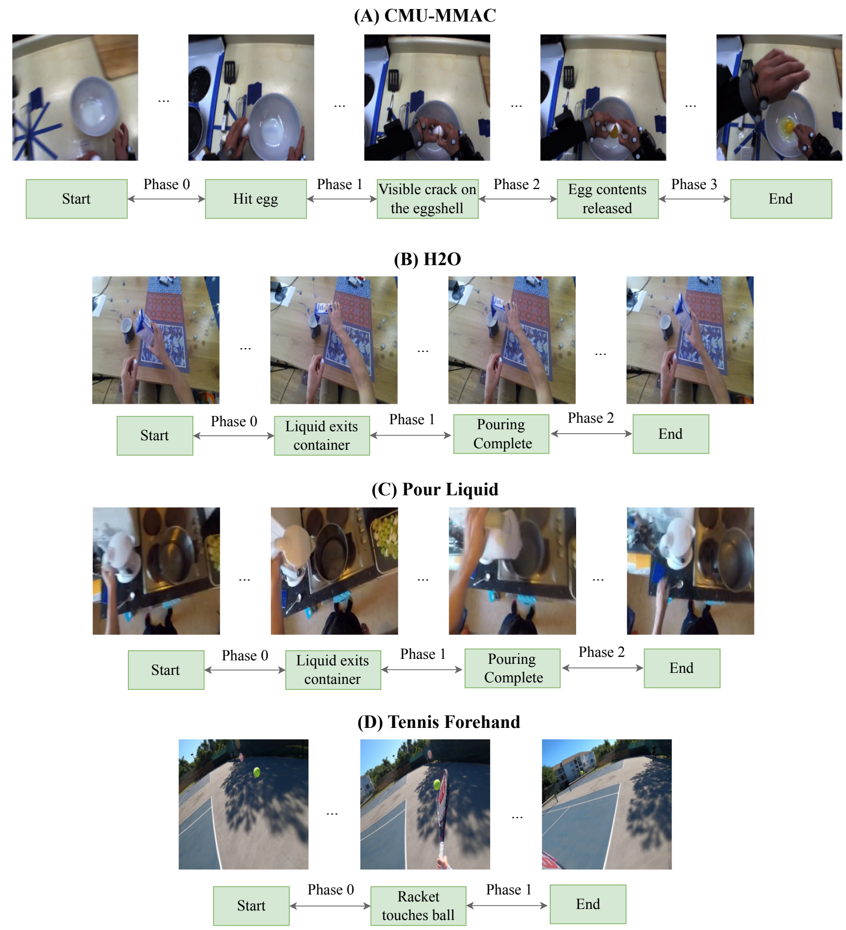

Figure 5: Example labels for all datasets. Key events are displayed in boxes below sequences, with the phase label assigned to each frame between two key events.

K frames have the same action phase labels as the query frame, divided by K. Furthermore, to evaluate view-invariance, we propose the cross-view frame retrieval task (*i.e.*, ego2exo and exo2ego frame retrieval). For each query in one view, the goal is to retrieve K frames from another view. No additional training is required for this task.

4. **Kendall's Tau**. This metric is calculated for every pair of test videos by sampling two frames in the first video and retrieving the corresponding nearest frames in the second video, then checking whether their orders are shuffled. It measures how well-aligned two sequences are in time. No additional training or frame-wise labels are necessary for this evaluation.

It is important to note that (1) due to label imbalance in these datasets, we opt for using the F1 score instead of accuracy for action phase classification to better account for the imbalance and provide a more meaningful performance evaluation. (2) Phase progression assumes a high level of consistency in actions, with noisy frames diminishing the performance greatly. Due to the challenging nature of Pour Liquid data, we observe a negative progression value for all approaches. Thus, we augment the resulting embeddings with a temporal dimension, as 0.001 times the time segment as the input so that the regression model can distinguish repetitive (or very similar) frames that differ in time. We report modified progression value for all baselines and our approach on this dataset. (3) Kendall's Tau assumes that there are no repetitive frames in a video. Since we adopt in-the-wild videos where strict monoticity is not guaranteed, this metric may not faithfully reflect the quality of representations. Nonetheless, we report them for completeness.

Table 5: Hyperparameters summary. 'Lr' stands for learning rate, and 'Wd' denotes weight decay.

Datasets	Optimizer		Transformer Encoder		Regularization		
	Lr	Wd	Hidden Dim.	# Layers	# Frames	# Pos. Frames	Ratio λ
(A) CMU-MMAC	5e-5	1e-5	256	1	32	32	1
(B) H2O	1e-4	1e-5	256	1	32	8	2
(C) Pour Liquid	5e-5	1e-5	128	3	32	16	2
(D) Tennis Forehand	5e-5	1e-5	128	1	20	10	4

Table 6: Results of few-shot action phase classification (F1 score) and frame retrieval (mAP@5,10,15).

Dataset	Method	Few-shot Cls.			Frame Retrieval		
		10%	50%	100%	mAP@5	mAP@10	mAP@15
(A)	Random Features	19.18	19.18	19.18	48.26	47.13	45.75
	ImageNet Features	46.15	48.80	50.24	49.98	50.49	50.08
	ActorObserverNet [66]	31.40	35.63	36.14	50.92	50.47	49.72
	single-view TCN [63]	52.30	54.90	56.90	52.82	53.42	53.60
	multi-view TCN [63]	56.88	59.25	59.91	59.11	58.83	58.44
	multi-view TCN (unpaired) [63]	56.13	56.65	56.79	58.18	57.78	57.21
	CARL [10]	39.18	41.92	43.43	47.14	46.04	44.99
	TCC [15]	57.54	59.18	59.84	59.33	58.75	57.99
	GTA [22]	56.89	56.77	56.86	62.79	61.55	60.38
	AE2 (ours)	**63.95**	**64.86**	**66.23**	**66.86**	**65.85**	**64.73**
(B)	Random Features	36.84	36.84	36.84	52.94	52.48	51.59
	ImageNet Features	39.29	40.83	41.59	53.32	54.09	54.06
	single-view TCN [63]	43.60	46.83	47.39	56.98	57.00	56.46
	CARL [10]	48.73	48.78	48.79	55.59	55.01	54.23
	TCC [15]	78.69	77.97	77.91	81.22	80.97	80.46
	GTA [22]	79.82	80.96	81.11	80.65	80.12	79.68
	AE2 (ours)	**85.17**	**85.12**	**85.17**	**85.25**	**84.90**	**84.55**
(C)	Random Features	45.26	45.26	45.26	49.69	49.83	49.18
	ImageNet Features	55.53	54.43	53.13	50.52	51.49	51.89
	single-view TCN [63]	54.62	55.08	54.02	48.50	48.83	49.03
	CARL [10]	51.68	55.67	56.98	55.03	55.29	54.93
	TCC [15]	52.37	51.70	52.53	62.93	62.33	61.44
	GTA [22]	55.91	56.87	56.92	62.83	62.79	62.12
	AE2 (ours)	**65.88**	**66.53**	**66.56**	**66.55**	**65.54**	**64.66**
(D)	Random Features	31.54	30.31	30.31	69.57	66.47	64.34
	ImageNet Features	65.48	68.03	69.15	78.11	76.96	75.84
	single-view TCN [63]	65.78	69.19	68.87	74.05	73.76	73.10
	CARL [10]	58.89	59.38	59.69	72.94	69.43	67.14
	TCC [15]	67.71	77.07	78.41	82.78	80.24	78.59
	GTA [22]	80.31	83.04	83.63	86.59	85.20	84.33
	AE2 (ours)	**85.24**	**85.72**	**85.87**	**87.94**	**86.83**	**86.05**

A.2.3 Implementation

For all video sequences, frames are resized to 224×224. During training, we randomly extract 32 frames from each video to construct a video sequence. We train the models for a total number of 300 epochs with a batch size of 4, using the Adam optimizer. The model checkpoint demonstrating the best performance on validation data is selected, and its performance on test data is reported. In terms of the encoder network, global features are taken from the output of $Conv4c$ layer in ϕ_{base}. Following [15], we stack the features of any given frame and its context frames along the dimension of time, followed by 3D convolutions for aggregating temporal information and 3D max pooling. In all experiments, the number and stride of context frames are set as 1 and 15, respectively. For local features, we take output of the $Conv1$ layer in ϕ_{base} and apply 3D max pooling to aggregate temporal information from the given frame and its context frame. The features are then fed as input of ROI Align.

Table 7: Results of cross-view frame retrieval (mAP@5,10,15).

Dataset	Method	Ego2exo Frame Retrieval mAP@5	mAP@10	mAP@15	Exo2ego Frame Retrieval mAP@5	mAP@10	mAP@15
(A)	Random Features	42.51	41.74	40.51	38.08	38.19	37.10
	ImageNet Features	33.32	33.09	32.78	38.99	37.80	36.71
	ActorObserverNet [66]	43.57	42.70	41.56	42.00	41.29	40.48
	single-view TCN [63]	31.12	32.63	33.73	34.67	34.91	35.31
	multi-view TCN [63]	46.38	47.04	46.96	52.50	52.68	52.43
	multi-view TCN (unpaired) [63]	55.34	54.64	53.75	58.79	57.87	57.07
	CARL [10]	37.89	37.38	36.57	40.37	39.94	39.38
	TCC [15]	62.11	61.11	60.33	62.39	62.03	61.25
	GTA [22]	57.11	56.25	55.10	54.47	53.93	53.22
	AE2 (ours)	**65.70**	**64.59**	**63.76**	**62.48**	**62.15**	**61.80**
(B)	Random Features	51.46	50.56	48.93	52.78	51.98	50.82
	ImageNet Features	25.72	27.31	28.57	41.50	43.21	43.06
	single-view TCN [63]	47.00	46.48	45.42	47.94	47.20	46.59
	CARL [10]	54.35	52.99	51.99	51.14	51.51	51.00
	TCC [15]	75.54	75.30	75.02	80.44	80.27	80.18
	GTA [22]	72.55	72.78	72.96	75.16	75.40	75.48
	AE2 (ours)	**78.21**	**78.48**	**78.78**	**83.88**	**83.41**	**83.05**
(C)	Random Features	55.78	55.44	54.77	56.31	55.75	54.56
	ImageNet Features	51.44	52.17	52.38	30.18	30.44	30.40
	single-view TCN [63]	53.60	55.28	55.46	29.16	31.15	31.95
	CARL [10]	59.59	59.37	59.19	34.73	36.80	38.10
	TCC [15]	55.98	56.08	56.13	**58.11**	**57.89**	**57.15**
	GTA [22]	57.03	58.52	59.00	51.71	53.32	53.54
	AE2 (ours)	**66.23**	**65.79**	**65.00**	57.42	57.35	57.03
(D)	Random Features	61.24	58.98	56.94	63.42	59.87	57.57
	ImageNet Features	69.34	66.90	64.95	61.61	60.31	58.55
	single-view TCN [63]	54.12	55.08	55.05	56.70	56.65	55.84
	CARL [10]	52.18	54.83	55.39	65.94	63.19	60.83
	TCC [15]	57.87	55.84	53.81	48.62	47.27	46.11
	GTA [22]	78.93	78.00	77.01	79.95	79.14	78.52
	AE2 (ours)	**82.58**	**81.46**	**80.75**	**82.82**	**82.07**	**81.69**

During evaluation, we freeze the encoder ϕ and use it to extract 128-dimensional embeddings for each frame. These representations are then assessed across a variety of downstream tasks (Sec. 4). Detailed hyperparameters specific to each dataset are provided in Table 5. Noteworthy adjustments include: (1) In the case of Tennis Forehand, we utilize a single object proposal, as the active object is only the tennis racket (the tennis ball is too small to be detected reliably). Furthermore, given the shorter video lengths, we sample 20 frames from each video as opposed to the usual 32. (2) For datasets featuring non-monotonic actions (*i.e.*, H2O and Pour Liquid), we construct the negative sequence by randomly reversing either the first or the last half of the sequence, rather than the whole sequence. This is due to the cyclic nature of the pouring action present in some videos within these datasets. All experiments are conducted using PyTorch [52] on 2 Nvidia V100 GPUs.

A.3 Further Results and Visualizations

Results Supplementing Table 1 in the main paper, Tables 6 and 7 present comprehensive results of AE2 and baseline models on few-shot action phase classification and frame retrieval tasks. For few-shot classification, we train the SVM classifier with 10% (or 50%) of the training data, averaging results over 10 runs. AE2 demonstrates superior performance in learning fine-grained, view-invariant ego-exo features when compared with ego-exo [66, 63], frame-wise contrastive learning [10], and alignment-based [15, 22] approaches.

On CMU-MMAC, AE2 greatly outperforms the multi-view TCN [63], which utilizes perfect ego-exo synchronization as a supervision signal. We hypothesize that the strict supervision requirement of TCN might be limiting, as it can not utilize as many ego-exo pairs as AE2 due to its reliance on

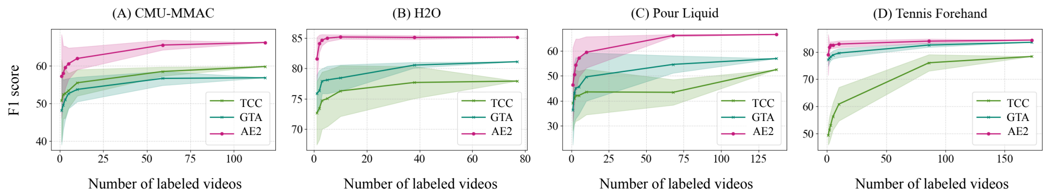

Figure 6: Few-shot action phase classification results. AE2 achieves superior performance across a wide range of labeled training videos, particularly under the most challenging conditions where less than 10 videos are labeled. Results are averaged over 50 runs, and confidence bars represent one standard deviation.

Table 8: Ablation Study of AE2.

Dataset	Method	Classification (F1 score) regular	ego2exo	exo2ego	Frame Retrieval (mAP@10) regular	ego2exo	exo2ego	Phase Progression	Kendall's Tau
(A)	Base DTW	58.53	57.78	54.23	58.36	55.36	58.95	0.1920	0.5641
	+ object	62.86	**60.88**	58.52	62.66	61.26	60.69	0.4235	0.5484
	+ object + contrast	**66.23**	57.41	**71.72**	**65.85**	**64.59**	**62.15**	**0.5109**	**0.6316**
(B)	Base DTW	82.91	81.82	81.83	81.49	74.63	80.21	0.7525	0.8199
	+ object	84.04	84.20	**84.23**	83.03	**81.42**	81.57	**0.7646**	0.8886
	+ object + contrast	**85.17**	**84.73**	82.77	**84.90**	78.48	**83.41**	0.7634	**0.9062**
(C)	Base DTW	59.66	55.48	59.49	52.57	54.12	52.23	0.0553	0.0609
	+ object	63.28	**57.60**	62.42	63.40	**67.15**	**63.05**	**0.2231**	**0.1339**
	+ object + contrast	**66.56**	57.15	**65.60**	**65.54**	65.79	57.35	0.1380	0.0934
(D)	Base DTW	79.56	81.38	72.54	82.65	75.36	76.74	0.4022	0.4312
	+ object	84.14	**85.36**	83.32	**88.22**	79.07	**82.61**	**0.5431**	**0.6477**
	+ object + contrast	**85.87**	84.71	**85.68**	86.83	**81.46**	82.07	0.5060	0.6171

ego-exo synchronization. In contrast, AE2 capitalizes on a broader set of unpaired ego-exo data. Even when we modify multi-view TCN to consider all potential ego-exo pairs as synchronously perfect (termed as multi-view TCN (unpaired) in the tables), it does not outperform its regular version, indicating a lack of robustness towards non-synchronized ego-exo pairs. Consequently, it appears that multi-view TCN is ill-equipped to learn desired view-invariant representations from unpaired, real-world ego-exo videos.

Few-shot Learning Besides the few-shot results in Table 6, we vary the number of labeled training videos, ranging from extremely sparse (a single labeled video) to the case where all training videos are labeled. Note that each labeled video equates to multiple labeled frames. AE2 is compared with top-performing baselines, TCC [15] and GTA [22], across all four datasets in Fig. 6. The results are averaged over 50 runs and include a +- one standard deviation error bar. As shown, AE2 excels in low-label scenarios. For instance, on H2O, a single labeled video yields an action phase classification F1 score over 80%. This suggests that AE2 effectively aligns representations across all training videos, enabling a robust SVM classifier for the downstream task, even with minimal labeling.

Ablation Table 8 presents an ablation of AE2 on all four datasets, which is a comprehensive version of Table 2 in the main paper. From the results, we can see that object-centric representations are instrumental in bridging the ego-exo gap, leading to substantial performance improvements. For instance, frame retrieval mAP@10 improves by +10.83% on Pour Liquid and +5.57% on Tennis Forehand. Furthermore, incorporating contrastive regularization provides additional performance boosts for several downstream tasks such as regular action phase classification. These results demonstrate the integral contributions of both components of AE2 to achieve optimal performance.

Visualizations In addition to the cross-view frame retrieval results for Pour Liquid and Tennis Forehand presented in the main paper (Fig. 4), we showcase results for the other two datasets (*i.e.*, CMU-MMAC and H2O) in Fig. 7. For any given query frame from one view, the retrieved

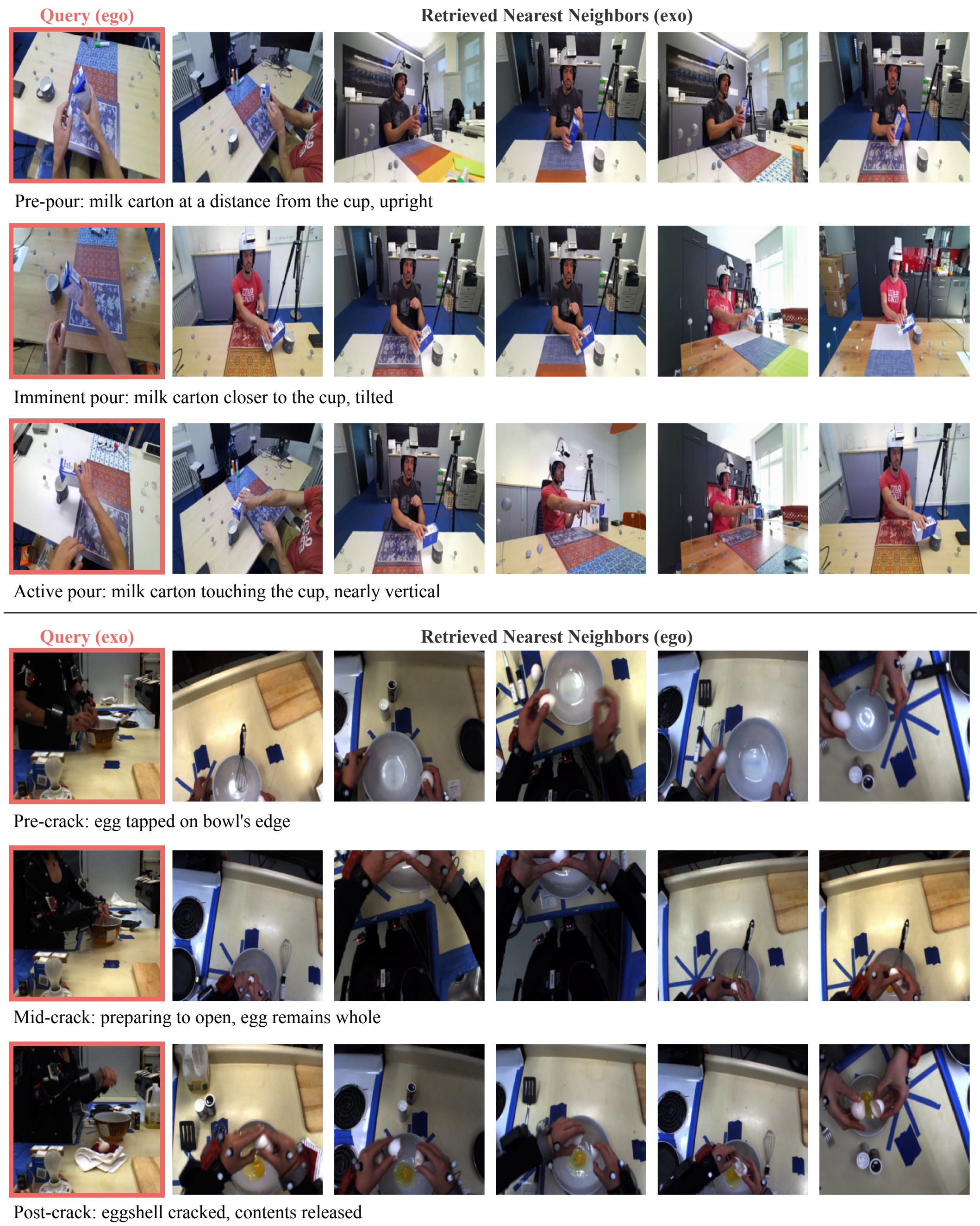

Figure 7: Cross-view frame retrieval results on H2O (rows 1-3) and CMU-MMAC (rows 4-6). AE2 leads to representations that encapsulate the fine-grained state of an action and are invariant to the ego-exo viewpoints.

nearest neighbors closely match the action stage of the query, regardless of substantial differences in viewpoints. These results underline AE2's efficacy in learning fine-grained action representations that transcend ego-exo viewpoint differences.

References

[1] Shervin Ardeshir and Ali Borji. An exocentric look at egocentric actions and vice versa. *Computer Vision and Image Understanding*, 171:61–68, 2018.

[2] Siddhant Bansal, Chetan Arora, and CV Jawahar. My view is the best view: Procedure learning from egocentric videos. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 657–675. Springer, 2022.

[3] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021.

[4] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994.

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[6] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10618–10627, 2020.

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[8] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3546–3555, 2019.

[9] Chao-Yeh Chen and Kristen Grauman. Inferring unseen views of people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2003–2010, 2014.

[10] Minghao Chen, Fangyun Wei, Chong Li, and Deng Cai. Frame-wise action representations for long videos via sequence contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13801–13810, 2022.

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[12] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022.

[13] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[15] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1801–1810, 2019.

[16] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[17] Ali Farhadi and Mostafa Kamali Tabrizi. Learning to recognize activities from the wrong view point. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*, pages 154–166. Springer, 2008.

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[19] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. 2017.

[20] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[21] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[22] Isma Hadji, Konstantinos G Derpanis, and Allan D Jepson. Representation learning via global temporal alignment and cycle-consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11068–11077, 2021.

[23] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5548–5558, 2021.

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[26] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *NeurIPS*, 2016.

[27] Alexandros Iosifidis, Anastasios Tefas, and Ioannis Pitas. View-invariant action recognition based on artificial neural networks. *IEEE transactions on neural networks and learning systems*, 23(3):412–424, 2012.

[28] D. Jayaraman and K. Grauman. Slow and steady feature analysis: Higher order temporal coherence in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] Bruno Korbar, Du Tran, and Lorenzo Torresani. Co-training of audio and video representations from self-supervised temporal synchronization. In *NeurIPS*, 2018.

[30] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

[31] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[32] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021.

[33] Taein Kwon, Bugra Tekin, Siyu Tang, and Marc Pollefeys. Context-aware sequence alignment using 4d skeletal augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8172–8182, 2022.

[34] Mandy Lange, Dietlind Zühlke, Olaf Holz, Thomas Villmann, and Saxonia-Germany Mittweida. Applications of lp-norms and their smooth approximations for gradient based learning vector quantization. In *ESANN*, pages 271–276. Citeseer, 2014.

[35] Jun Li and Sinisa Todorovic. Action shuffle alternating learning for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12628–12636, 2021.

[36] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6943–6953, 2021.

[37] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

[38] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13853–13863, 2022.

[39] Jingen Liu, Mubarak Shah, Benjamin Kuipers, and Silvio Savarese. Cross-view action recognition via view knowledge transfer. In *CVPR 2011*, pages 3209–3216. IEEE, 2011.

[40] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022.

[41] Weizhe Liu, Bugra Tekin, Huseyin Coskun, Vibhav Vineet, Pascal Fua, and Marc Pollefeys. Learning to align sequential actions in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2181–2191, 2022.

[42] YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1118–1125. IEEE, 2018.

[43] Jian Ma and Dima Damen. Hand-object interaction reasoning. In *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–8. IEEE, 2022.

[44] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Towards universal visual reward and representation via value-implicit pre-training. In *ICLR*, 2023.

[45] Behrooz Mahasseni and Sinisa Todorovic. Latent multitask learning for view-invariant action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3128–3135, 2013.

[46] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020.

[47] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[48] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, 2016.

[49] Pedro Morgado, Yi Li, and Nuno Nvasconcelos. Learning representations from audio-visual spatial alignment. In *NeurIPS*, 2020.

[50] Tushar Nagarajan and Kristen Grauman. Shaping embodied agent behavior with activity-context priors from egocentric video. *Advances in Neural Information Processing Systems*, 34:29794–29805, 2021.

[51] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018.

[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[53] AJ Piergiovanni and Michael S Ryoo. Recognizing actions in videos from unseen viewpoints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4124–4132, 2021.

[54] Senthil Purushwalkam, Tian Ye, Saurabh Gupta, and Abhinav Gupta. Aligning videos in space and time. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 262–278. Springer, 2020.

[55] Hossein Rahmani and Ajmal Mian. Learning a non-linear knowledge transfer model for cross-view action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2458–2466, 2015.

[56] Hossein Rahmani and Ajmal Mian. 3d action recognition from novel viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2016.

[57] Cen Rao and Mubarak Shah. View-invariance in action recognition. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001.

[58] Cen Rao, Alper Yilmaz, and Mubarak Shah. View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203, 2002.

[59] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *International Journal of Computer Vision*, 119:346–373, 2016.

[60] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

[61] Madeline C. Schiappa, Yogesh S. Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *arXiv:2207.00419*, 2022.

[62] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022.

[63] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018.

[64] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

[65] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625, 2020.

[66] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7396–7404, 2018.

[67] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

[68] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016.

[69] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[70] Bilge Soran, Ali Farhadi, and Linda Shapiro. Action recognition in the presence of one egocentric and multiple static cameras. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part V 12*, pages 178–193. Springer, 2015.

[71] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 53–70. Springer, 2020.

[72] Reuben Tan, Bryan Plummer, Kate Saenko, Hailin Jin, and Bryan Russell. Look at what i'm doing: Self-supervised spatial grounding of narrations in instructional videos. *Advances in Neural Information Processing Systems*, 34:14476–14487, 2021.

[73] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019.

[74] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[76] Daniel Weinland, Mustafa Özuysal, and Pascal Fua. Making action recognition robust to occlusions and viewpoint changes. In *European Conference on Computer Vision*, 2010.

[77] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding (CVIU)*, 2006.

[78] Xinxiao Wu, Han Wang, Cuiwei Liu, and Yunde Jia. Cross-view action recognition over heterogeneous feature spaces. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–616, 2013.

[79] Yuncong Yang, Jiawei Ma, Shiyuan Huang, Long Chen, Xudong Lin, Guangxing Han, and Shih-Fu Chang. Tempclr: Temporal alignment representation with contrastive learning. *arXiv preprint arXiv:2212.13738*, 2022.

[80] Huangyue Yu, Minjie Cai, Yunfei Liu, and Feng Lu. What i see is what you see: Joint attention learning for first and third person video co-analysis. *ACM MM*, 2019.

[81] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE international conference on computer vision*, pages 2117–2126, 2017.

[82] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 2248–2255, 2013.

[83] Zhong Zhang, Chunheng Wang, Baihua Xiao, Wen Zhou, Shuang Liu, and Cunzhao Shi. Cross-view action recognition via a continuous virtual path. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2690–2697, 2013.

[84] Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. *arXiv preprint arXiv:2303.18230*, 2023.

[85] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. *arXiv preprint arXiv:2210.11339*, 2022.

[86] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019.