Bootstrap Your Own Views: Masked Ego-Exo Modeling for Fine-grained View-invariant Video Representations

Jungin Park[1] Jiyoung Lee[2,3*] Kwanghoon Sohn[1,4*]
[1]Yonsei University [2]Ewha Womans University
[3]NAVER AI Lab [4]Korea Institute of Science and Technology (KIST)
{newrun, khsohn}@yonsei.ac.kr lee.jiyoung@ewha.ac.kr

Appendix

In this document, we provide more concrete details of the AE2 benchmark [15] in Appendix A, additional experimental results in Appendix B, including the results using different backbones (CLIP pretrained ViT-L/14 [5] and ResNet-50 [8]), few-shot classification and frame retrieval performance, ablation studies for each hyper-parameter, and analysis for the failure case. Finally, we present the broader impact of our **BYOV** in Appendix C.

A. Benchmark Details

In this section, we provide a detailed explanation of AE2 benchmark [15] and the evaluation details of four downstream tasks, including action phase classification, frame retrieval, phase progression, and Kendall's τ.

A.1. Datasets

The AE2 benchmark [15] contains four datasets: (1) Break Eggs; (2) Pour Mild; (3) Pour Liquid; and (4) Tennis Forehand. The summary of each dataset is shown in Tab. 1.

- **Break Eggs** sampled from the CMU-MMAC dataset [11] contains 5 different cooking recipes (brownies, pizza, sandwiches, salad, and scrambled eggs) captured by 43 users. While the ego and exo videos are strictly synchronized (i.e., capturing the same scene), we do not use the correspondence between videos for training.
- **Pour Milk** sampled from the H2O dataset [10] contains the scene of 10 users interacting with a milk carton using their hands. The dataset provides one egocentric video and four exocentric static videos for each scene. Some ego and exo video pairs are synchronized and the rest are asynchronous.
- **Pour Liquid** assumes a more challenging scenario as the ego and exo videos are sampled from different datasets. Therefore, those videos are fully asynchronous and captured from different environments. The ego videos consist of the "pour water" class in EPIC-Kitchens [3] and the exo videos are the "pour" category in HMDB51 [9].
- **Tennis Forehand** includes outdoor activity videos. The exocentric videos of the tennis forehand action are sampled from the Penn Action [16] dataset and the egocentric videos are collected from 12 players using the Go Pro HERO8 camera. The videos are asynchronous, covering real-world scenarios.

A.2. Downstream tasks

- **Action phase classification** aims to predict an atomic action phase label corresponding to a given frame. The Break Eggs dataset contains four action phases between 'start', 'hit egg', 'visible crack on the eggshell', 'egg contents released', and 'end.' The Pour Milk and Pour Liquid datasets contain three phases between 'start', 'liquid exits container', 'pouring complete', and 'end.' The Tennis Forehand dataset has only two phases between 'start', 'racket touches ball', and 'end.' In this document, we additionally provide a few-shot classification performance to validate the robustness of **BYOV**.
- **Frame retrieval** selects frames corresponding to a given frame using the NN search. We evaluate this task with mean average precision (mAP)@K (K=5,10,15) in the regular and cross-view settings.
- **Phase progression** quantifies how effectively the learned representations imply the progression of an action. The progression value within each phase is defined as the normalized temporal difference between the timestamp of a given frame and those of key events, scaled by the total number of frames in the video. A linear regressor is then employed to predict the phase progression values from the embeddings, where our encoders are frozen. The performance is evaluated using the average R-squared value as follows:

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2},$$

where y_t is the ground truth phase progress value, $\bar{y}$ is the

Table 1. Performance with respect to variants of the components in **BYOV**. We report the performance evaluated on the Break Eggs dataset.

Dataset	# Train		# Val		# Test		Fixed exo-view	Sync. ego-exo
	Ego	Exo	Ego	Exo	Ego	Exo		
(A) Break Eggs	61	57	5	5	10	10	✓	✓
(B) Pour Milk	29	48	4	8	7	16	✓	✗
(C) Pour Liquid	70	67	10	9	19	18	✗	✗
(D) Tennis Forehand	94	79	25	24	50	50	✗	✗

average value of all y_t, and $\hat{y}_t$ is the prediction from the linear regressor. The maximum value of R^2 is 1.

- **Kendall's τ** assesses the temporal alignment between two sequences by comparing the order of frames. Specifically, we first sample a pair of frames from one video, (u_i, u_j), and retrieve their nearest corresponding frames in the other video, (v_p, v_q). A set of frame indices (i, j, p, q) is treated as 'matched' if the temporal order of u_i and u_j and that of v_p and v_q are the same. Kendall's τ is then computed by,

$$\tau = \frac{\#\text{matched pairs} - \#\text{not matched pairs}}{\#\text{all possible pairs}}.$$

A value of 1 means the frame representations are perfectly aligned while -1 indicates the representations are aligned in the reverse order.

B. Additional Experiments

B.1. Results with different frame encoders

We mainly used the CLIP pretrained ViT-B/16 [5] to encode each frame in the main paper. To demonstrate the robustness of **BYOV** according to the frame encoders, we train **BYOV** with the CLIP pretrained ViT-L/14 [5] and ResNet-50 [8], and evaluate the performance on four tasks for four datasets. Implementation details for each frame encoder are as follows.

- **CLIP ViT-L/14** [5] pretrained on LAION-400M [12] projects each frame into 1024-dimensional 256 token embeddings different from the ViT-B/16, which has 768-dimensional 196 token embeddings. We keep the number of layers of the encoder $g_\phi(\cdot)$ and the decoder $h_\psi(\cdot)$ as 12 and 4 while setting the size of the latent space to 512. The number of trainable parameters is 51.8M (38.4M for the encoder and 13.4M for the decoder, respectively). The token selection ratio in selective token merging (STM), and the masking ratio in masked self-view modeling (MSM) and masked cross-view modeling (MCM) are set to 0.3, 0.4, and 0.8 as with the ViT-B/16.
- **ResNet-50** [8] employs convolutional neural network, and is pretrained on ImageNet-1K [4]. We extract the feature for each frame from a *Conv4c* layer of ResNet-50, which has 14×14 resolution with 1024 dimensions. We also perform the selective token merging to keep the overall framework of **BYOV**. The receptive field of each 1024-d embedding is 55×55 pixels, which is wider than 16×16 in ViT-B/16. Therefore, we reduce the selection ratio to 0.1. The masking ratio in MSM and MCM are set to 0.4 and 0.8, respectively. Similar to the ViT-L/14, we use 512-dimensional latent space for the encoder $g_\phi(\cdot)$ and the decoder $h_\psi(\cdot)$.

In Tab. 2, we first provide the zero-shot performance of the ResNet-50 (ImageNet features), CLIP ViT-B/16, and CLIP ViT-L/14. While ViT-L/14 (303M) has about three times more parameters than ViT-B/16 (86M), comparisons between the two frame encoders show that generalization capability is not dependent on the model size. Meanwhile, our **BYOV** with various frame encoders consistently outperforms the state-of-the-art [15] across tasks and datasets. In practice, **BYOV** with the ResNet-50 surpasses AE2 [15] without any additional information such as bounding boxes from the hand-object detector as in [15]. It demonstrates the robustness of the framework of our **BYOV**.

B.2. Few-shot classification

Following [15], we compare few-shot classification performance with the state-of-the-art methods [1, 6, 7, 13–15] in Tab. 3. We train the SVM classifier using 10% (or 50%) of the latents from the training data and evaluate the classification performance. Note that we train **BYOV** ten times on non-overlapped few-shot training data and report the average performance. Tab. 3 demonstrates the superior performance of **BYOV**, showing significant performance gaps to the existing works across all datasets. In particular, **BYOV** trained with only 10% training data significantly outperforms the prior best performance [15] trained with 100% training data by a large margin of 12.54 on the Pour Liquid dataset.

B.3. Frame retrieval

In Tab. 3 and Tab. 4, we report the frame retrieval performance, evaluated in both regular and cross-view settings. Comparisons with the existing methods consistently demonstrate the effectiveness of **BYOV** in both regular and cross-view retrieval across all datasets, showing an average performance improvement of about 10%.

Table 2. Performance comparison with various frame encoders on the AE2 benchmark [15]. The benchmark consists of four sub-tasks: (A) Break Eggs, (B) Pour Milk, (C) Pour Liquid, and (D) Tennis Forehand. The top results are highlighted in **bold** and the second-best results are underlined.

Task	Method	Classification (F1 score)			Frame Retrieval (mAP@10)			Phase progression	Kendall's τ
		Regular	Ego2Exo	Exo2Ego	Regular	Ego2Exo	Exo2Ego		
(A)	Random features	19.18	18.93	19.45	47.13	41.74	37.19	-0.0572	0.0018
	ImageNet features	50.24	21.48	32.25	50.49	33.09	37.80	-0.1446	0.0188
	CLIP ViT-B/16	51.66	27.97	26.24	44.46	35.85	35.70	0.0402	0.0168
	CLIP ViT-L/14	54.24	41.56	38.31	38.14	38.96	34.99	0.1672	0.0483
	AE2 [15]	66.23	57.41	71.72	65.85	64.59	62.15	0.5109	0.6316
	BYOV (ResNet-50)	72.57	67.91	70.74	68.42	63.27	63.85	0.7751	0.7463
	BYOV (ViT-B/16)	**74.30**	**75.01**	71.28	67.17	**70.65**	**69.02**	**0.8533**	**0.9451**
	BYOV (ViT-L/14)	72.41	70.11	**72.92**	**75.59**	67.73	67.55	0.8272	0.8940
(B)	Random features	36.84	33.96	41.97	52.48	50.56	51.98	-0.0477	0.0050
	ImageNet features	41.59	39.93	45.52	54.09	27.31	43.21	-2.6681	0.0115
	CLIP ViT-B/16	43.24	49.21	30.94	52.16	46.39	40.34	-4.0754	0.0046
	CLIP ViT-L/14	46.65	46.79	17.77	46.20	44.32	53.75	-0.4735	0.0503
	AE2 [15]	85.17	84.73	82.77	84.90	78.48	83.41	0.7634	0.9062
	BYOV (ResNet-50)	**86.84**	83.83	**87.00**	87.17	79.27	79.87	0.8082	0.9152
	BYOV (ViT-B/16)	86.46	85.09	86.61	**89.42**	**87.73**	**85.06**	**0.8992**	**0.9466**
	BYOV (ViT-L/14)	86.76	**85.54**	86.58	87.35	82.51	82.61	0.8407	0.9448
(C)	Random features	45.26	47.45	44.33	49.83	55.44	55.75	-0.1303	-0.0072
	ImageNet features	53.13	22.44	44.61	51.49	52.17	30.44	-1.6329	-0.0053
	CLIP ViT-B/16	60.60	36.97	48.43	43.63	47.58	37.02	-0.3139	-0.0048
	CLIP ViT-L/14	54.38	6.83	51.69	50.01	31.82	54.61	-0.2066	-0.0052
	AE2 [15]	66.56	57.15	65.60	65.54	65.79	57.35	0.1380	0.0934
	BYOV (ResNet-50)	78.63	**73.67**	76.53	**71.47**	66.74	71.17	0.3982	0.2883
	BYOV (ViT-B/16)	**79.48**	71.83	76.23	71.06	75.03	70.03	0.4483	0.3052
	BYOV (ViT-L/14)	**79.48**	71.49	**76.61**	70.36	**76.48**	**73.38**	**0.4534**	**0.3084**
(D)	Random Features	30.31	33.42	28.10	66.47	58.98	59.87	-0.0425	0.0177
	ImageNet Features	69.15	42.03	58.61	76.96	66.90	60.31	-0.4143	0.0734
	CLIP ViT-B/16	67.81	43.41	44.22	74.54	59.57	52.02	-0.4996	0.0618
	CLIP ViT-L/14	64.40	47.53	47.50	74.26	67.19	58.73	-0.4126	0.0302
	AE2 [15]	85.87	84.71	85.68	86.83	81.46	82.07	0.5060	0.6171
	BYOV (ResNet-50)	89.34	**94.83**	84.96	89.83	86.71	82.68	0.7588	0.7599
	BYOV (ViT-B/16)	89.12	94.47	85.73	90.61	**88.34**	**88.94**	**0.7881**	0.7852
	BYOV (ViT-L/14)	**89.56**	94.48	**86.51**	**91.21**	87.04	88.33	0.7653	**0.8101**

We illustrate examples of cross-view frame retrieval from the Pour Milk and Tennis Forehand datasets in Fig. 1. Given the query frame (blue box) from one view, we retrieve the frames (red box) from the other view videos using NN search. The results show that the query and retrieved frames are contextually well-aligned through the action states. In addition, properly retrieved frames demonstrate that **BYOV** captures contexts over time. For example, the frames with the action phases of 'pre-pour' and 'pouring complete' are visually similar, however, **BYOV** successfully performs frame retrieval by capturing the context with respect to the action state over time. In this regard, we further analyze the effectiveness of **BYOV** by visualizing the frame embeddings in the following section.

B.4. Ablation study

We analyze the effectiveness of each component in **BYOV**, including the size of latent space, token selection ratio in STM, and masking ratio in MSM and MCM. Note that we use the CLIP pretrained ViT-B/16 as the frame encoder for the following experiments.

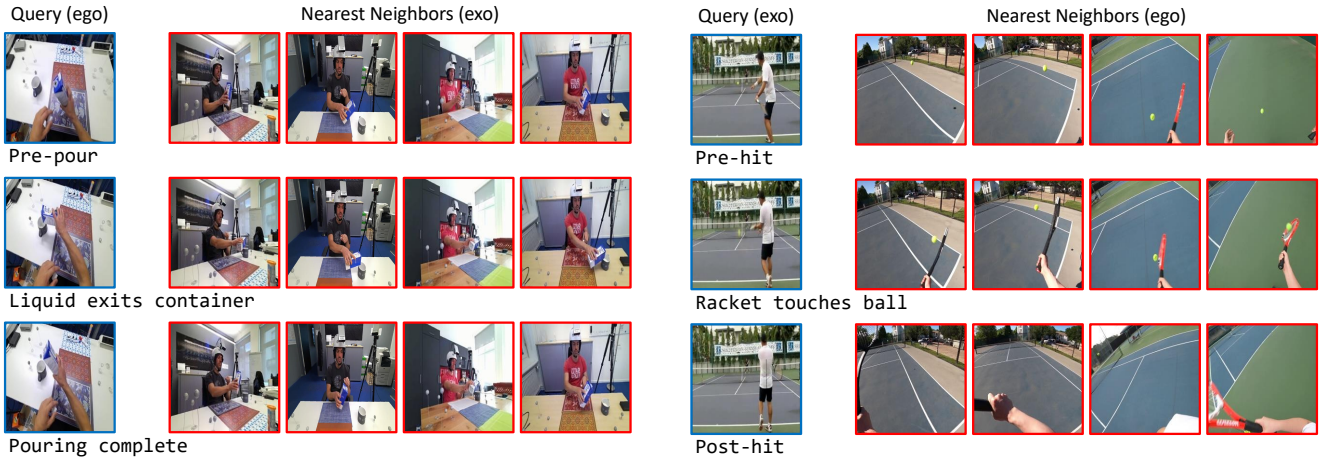
Hidden dimension of autoencoders. The encoder $g_\phi(\cdot)$ maps the frame token embeddings into the 256-dimensional latents, such that the encoder and decoder have 9.7M and 2.6M trainable parameters, respectively. To assess the impact of latent space size on performance, we train **BYOV**

Table 3. Performance comparison for few-shot classification and regular frame retrieval on the AE2 benchmark [15]. The benchmark consists of four sub-tasks: (A) Break Eggs, (B) Pour Milk, (C) Pour Liquid, and (D) Tennis Forehand. We report the few-shot classification (F1 score) and regular frame retrieval (mAP@5, mAP@10, and mAP@15) performance. The top results are highlighted in **bold** and the second-best results are underlined.

Task	Method	Few-shot Classification (F1 score)			Regular 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
	CLIP ViT-B/16	46.46	49.18	51.66	44.89	44.46	43.44
	CLIP ViT-L/14	47.36	51.80	54.24	38.47	38.14	37.72
	ActorObserverNet [14]	31.40	35.63	36.14	50.92	50.47	49.72
	TCN [13] (single-view)	52.30	54.90	56.90	52.82	53.42	53.60
	TCN [13] (multi-view)	56.88	59.25	59.91	59.11	58.83	58.44
	TCN [13] (unpaired multi-view)	56.13	56.65	56.79	58.18	57.78	57.21
	CARL [1]	39.18	41.92	43.43	47.14	46.04	44.99
	TCC [6]	57.54	59.18	59.84	59.33	58.75	57.99
	GTA [7]	56.89	56.77	56.86	62.79	61.55	60.38
	AE2 [15]	63.95	64.86	66.23	66.86	65.85	64.73
	BYOV (ViT-B/16)	**71.84**	**73.92**	**74.30**	**67.28**	**67.17**	**66.40**
(B)	Random features	36.84	33.96	41.97	52.48	50.56	51.98
	ImageNet features	41.59	39.93	45.52	54.09	27.31	43.21
	CLIP ViT-B/16	39.44	38.90	43.24	53.29	52.16	51.55
	CLIP ViT-L/14	42.68	39.91	46.65	46.20	46.20	53.75
	TCN [13] (single-view)	43.60	46.83	47.39	56.98	57.00	56.46
	CARL [1]	48.73	48.78	48.79	55.29	55.01	54.23
	TCC [6]	78.69	77.97	77.91	81.22	80.97	80.46
	GTA [7]	79.82	80.96	81.11	80.65	80.12	79.68
	AE2 [15]	85.17	85.12	85.17	85.25	84.90	84.55
	BYOV (ViT-B/16)	**86.12**	**86.44**	**86.46**	**90.99**	**89.42**	**88.98**
(C)	Random features	45.26	47.45	44.33	49.83	55.44	55.75
	ImageNet features	53.13	22.44	44.61	51.49	52.17	30.44
	CLIP ViT-B/16	57.21	35.46	60.60	42.34	43.63	44.03
	CLIP ViT-L/14	51.72	28.30	54.38	48.56	50.01	50.52
	TCN [13] (single-view)	54.62	55.08	54.02	48.50	48.83	49.03
	CARL [1]	51.68	55.67	56.98	55.03	55.29	54.93
	TCC [6]	52.37	51.70	52.53	62.93	62.33	61.44
	GTA [7]	55.91	56.87	56.92	62.83	62.79	62.12
	AE2 [15]	65.88	66.53	66.56	66.55	65.54	64.66
	BYOV (ViT-B/16)	**79.10**	**79.28**	**79.48**	**73.89**	**71.06**	**67.83**
(D)	Random Features	30.31	33.42	28.10	66.47	58.98	59.87
	ImageNet Features	69.15	42.03	58.61	76.96	66.90	60.31
	CLIP ViT-B/16	70.37	48.01	67.81	76.45	74.54	73.15
	CLIP ViT-L/14	68.44	42.95	64.40	75.61	74.26	73.14
	TCN [13] (single-view)	65.78	69.19	68.87	74.05	73.76	73.10
	CARL [1]	58.89	59.38	59.69	72.94	69.43	67.14
	TCC [6]	67.71	77.07	78.41	82.78	80.24	78.59
	GTA [7]	80.31	83.04	83.63	86.59	85.20	84.33
	AE2 [15]	85.24	85.72	85.87	87.94	86.83	86.05
	BYOV (ViT-B/16)	**88.78**	**89.01**	**89.12**	**90.89**	**90.61**	**90.87**

Table 4. Performance comparison for cross-view retrieval on the AE2 benchmark [15]. The benchmark consists of four sub-tasks: (A) Break Eggs, (B) Pour Milk, (C) Pour Liquid, and (D) Tennis Forehand. We report the cross-view frame retrieval (mAP@5, mAP@10, and mAP@15) performance. The top results are highlighted in **bold** and the second-best results are underlined.

Task	Method	Ego2Exo Frame Retrieval			Exo2Ego Frame Retrieval		
		mAP@5	mAP@10	mAP@15	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
	CLIP ViT-B/16	35.80	35.85	34.92	34.91	35.70	35.96
	CLIP ViT-L/14	39.30	38.94	38.14	35.23	34.99	33.98
	ActorObserverNet [14]	43.57	42.70	41.56	42.00	41.29	40.48
	TCN [13] (single-view)	31.12	32.63	33.73	34.67	34.91	35.31
	TCN [13] (multi-view)	46.38	47.04	46.96	52.50	52.68	52.43
	TCN [13] (unpaired multi-view)	55.34	54.64	53.75	58.79	57.87	57.07
	CARL [1]	37.89	37.38	36.57	40.37	39.94	39.38
	TCC [6]	62.11	61.11	60.33	62.39	62.03	61.25
	GTA [7]	57.11	56.25	55.10	54.47	53.93	53.22
	AE2 [15]	65.70	64.59	63.76	62.48	62.15	61.80
	BYOV (ViT-B/16)	**72.76**	**70.65**	**70.27**	**71.79**	**69.02**	**68.94**
(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
	CLIP ViT-B/16	46.37	46.39	46.86	41.28	40.34	39.86
	CLIP ViT-L/14	43.71	44.32	44.20	55.55	53.75	53.10
	TCN [13] (single-view)	47.00	46.48	45.42	47.94	47.20	46.59
	CARL [1]	54.35	52.99	51.99	51.14	51.51	51.00
	TCC [6]	75.54	75.30	75.02	80.44	80.27	80.18
	GTA [7]	72.55	72.78	72.96	75.16	75.40	75.48
	AE2 [15]	78.21	78.48	78.78	83.88	83.41	83.05
	BYOV (ViT-B/16)	**85.15**	**87.73**	**87.80**	**85.48**	**85.06**	**85.00**
(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
	CLIP ViT-B/16	42.08	47.58	49.78	35.14	37.02	36.71
	CLIP ViT-L/14	32.33	31.82	31.59	54.01	54.61	54.64
	TCN [13] (single-view)	53.60	55.28	55.46	29.16	31.15	31.95
	CARL [1]	59.59	59.37	59.19	34.73	36.80	38.10
	TCC [6]	55.98	56.08	56.13	58.11	57.89	57.15
	GTA [7]	57.03	58.52	59.00	51.71	53.32	53.54
	AE2 [15]	66.23	65.79	65.00	57.42	57.35	57.03
	BYOV (ViT-B/16)	**79.06**	**75.03**	**72.73**	**76.21**	**70.03**	**69.44**
(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
	CLIP ViT-B/16	60.63	59.57	58.46	52.25	52.02	52.12
	CLIP ViT-L/14	69.02	67.19	65.44	61.83	58.73	57.05
	TCN [13] (single-view)	54.12	55.08	55.05	56.70	56.65	55.84
	CARL [1]	52.18	54.83	55.39	65.94	63.19	60.83
	TCC [6]	57.87	55.84	53.81	48.62	47.27	46.11
	GTA [7]	78.93	78.00	77.01	79.95	79.14	78.52
	AE2 [15]	82.58	81.46	80.75	82.82	82.07	81.69
	BYOV (ViT-B/16)	**88.55**	**88.34**	**87.98**	**90.64**	**88.94**	**87.26**



(a) Ego2Exo frame retrieval on Pour Milk

(b) Exo2Ego frame retrieval on Tennis Forehand

Figure 1. Qualitative examples of frame retrieval from the Pour Milk and Tennis Forehand datasets. We retrieve the nearest neighbor frames (red box) corresponding to the given query frame (blue box).

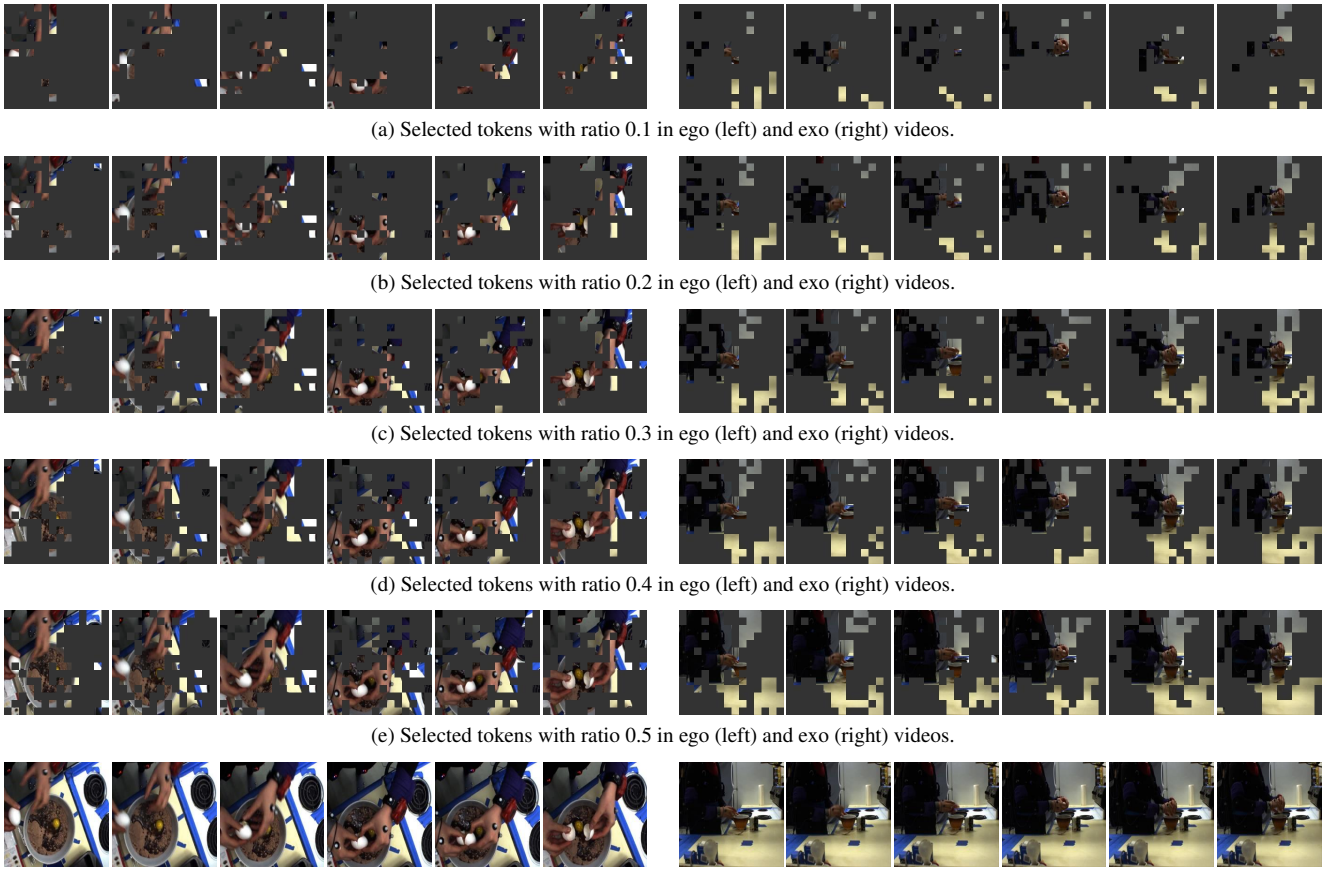
Table 5. Performance comparison according to various sizes of latent space in **BYOV**. We evaluate the performance on the Break Eggs dataset.

Latent Size	Trainable Params	Classification (F1 score)			Frame Retrieval (mAP@10)			Phase progression	Kendall's τ
		Regular	Ego2Exo	Exo2Ego	Regular	Ego2Exo	Exo2Ego		
64	0.9M	71.55	71.34	69.36	65.87	64.09	68.16	0.8362	0.8943
128	3.4M	70.84	72.74	69.71	67.07	66.70	68.45	0.8407	0.9240
256	12.3M	**74.30**	**75.01**	**71.28**	67.17	70.65	69.02	**0.8533**	**0.9451**
512	51.5M	70.89	70.19	68.72	**68.70**	**73.29**	**74.45**	0.8107	0.9240

with various latent sizes and evaluate the performance on the Break Eggs dataset. Tab. 5 summarizes the results, including the performance on downstream tasks and the number of trainable parameters corresponding to each latent size. Naturally, large latent spaces enhance representation capability but lead to more trainable parameters (e.g. 51.5M parameters with a 512-dimensional latent space for 12 encoder and 4 decoder layers) and require more extensive training data. The results indicate that increasing the latent size from 64 to 256 consistently improves performance. However, a further increase to a 512-dimensional latent space leads to performance degradation, attributed to the limited availability of training data.

Token selection ratio. Selective token merging (STM) allows **BYOV** to effectively capture action-related regions while excluding noisy regions without any training as shown in the main paper. We provide the performance of **BYOV** with various token selection ratios in the first panel of Tab. 6 and depict the selected tokens corresponding to each selection ratio in Fig. 2. The results show that the token selection ratio significantly affects the performance due to the difference in the field of view between ego and exo videos. In other words, a low selection ratio is insufficient to cover the action-related regions in ego videos (see Fig. 2a and Fig. 2b), while a high selection ratio makes noisy tokens be included in exo videos (see Fig. 2e). To balance the lack of information in the ego video and the unnecessary noise in the exo video, we set the token selection ratio to 0.3.

Masking ratio. We validate the effectiveness of the masking ratio in masked self-view modeling (MSM) and masked cross-view modeling (MCM) in the second and third panels of Tab. 6. In MSM, we can guess a low masking ratio enables the model to easily solve each masked modeling problem, leading to insufficient causality learning. In practice, the results show significant performance drops in phase progression and Kendall's τ. An extremely high masking ratio in MSM makes learning the causality between frames hard as the decoder takes only a few clean tokens (or only masked tokens with a 100% masking ratio). The low masking ratio in MCM degrades performance for a similar reason as in MSM. Meanwhile, a high masking ratio in MCM makes the masked cross-view modeling significantly difficult to solve with limited training data, showing performance drops across all downstream tasks. **BYOV** trained with the masking ratio of 0.4 and 0.8 in MSM and MCM achieves to produce the effective fine-grained view-



(a) Selected tokens with ratio 0.1 in ego (left) and exo (right) videos.

(b) Selected tokens with ratio 0.2 in ego (left) and exo (right) videos.

(c) Selected tokens with ratio 0.3 in ego (left) and exo (right) videos.

(d) Selected tokens with ratio 0.4 in ego (left) and exo (right) videos.

(e) Selected tokens with ratio 0.5 in ego (left) and exo (right) videos.

(f) Complete frames from ego (left) and exo (right) videos.

Figure 2. Visualization of selected tokens at each frame sampled from ego (left) and exo (right) videos. Note that complete frames are identical with the token selection ratio of 1.0.

invariant video representations.

B.5. Failure cases

While our **BYOV** significantly improves the performance across various benchmarks and experimental protocols, we observed that most failure cases occur in videos with slow movement transitions, particularly in exocentric videos. In such cases, frame embeddings tend to attend to each other uniformly, reducing the model's ability to capture meaningful temporal dependencies. Fig. 3 illustrates a visualization of the softmax similarity score between the final frame embeddings for a failure case from the *Pour Liquid* benchmark. Despite introducing positional embeddings and selective token merging, the embedding feature for a reference frame (red box) attends to all other embeddings similarly, resulting in less informative final representations. Beyond simple token selection of **BYOV**, learning-based token selection approach [2] may further improve the robustness of learned representations.

C. Broader Impact

By achieving robust, view-invariant learning from unpaired ego-exo videos, **BYOV** can significantly advance the ability of AI to understand human actions and interactions across diverse perspectives, contributing to a wide range of real-world applications such as robotics, augmented and virtual reality, and assistive technologies. Moreover, this research can facilitate new related research as follows;

- Cross-view video generation: The video representations learned by **BYOV** contain fine-grained action context. In addition, the decoders used during training show a high recovery rate. This shows that it is possible to generate videos across views, which can be used to generate educational or instructional videos.
- Multi-view activity tracking: The view-consistent representations can be used in continuously tracking a person or object across various camera views (ego and exo) to maintain consistent identity and action recognition across perspectives, useful for applications in security and autonomous vehicles.

Table 6. Performance comparison according to variants of the hyperparameters in **BYOV**. We report the performance evaluated on the Break Eggs dataset.

Ratio (%)			Classification (F1 score)			Frame Retrieval (mAP@10)			Phase progression	Kendall's τ
STM	MSM	MCM	Regular	Ego2Exo	Exo2Ego	Regular	Ego2Exo	Exo2Ego		
Effectiveness of token selection ratio										
10	40	80	41.45	21.13	20.03	56.05	46.06	46.85	0.1858	0.0157
20	40	80	70.97	69.60	66.27	65.05	71.13	64.52	0.6597	0.7978
30	40	80	**74.30**	**75.01**	**71.28**	67.17	70.65	**69.02**	**0.8533**	**0.9451**
40	40	80	72.39	72.59	69.19	**68.20**	**73.79**	67.44	0.8299	0.8963
50	40	80	71.56	69.05	68.79	**68.20**	72.79	67.20	0.8299	0.8926
100	40	80	71.34	72.58	65.07	67.44	69.32	67.87	0.7894	0.8957
Effectiveness of masking ratio in MSM										
30	10	80	70.71	69.51	66.20	67.67	66.10	63.89	0.5228	0.6724
30	20	80	71.22	70.38	69.81	67.67	68.27	65.83	0.8134	0.9126
30	30	80	72.28	73.21	70.22	**67.28**	70.21	68.15	0.8330	0.9337
30	40	80	**74.30**	**75.01**	**71.28**	67.17	70.65	69.02	**0.8533**	**0.9451**
30	50	80	72.87	73.71	70.87	**67.28**	**71.21**	**70.15**	0.8398	0.9410
30	100	80	66.65	69.97	68.24	65.01	67.48	66.86	0.6916	0.7818
Effectiveness of masking ratio in MCM										
30	40	0	67.23	66.65	67.10	60.38	58.44	56.97	0.7019	0.8040
30	40	20	71.40	69.06	70.19	64.98	62.09	61.27	0.8269	0.9112
30	40	40	73.23	73.81	71.17	**68.84**	65.94	68.22	0.8133	0.9247
30	40	60	73.33	74.54	**71.32**	67.21	**70.65**	**69.02**	0.8480	0.9440
30	40	80	**74.30**	**75.01**	71.28	67.17	**70.65**	**69.02**	**0.8533**	**0.9451**
30	40	100	71.09	70.01	70.47	65.34	66.50	68.63	0.7435	0.8354

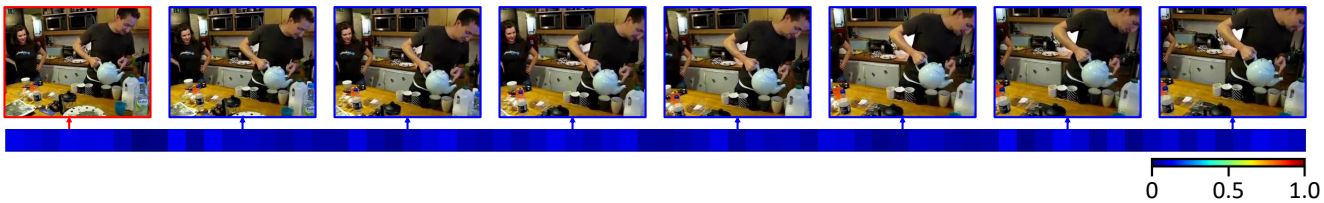


Figure 3. Visualization of the softmax similarity between final frame embeddings for a failure case from the *Pour Liquid* benchmark. We depict the similarity score between only one reference token embedding (red box) and other token embeddings (blue boxes) for visibility.

References

[1] Minghao Chen, Fangyun Wei, Chong Li, and Deng Cai. Frame-wise action representations for long videos via sequence contrastive learning. In *CVPR*, 2022. 2, 4, 5

[2] Rohan Choudhury, Guanglei Zhu, Sihan Liu, Koichiro Niinuma, Kris Kitani, and László Jeni. Don't look twice: Faster video transformers with run-length tokenization. In *NeurIPS*, pages 28127–28149, 2024. 7

[3] 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 *ECCV*, 2018. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2

[6] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *CVPR*, 2019. 2, 4, 5

[7] Isma Hadji, Konstantinos G. Derpanis, and Allan D. Jepson. Representation learning via global temporal alignment and cycle-consistency. In *CVPR*, 2021. 2, 4, 5

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2

[9] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011. 1

[10] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and

Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021. 1

[11] 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. *Tech. report CMU-RI-TR-08-22*, 2009. 1

[12] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. In *NeurIPS Workshop*, 2021. 2

[13] 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 *ICRA*, 2018. 2, 4, 5

[14] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *CVPR*, 2018. 4, 5

[15] Zihui Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In *NeurIPS*, 2023. 1, 2, 3, 4, 5

[16] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly supervised representation for detailed action understanding. In *ICCV*, 2013. 1