FILS: Self-Supervised Video Feature Prediction In Semantic Language Space

BMVC 2024
m.ahmadian@surrey.ac.uk
Code: github.com/Moohnai
Mona Ahmadian Frank Guerin Andrew Gilbert
UNIVERSITY OF SURREY

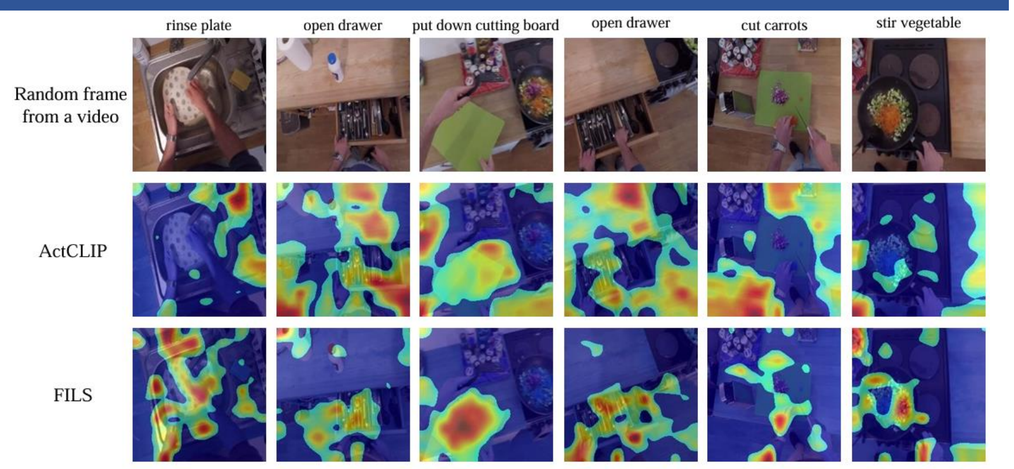
Contributions

Motivation behind FILS is to bridge vision and language, leveraging text descriptions to guide video representation, enabling a more semantic comprehension of action and context in video data.

• Feature prediction strategy on masked video and patch-wise contrastive learning within potential action

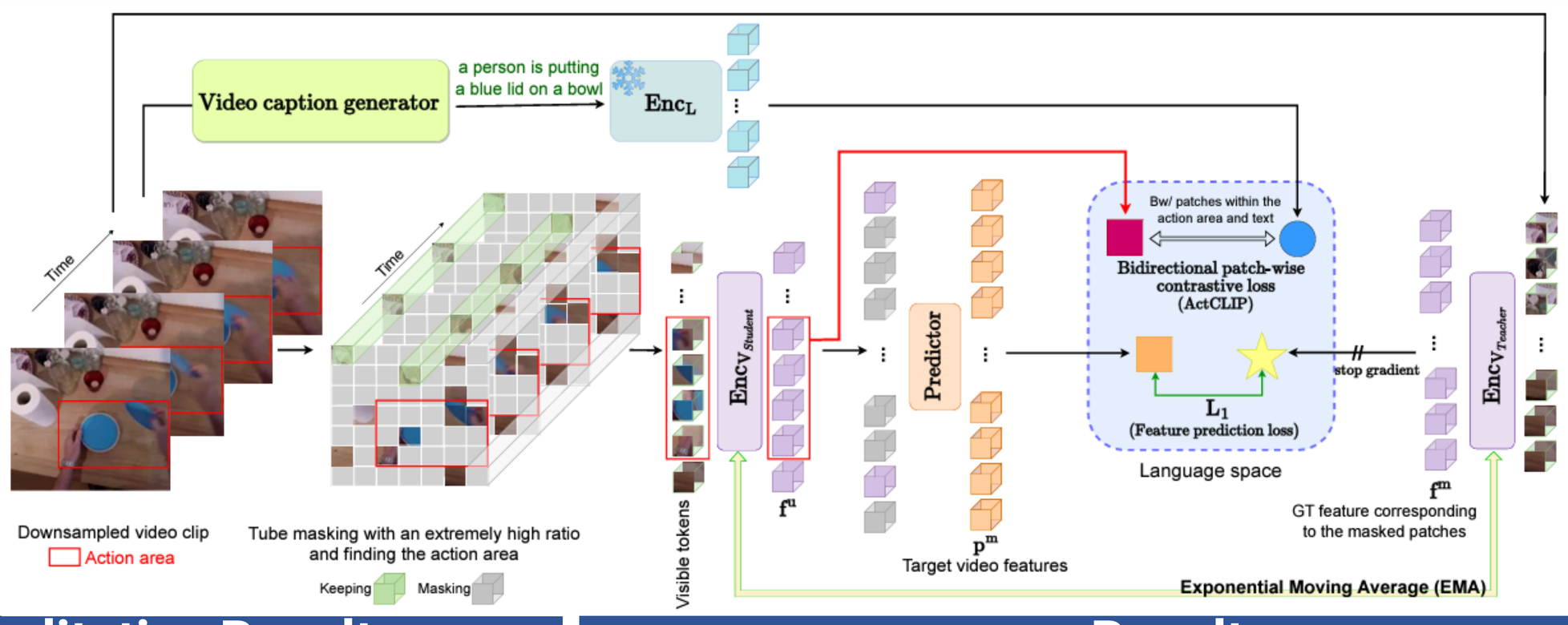
• **ActCLIP** is a technique used within FILS that applies CLIP-based contrastive learning specifically to patches in video action areas, aligning motion regions with their corresponding text descriptions

• Demonstrating the effectiveness of using mutual information to prioritize patches that convey semantic and action-rich details, enhancing learned representations for various tasks



Visualization of the similarity between text (action label) and video features

Method



Qualitative Results

FILS represents the potential semantic region in the video and acquires an understanding of spatiotemporal relationships

Results

Superior performance across all metrics on the action recognition datasets

Epic-Kitchens

Method	Backbone	p-data	L	Verb Top-1	Noun Top-1	Action Top-1
SlowFast [13]	ResNet101	K400	×	65.6	50.0	38.5
TSM [28]	ResNet50	IN-1K	×	67.9	49.0	38.3
Mformer [33]	ViT-L	IN-21K+K400	×	67.1	57.6	44.1
Video Swin [31]	Swin-B	K400	×	67.8	57.0	46.1
ViViT FE [2]	ViT-L	IN-21k+K400	×	66.4	56.8	44.0
IPL [51]	I3D	K400	✓	68.6	51.2	41.0
Omnivore [15]	Swin-B	IN+K400+SUN	×	69.5	61.7	49.9
MeMViT [53]	ViT-B	K600	×	71.4	60.3	48.4
MTV [55]	MTV-B	WTS-60M	×	69.9	63.9	50.5
LaViLa [63]	TSF-B	WIT+Ego4D	✓	69.0	58.4	46.9
AVION [62]	ViT-B	WIT+Ego4D	✓	70.0	59.8	49.1
VideoMAE* [46]	ViT-B	EK100	✓	-	-	48.5
FILS(ours)	ViT-B	EK100	✓	72.2	61.7	51.0

Something-Something V2

Method	Backbone	p-data	L	Top-1
SlowFast [13]	ResNet101	K400	×	63.1
TSM [28]	ResNet50	K400	×	63.4
TimeSformer [6]	ViT-L	IN-21K	×	62.4
Mformer [33]	ViT-L	IN-21K+K400	×	68.1
Video Swin [31]	Swin-B	K400	×	69.6
ViViT FE [2]	ViT-L	IN-21k+K400	×	65.9
VIMPAC [45]	ViT-L	HowTo100M	✓	68.1
BEVT [50]	Swin-B	IN-1K + K400	×	70.6
VideoMAE [46]	ViT-B	SSV2	×	70.8
OmniMAE [16]	ViT-B	IN-1K + SSv2	×	69.5
Omnivore [15]	Swin-B	IN-21k+K400	×	71.4
VideoMAE V2 [47]	ViT-B	UnlabeledHybrid	×	71.2
FILS(ours)	ViT-B	SSV2	✓	72.1

Charades-Ego

Method	Backbone	p-data	L	mAP
ActorObserverNet [17]	ResNet-152	Charades	×	20
SSDA [4]	I3D	Charades-Ego	×	25.8
Ego-Exo [10]	SlowFast-R101	Kinetics-400	×	30.1
EgoVLP [12]	TSF-B	Ego4D	✓	32.1
HierVL-Avg [1]	ViT-Base	Ego4D	✓	32.6
HierVL-SA [1]	ViT-Base	Ego4D	✓	33.8
EgoVLPv2 [14]	TSF-B	EgoClip	✓	34.1
LaViLA [24]	TSF-B	WIT+Ego4D	✓	33.7
FILS(ours)	ViT-Base	EK100	✓	**34.4**
FILS(ours)	ViT-Base	SSV2	✓	**34.2**

EGTEA

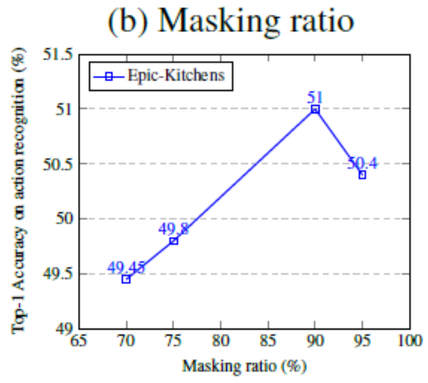
Method	Backbone	p-data	L	Top-1 Acc.	Mean Acc.
Li et al. [11]	I3D	K400	×	-	53.30
LSTA [19]	ConvLSTM	IN-1k	×	61.86	53.00
IPL [21]	I3D	K400	✓	-	60.15
MTCN [9]	SlowFast	K400+VGG-S	✓	73.59	65.87
LaViLA [24]	TSF-B	WIT+Ego4D	✓	77.45	70.12
FILS(ours)	ViT-Base	EK100	✓	**78.48**	**71.20**
FILS(ours)	ViT-Base	SSV2	✓	**78.57**	**71.31**

Ablation Studies

(a) ActCLIP strategies

Method	strategy	number of iteration	Action Top-1 Acc.
FILS	patch	10	38.0
FILS	patch	30	43.5
FILS	patch	50	46.3
FILS	patch-average	1	51.0

(b) Masking ratio



Synthetic captions