Implicit Temporal Modeling with Learnable Alignment for Video Recognition

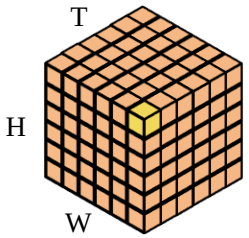
Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, Yu-Gang Jiang

Fudan University, Microsoft Research, CMU

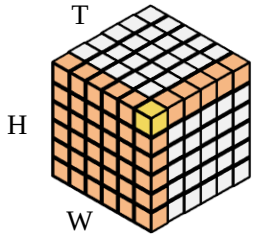
Microsoft Research

Carnegie Mellon University

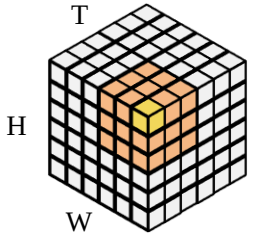
Motivation



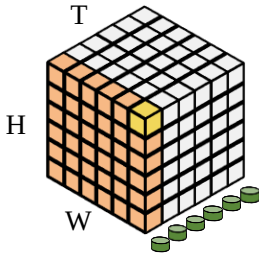
Joint Space-Time Attention



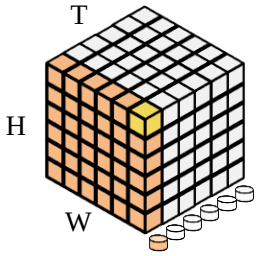
Divided Space-Time Attention (TimeSformer)



Local Space-Time Attention (VideoSwin)



Cross-frame Attention (X-CLIP)

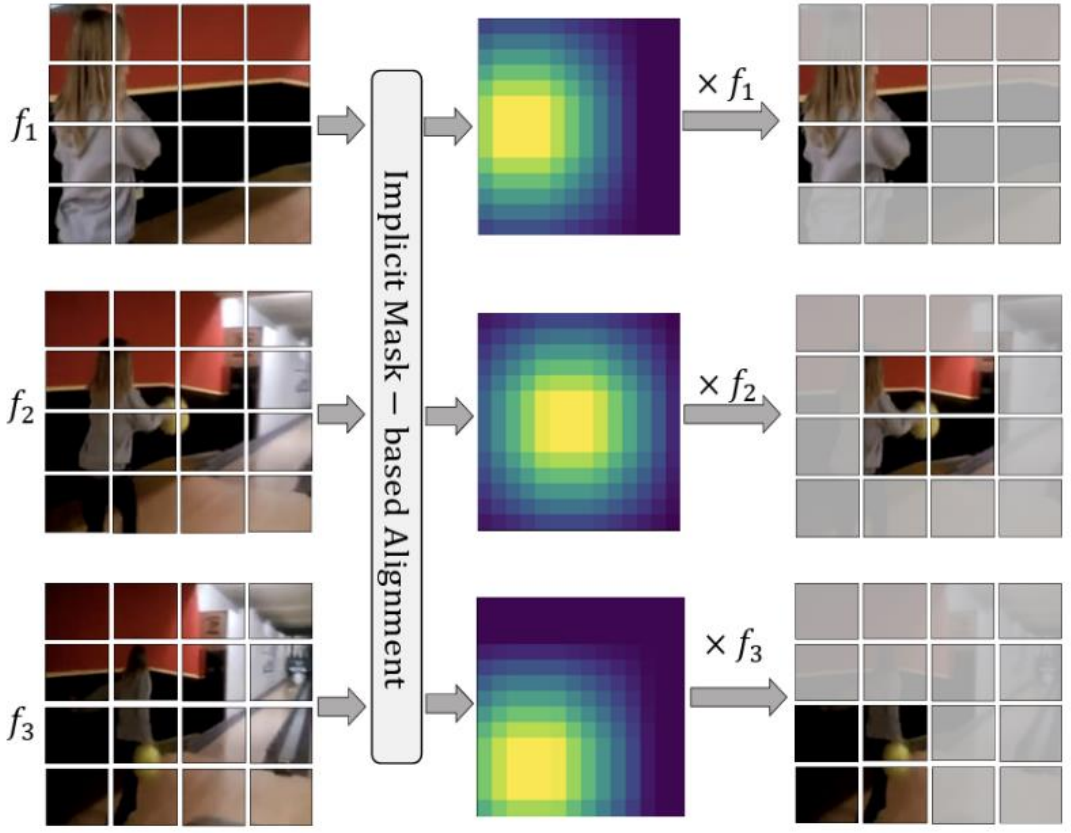


Implicit Temporal Modeling (Ours)

- Current temporal modeling approaches still depend on the complex self-attention.
- These methods have suffered from either high computation cost or insufficient modeling abilities.

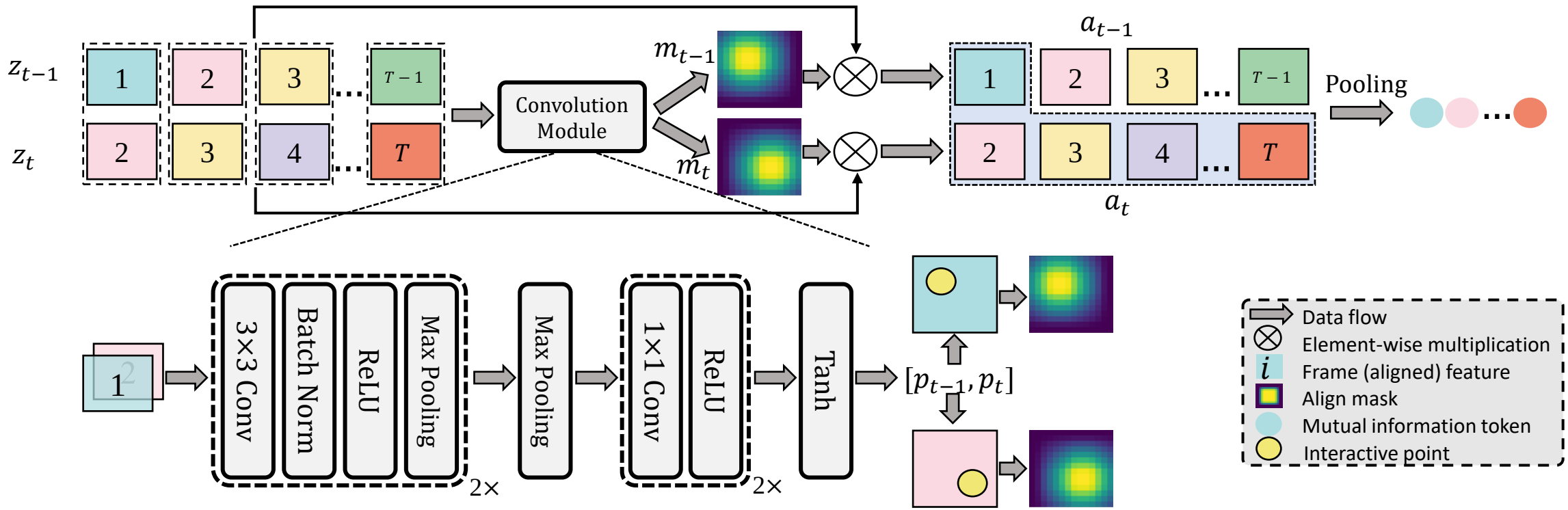
Motivation

- Important motion and action clues can be derived when performing alignment of pairwise frames.
- Simple implicit and coarse alignment is sufficient.
- Temporal attention is unnecessary.



Method

Implicit Mask-based Alignment



$$\boldsymbol{p}_t^{(\ell)},\ \boldsymbol{p}_{t-1}^{(\ell)} = \texttt{Conv}(\texttt{Concat}(\mathbf{z}_t^{(\ell-1)}, \mathbf{z}_{t-1}^{(\ell-1)}))$$

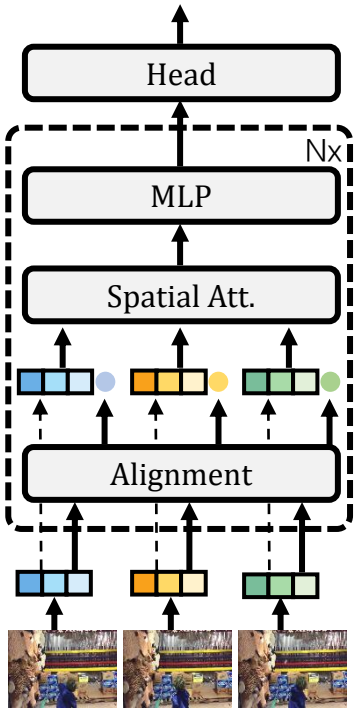
$$s = \texttt{dist}(\boldsymbol{u}, \boldsymbol{p}_t^{(\ell)}),$$

$$w_{\boldsymbol{u}} = \begin{cases} \eta, & \text{if } s \le \delta, \\ \max\left(0, \eta - \beta\left(s - \delta\right)\right), & \text{if } s > \delta, \end{cases}$$

$$\mathbf{a}_t^{(\ell)} = \mathbf{m}_t^{(\ell)} \mathbf{z}_t^{(\ell-1)},$$
$$\mathbf{a}_{t-1}^{(\ell)} = \mathbf{m}_{t-1}^{(\ell)} \mathbf{z}_{t-1}^{(\ell-1)}.$$

Method

Block Design



Implicit Spatio-Temporal attention block

Pairwise Alignment: $\mathbf{a}_t^{(\ell)},\ \mathbf{a}_{t-1}^{(\ell)} = \texttt{Align}(\mathbf{z}_t^{(\ell-1)}, \mathbf{z}_{t-1}^{(\ell-1)})$

Average Pooling: $\hat{\mathbf{z}}_{t,mut}^{(\ell)} = \texttt{Avg}(\mathbf{a}_t^{(\ell)}),$

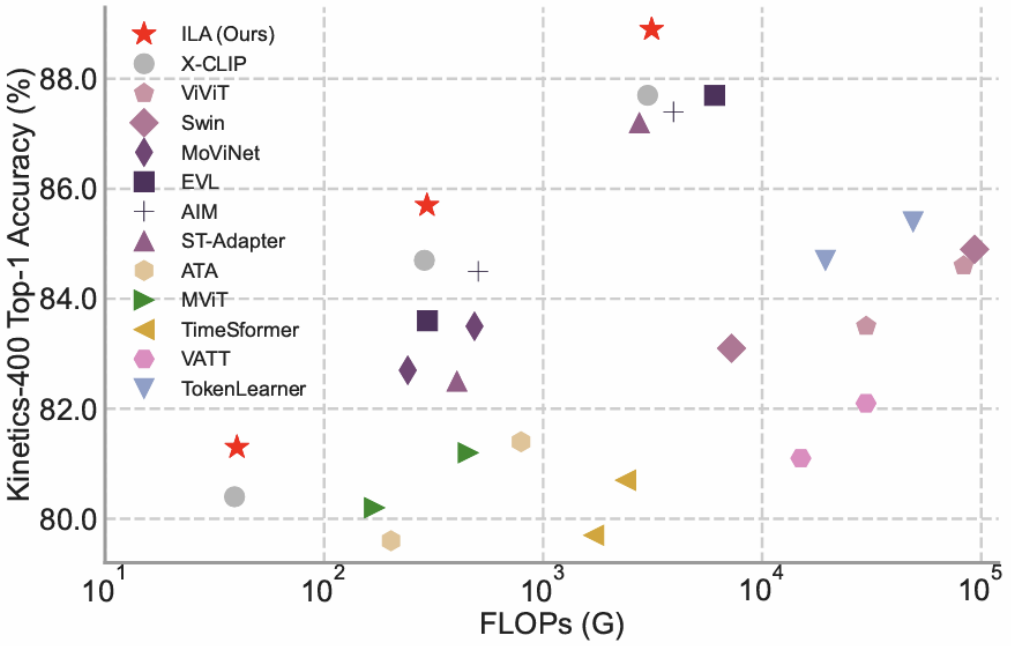
Spatial Attention: $[\tilde{\mathbf{z}}_t^{(\ell)},\ \tilde{\mathbf{z}}_{t,mut}^{(\ell)}] = \texttt{MSA}(\texttt{LN}([\mathbf{z}_t^{(\ell-1)}, \hat{\mathbf{z}}_{t,mut}^{(\ell)}])) + [\mathbf{z}_t^{(\ell-1)}, \hat{\mathbf{z}}_{t,mut}^{(\ell)}]$

Feed-Forward: $\mathbf{z}_t^{(\ell)} = \texttt{MLP}(\texttt{LN}(\tilde{\mathbf{z}}_t^{(\ell)})) + \tilde{\mathbf{z}}_t^{(\ell)}.$

Experiments

Results on Kinetics-400

Model	Pretrain	Frames	Top-1	Top-5	Views	FLOPs (G)
TimeSformer-L	IN-21K	96	80.7	94.7	1x3	2380
Swin-L@384	IN-21K	32	84.9	96.7	10x5	2107
MViTv2-L@312	IN-21K	40	86.1	97.0	5x3	2828
EVL-B/16	CLIP-400M	16	83.6	-	1x3	296
EVL-L/14@336	CLIP-400M	32	87.7	-	1x3	6068
XCLIP-B/16	CLIP-400M	16	84.7	96.8	4x3	287
XCLIP-L/14@336	CLIP-400M	16	87.7	97.4	4x3	3086
AIM-L/14	CLIP-400M	16	87.3	97.6	1x3	1868
ST-Ada-L/14	CLIP-400M	16	86.9	97.6	1x3	1375
ILA-B/16	CLIP-400M	16	85.7	97.2	4x3	295
ILA-L/14	CLIP-400M	8	88.0	98.1	4x3	673
ILA-L/14@336	CLIP-400M	16	88.7	97.8	4x3	3130



Experiments

Results on Something-Something V2

Model	Pretrain	Frames	Top-1	Top-5	Views	FLOPs (G)
ViViT-L	IN-21K+K400	16	65.4	89.8	1x3	903
TimeSformer-L	IN-21K	96	62.4	81.0	1x3	2380
MViTv1-B	K400	16	64.7	89.2	1x3	70.5
EVL-B/16	CLIP-400M	16	61.7	-	1x3	345
EVL-L/14	CLIP-400M	32	66.7	-	1x3	3216
EVL-L/14@336	CLIP-400M	32	68.0	-	1x3	8090
XCLIP-B/16	CLIP-400M	8	57.8	84.5	4x3	145
AIM-B/16	CLIP-400M	8	66.4	90.5	1x3	208
AIM-L/14	CLIP-400M	32	69.4	92.3	1x3	3836
ILA-B/16	CLIP-400M	16	66.8	90.3	4x3	438
ILA-L/14	CLIP-400M	8	67.8	90.5	4x3	907
ILA-L/14@336	CLIP-400M	16	70.2	91.8	4x3	3723

Experiments

Ablation Study

Generalization to different backbones

Model	Pre-training	Acc. (%)	FLOPs
EVL [35]	CLIP-400M	82.9	**150G**
EVL + ILA	CLIP-400M	**83.5**	162G
TimeSformer [5]	IN-21K	78.0	196G
TimeSformer + ILA	IN-21K	**79.8**	**164G**

Comparison of mutual information

Model	Acc. (%)	MI (EMD)
Baseline	79.8	0.56
X-CLIP [38]	80.4	0.51
CLIP + Divided ST Attention [5]	80.6	0.47
CLIP + ATA [70]	81.0	0.30
CLIP + ILA	**81.3**	**0.13**

Effectiveness of implicit alignment

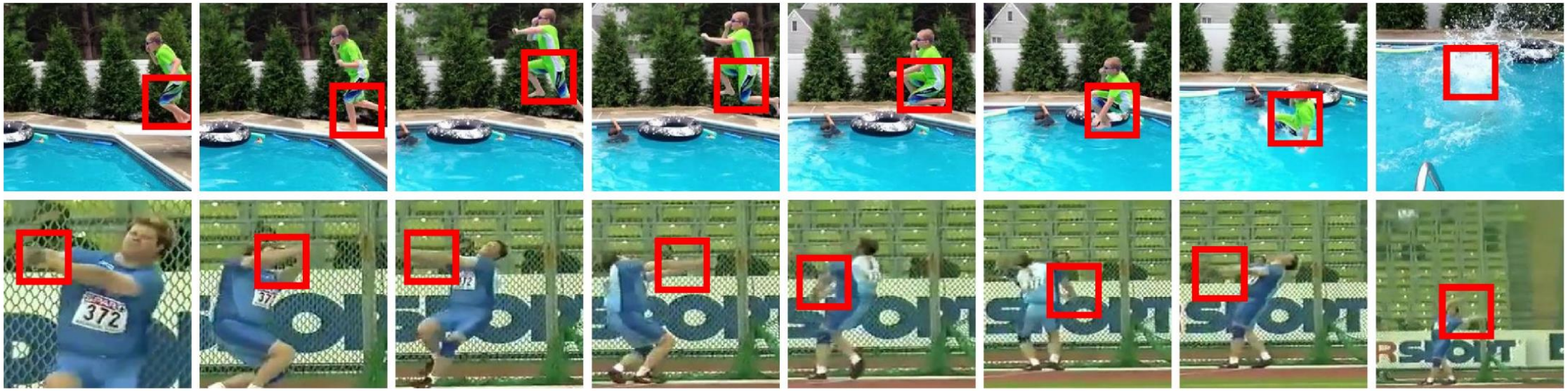
Model	Acc. (%)	FLOPs
Baseline	79.8	37G
X-CLIP [38]	80.4	39G
CLIP + Divided ST Attention [5]	80.6	58G
CLIP + Temporal Shift [56]	80.1	**37G**
CLIP + ATA [70]	81.0	60G
CLIP + Average Pooling	80.4	39G
CLIP + ILA	**81.3**	40G

Impact of different aligning strategies

Aligning Strategy	Top1. (%)	Top5. (%)
Align-First	80.7	94.5
Align-Middle	80.8	94.6
Adjacent frame	**81.3**	**95.0**

Experiments

Visualization of Interactive Points over time



Experiments

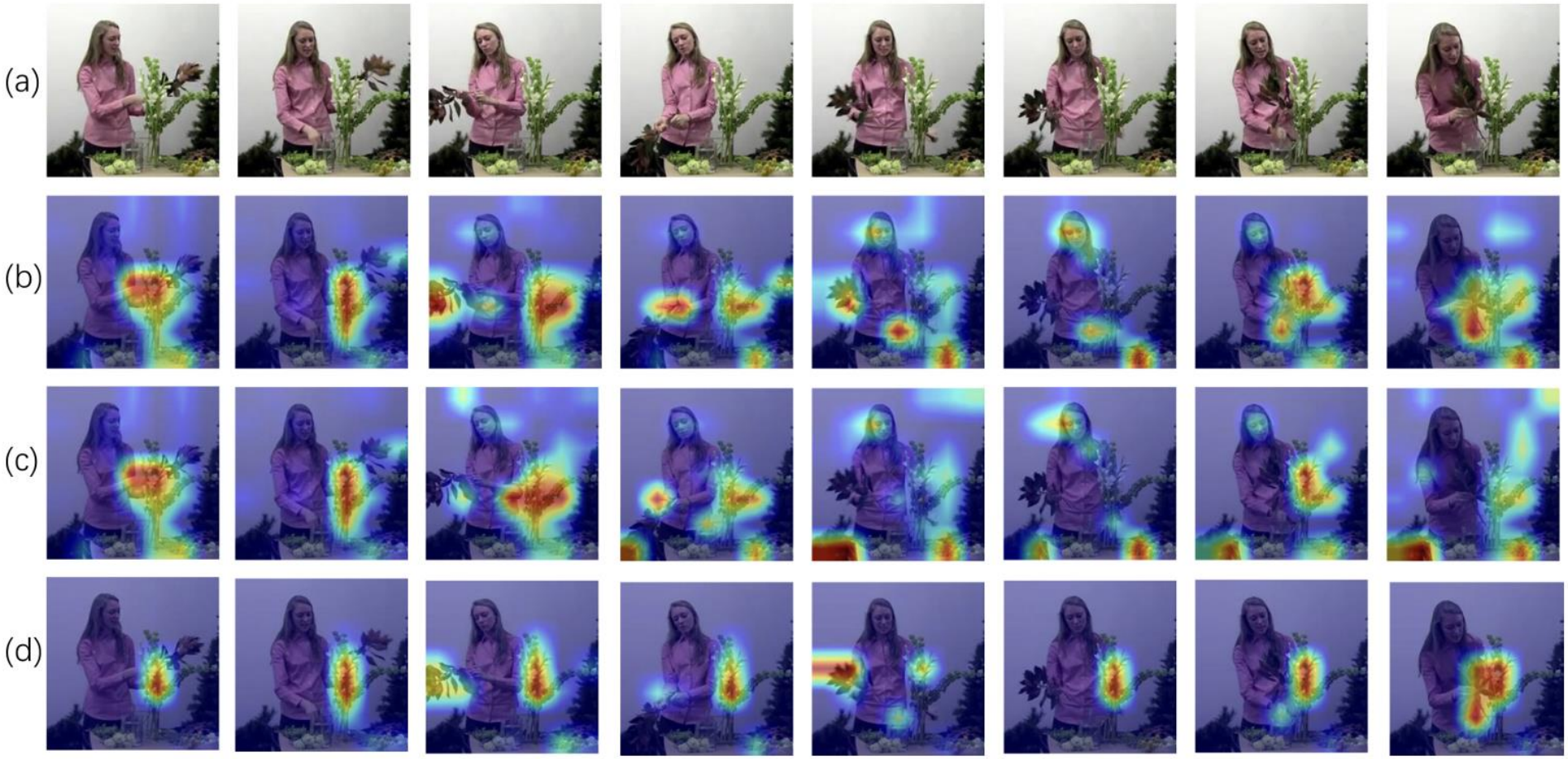
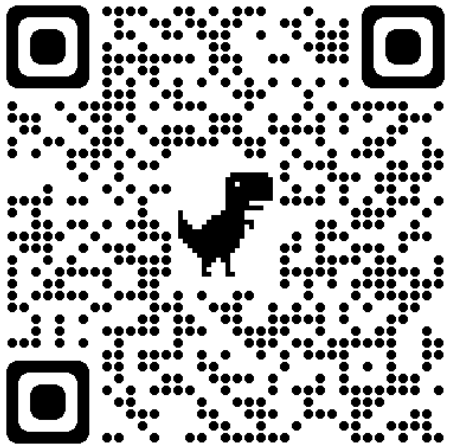


Figure 2. Visualization of the last feature map of different temporal modeling approaches on Kinetics-400. (a) refers to raw frames. (b), (c) and (d) refer to Divided ST Attention, ATA and ILA respectively.

Code & Models Available at
https://github.com/Francis-Rings/ILA

Thanks!



Microsoft Research

Carnegie Mellon University