Supplementary Material for Implicit Temporal Modeling with Learnable Alignment for Video Recognition

Shuyuan Tu[1,2] Qi Dai[3] Zuxuan Wu[1,2] * Zhi-Qi Cheng[4] Han Hu[3] Yu-Gang Jiang[1,2]

[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
[3]Microsoft Research Asia [4]Carnegie Mellon University

1. Implementation Details of ILA

Training Details. The experiments are conducted on 8 NVIDIA 32G V100 GPUs. The training configuration is listed in Table 1. It is worth noting that our sampling strategies for Kinetics-400 and Something-Something-V2 are different during the training phase. We implement the sparse sampling strategy on Kinetics-400. For SSv2, we uniformly sample the entire video at predefined temporal intervals without group division. In term of the training on Kinetics-400, the base learning rate indicates the learning rate of the original CLIP parameters. The learning rate for other additional parameters is 10× larger than the base learning rate. In term of the training on SSv2, we exclude the prompt branch and freeze the weights of CLIP visual branch for training stability. Thus the base learning rate is used for the rest parameters.

Convolution Module in SSv2. In SSv2, we increase the number of convolution layers in alignment. Particularly, two additional 3×3 convolution layers plus batch normalization and ReLU are added. In comparison to the original convolution module, it can bring 0.6% improvement on top-1 accuracy.

Table 1. Default implementation details of our method.

Training Configuration	Kinetics-400	Something-Something v2
Optimisation		
Optimizer	AdamW	
Optimizer betas	(0.9,0.98)	
Batch size	256	
Learning rate schedule	Cosine	
Learning warmup epochs	5	
Base learning rate	8e-6	5e-4
Minimal learning rate	8e-8	5e-6
training steps	50000	30000
Data augmentation		
RandomFlip	0.5	
MultiScaleCrop	(1, 0.875, 0.75, 0.66)	
ColorJitter	0.8	
GrayScale	0.2	
Label smoothing	0.1	
Mixup	0.8	
Cutmix	1.0	
Other regularisation		
Weight decay	0.003	0.01

2. Complexity of ILA

We analyze various temporal modeling methods (Spatial Attention [1], Joint Attention [1], Divided ST Attention [1], ATA [3], X-CLIP [2] and our proposed ILA) in terms of complexity, as shown in Table 2. The complexity of our alignment process is $O(Thwk^2d)$ due to the 2D convolution-based operations. The complexity of the whole ILA consists of the implicit alignment $O(Thwk^2d)$ and the spatial attention $O(Th^2w^2d)$. In terms of Joint Attention and Divided Spatiotemporal Attention, Joint Attention requires more computational memory since it takes all patches into consideration. Divided ST Attention only considers the temporal attention along the time axis. In terms of ATA, ATA is based on Hungarian Algorithm whose complexity is $O(N^3)$. In practice, the complexity of Hungarian matching is $O(Th^3w^3d)$ in video domain. Moreover, ATA requires additional temporal attention with complexity $O(T^2hwd)$. X-CLIP adopts a frame-level temporal attention with complexity $O(T^2d)$, which however obtains suboptimal result. We can observe that our proposed ILA can have better performance in low complexity.

3. Qualitative Analysis

In order to investigate the quality of three temporal modeling approaches (Divided ST Attention [1], ATA [3], and ILA), we visualize their intermediate and last feature maps respectively, as shown in Figure 1 and Figure 2. According to the illustrations, all three approaches capture the static semantic features, such as static flowers on the desk. More-

[*] Corresponding author

Table 2. Complexities of different methods, with results on Kinetics-400. T, h, w, d, and k refer to temporal size, spatial height of input, spatial width of input, channel depth of input, and kernel size of convolution, respectively.

Temporal Modeling	Complexity	Acc.(%)	FLOPs
Spatial Attention [1]	$O(Th^2w^2d)$	79.8	37G
Joint Attention [1]	$O(T^2h^2w^2d)$	80.4	71G
Divided ST Attention [1]	$O(T^2hwd + Th^2w^2d)$	80.6	58G
ATA [3]	$O(Th^3w^3d + T^2hwd + Th^2w^2d)$	81.0	60G
X-CLIP [2]	$O(T^2d + Th^2w^2d)$	80.4	39G
ILA	$O(Thwk^2d + Th^2w^2d)$	81.3	40G

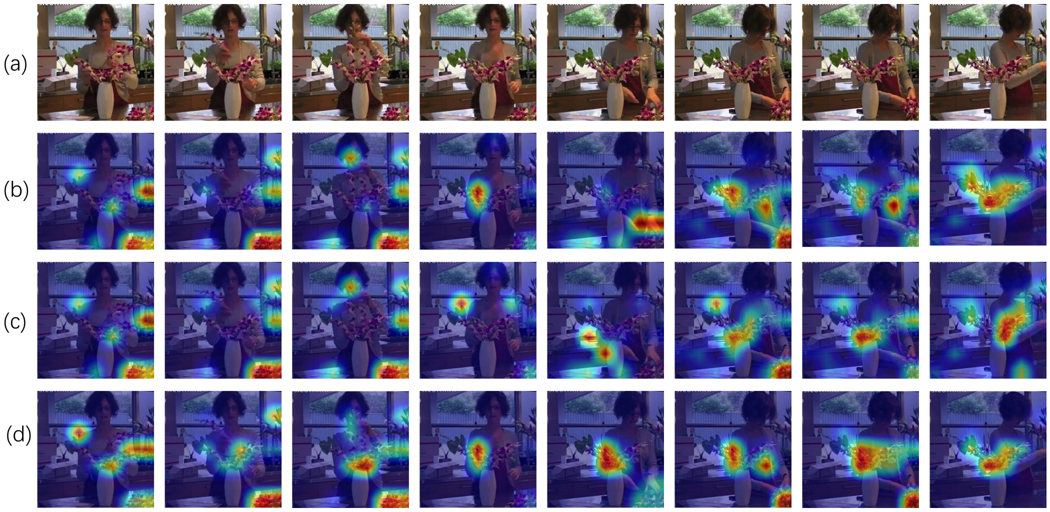

Figure 1. Visualization of intermediate feature map of different temporal modeling approaches on Kinetics-400. (a) refers to raw frames. (b), (c) and (d) refer to Divided ST Attention, ATA and ILA respectively.

over, our proposed ILA pays more attention to the action area of arranging flowers (*e.g.* the 5-th frame in the last row of Figure 2) instead of the static flowers on the desk. It indicates that our ILA can leverage the learnable mask to achieve implicit temporal modeling, focusing on the vital motion region. For divided ST attention, the model prefers to focus on static object instead of significant actions. While in ATA, the model attempts to concentrate on discontinuous regions with inaccurate positions. The plausible reason is that ATA utilizes patch movement-based alignment, which may destroys the continuity of semantic distribution.

4. Key differences between ILA and ATA

ATA adopts an explicit patch-level alignment with Hungarian matching, aiming at modeling temporal attention within aligned patches, which has poor efficiency due to the frame-by-frame serial alignment. Our ILA is fundamentally different as we utilize learnable masks to obtain implicit and coarse semantic-level alignment, which attempts to enhance favorable mutual information and can be performed in parallel with high efficiency.

It exits three fundamental different aspects. First, ATA can only align the collection of frames representations in serial frame-by-frame mode due to the limitation of KMA, while our ILA can utilize learnable masks to align semantical correspondences between two neighboring frames in parallel resulting in faster inference. Second, the complexity of ATA is $O(N^3)$ and ATA is unlearnable resulting in difficult optimization. Computational complexity of ILA is $O(N^2)$. Third, the core idea of ATA is to implement KMA algorithm to find out the optimal patch-level movement scheme capturing temporal correspondences, while the core idea of our ILA is to utilize specific masks to suppress irrelevant redundant information and enhance task-related mutual information among frames resulting in implicit alignment. Therefore, ATA still preserve the original irrelevant redundant information, while our ILA has the suppression of irrelevant redundant information due to principle of masks.

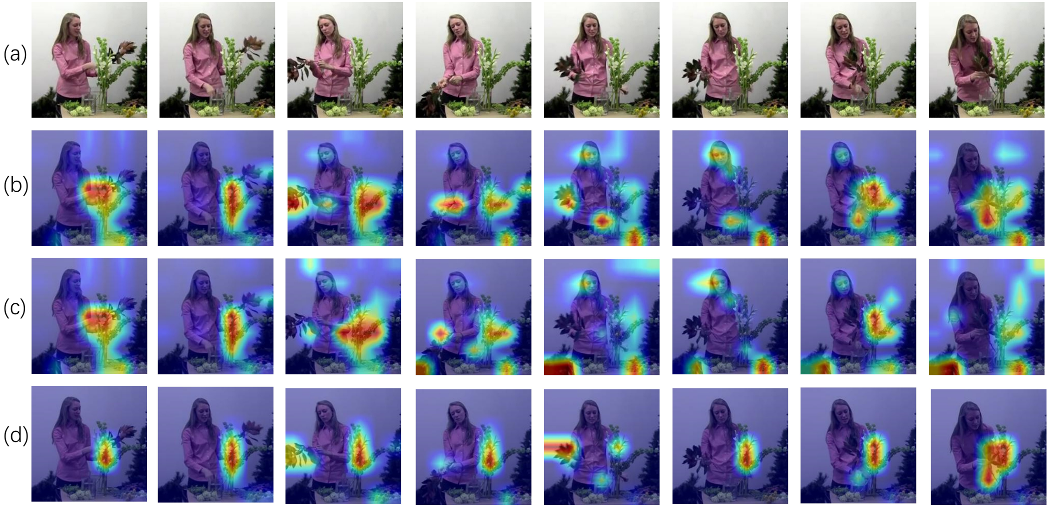

Figure 2. Visualization of the last feature map of different temporal modeling approaches on Kinetics-400. (a) refers to raw frames. (b), (c) and (d) refer to Divided ST Attention, ATA and ILA respectively.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 1, 2

[2] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, 2022. 1, 2

[3] Yizhou Zhao, Zhenyang Li, Xun Guo, and Yan Lu. Alignment-guided temporal attention for video action recognition. In *NeurIPS*, 2022. 1, 2