TAM: Temporal Adaptive Module for Video Recognition **Supplementary Material**

Zhaoyang Liu[1, 2] Limin Wang[1] ✉ Wayne Wu[2] Chen Qian[2] Tong Lu[1]
[1] State Key Lab for Novel Software Technology, Nanjing University, China
[2] SenseTime Research
zyliumy@gmail.com lmwang@nju.edu.cn {wuwenyan,qianchen}@sensetime.com lutong@nju.edu.cn

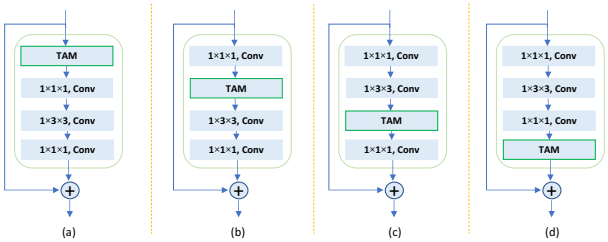

Figure 1. **The four styles of TA-Block.** The (b) is actually the model we used in the main text.

.

1. TAM in the different position.

We here introduce the four different exemplars of TANet. TANet-a, TANet-b, TANet-c, and TANet-d denote the TAM is inserted before the first convolution, after the first convolution, after the second convolution, and after the last convolution in the block, respectively. These four styles are graphically presented in Fig. 1 which were mentioned in main text.

2. Visualizations of Learned Kernel

In the supplementary material, we are ready to add more visualizations of distribution for *importance map* V in local branch and *video adaptive kernel* Θ in global branch. The $3\times1\times1$ convolution kernels in $\text{I3D}_{3\times1\times1}$ are also visualized to study their intentions in inference. To probe into the effects on learning kernels in the different stages, the visualized kernels are chosen in stage4_6b and stage5_3b, respectively. Some videos are randomly selected from Kinetics-400 and Sth-Sth V2 to show the diversities in different video datasets.

Firstly, as depicted in Fig. 2 and Fig. 3, We can observe that the distributions of importance map V in local branch are smoother than the kernel Θ in global branch, and local branch pays different attention to each video when modeling the temporal relations. Then, the kernel Θ in global branch performs the adaptive aggregation to learn the temporal diversities in videos. The visualized kernels in $\text{I3D}_{3\times1\times1}$ can make a direct comparison with the kernel Θ, and we find that the distributions of kernel in $\text{I3D}_{3\times1\times1}$ are extremely narrow whether on Kinetics-400 or on Sth-Sth V2. Finally, our learned kernels visualized in figures have exhibited the clear differences between two datasets (Kinetics-400 vs. Sth-Sth V2). This fact is in line with our prior knowledge that there is an obvious domain shift between two datasets. The Kinetics-400 mainly focuses on appearance and Sth-Sth V2 is a motion dominated dataset. However, this point can not be easily summarized from the kernels in $\text{I3D}_{3\times1\times1}$, because the overall distributions of kernels in $\text{I3D}_{3\times1\times1}$ on two datasets show minor differences.

Generally, the diversities in our learned kernels have demonstrated that the diversities are indeed existing in videos, and it is reasonable to learn spatio-temporal representation in an adaptive scheme. These findings are again in line with our motivation claimed in the paper.

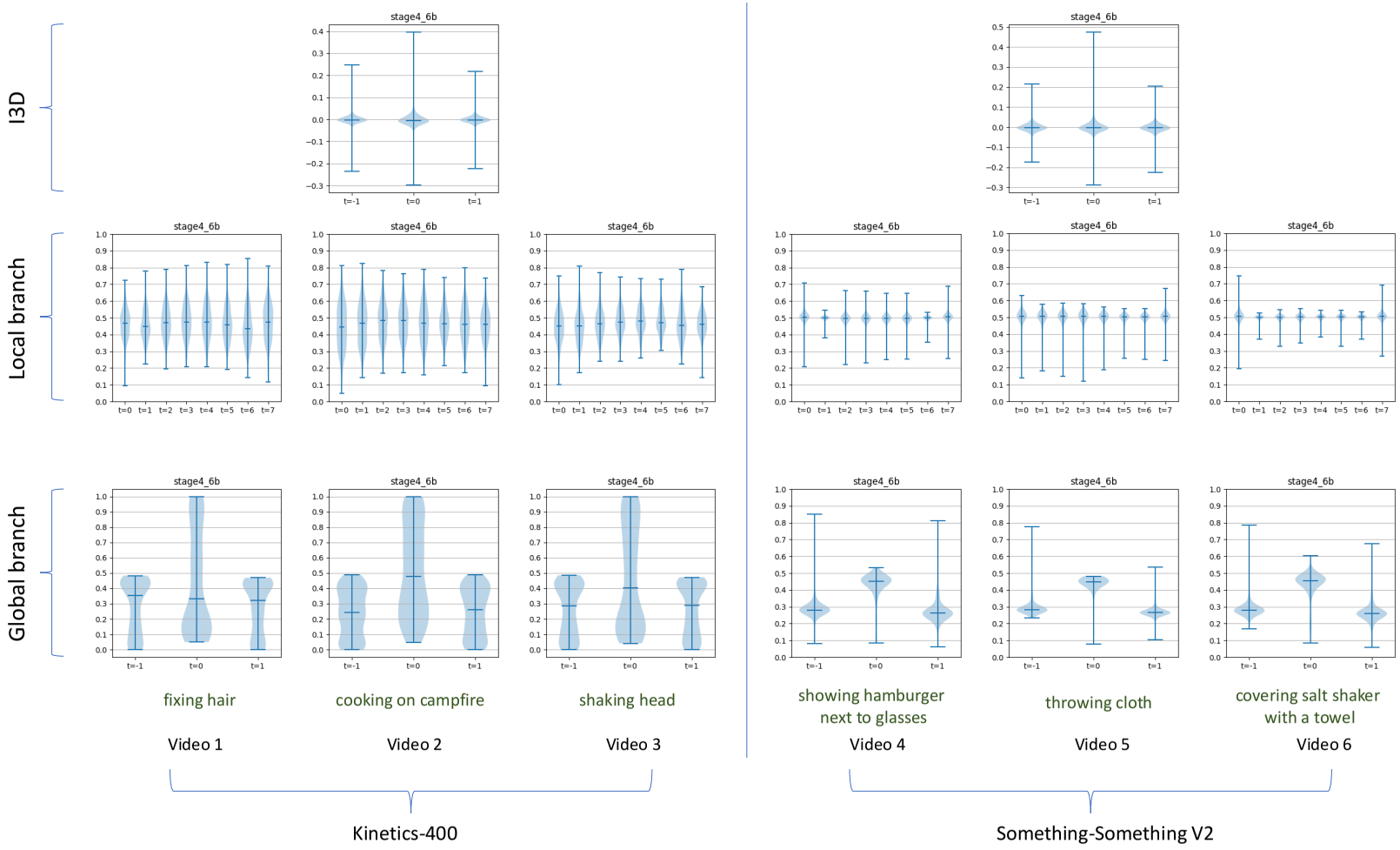

Figure 2. The distribution of learned kernel V and Θ in the stage4_6b.

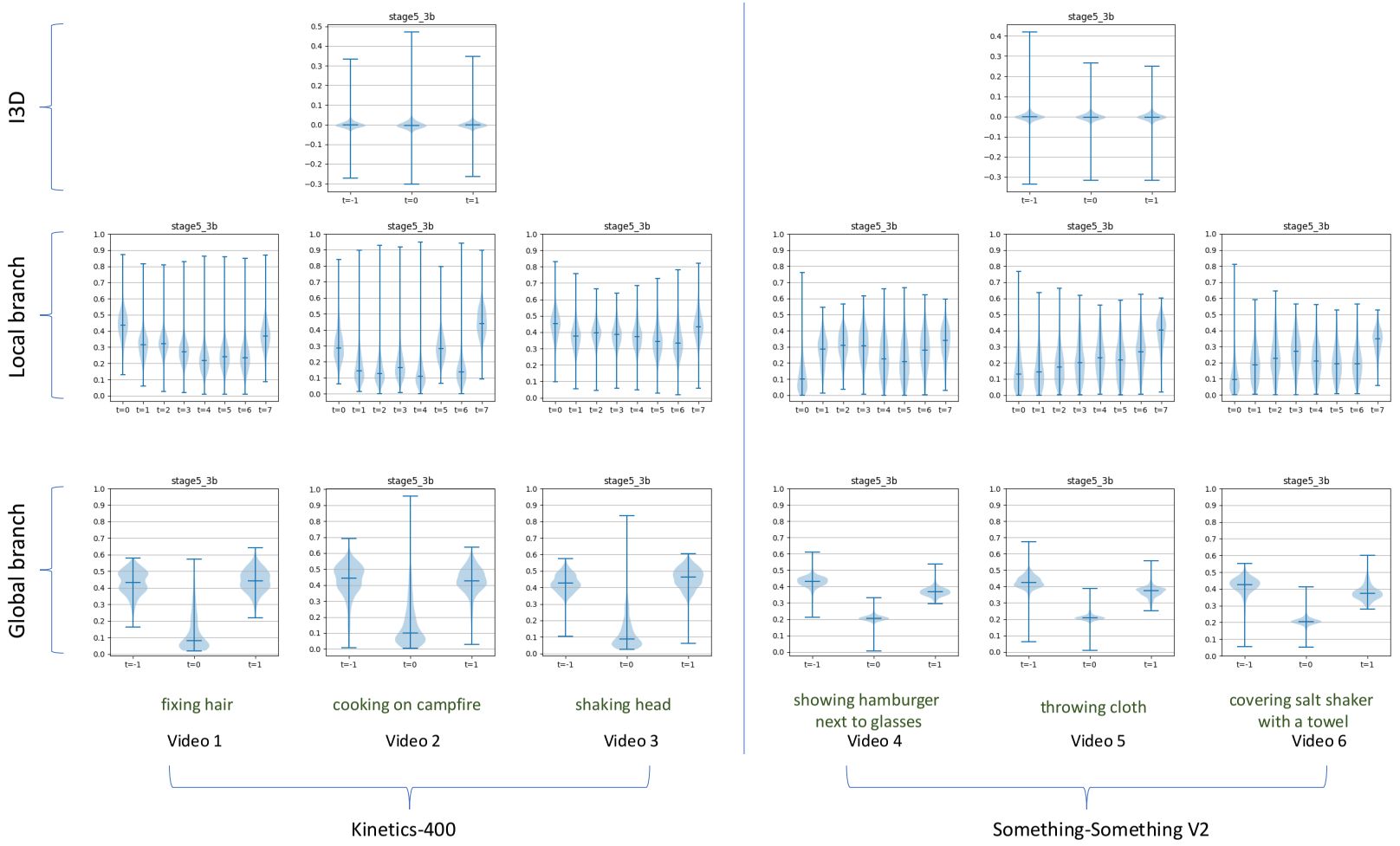

Figure 3. The distribution of learned kernel V and Θ in the stage5_3b.