Appendix

This appendix details the architectures of the other calibrators in the ablation study. The results by inserting GCs on different residual stages are provided. Visualization examples of class activation maps on Something-Something V1 dataset are given. Gating weight distributions of different calibrators on Something-Something V1 and Kinetics-400 datasets are shown. More experimental results on datasets EGTEA Gaze+, Diving48 and Basketball-8&Soccer-10 are presented.

A. Architectures of SE3D, GE3D-G/C and S3D-G

Figure 7 shows the detailed architectures of SE3D [16], GE3D-G/C [15] and S3D-G [50]. In the implementation, the calibrators are also densely inserted into the TSN and TSM backbones. We report their performances on Something-Something V1 (see ablation study).

B. Results on different residual stages

Here, we also investigate which residual stage(s) to add the GC block in ResNet-50 using TSN as backbone. Table 6 compares both a single and multiple GC blocks added to different stages of ResNet. Overall, a GC block significantly improve the performance (19.7%) of the original TSN. Specifically, the results on the deeper Res3 and Res4 are better than those on Res1 and Res2, and the improvement reaches the highest of 45.9% on Res3. The possible explanation is that deeper layers can provide more high-level features which are precise for context modeling. But the last residual block Res4 has a small spatial size (7×7) and it limits the precision of spatial information. Moreover, densely incorporating multiple GC blocks into ResNet exhibits better performance than the single version.

Model	Residual Blocks				Top1 (%)
	Res1	Res2	Res3	Res4	
TSN	✔	✗	✗	✗	42.2
	✗	✔	✗	✗	43.7
	✗	✗	✔	✗	45.9
	✗	✗	✗	✔	44.2
	✔	✔	✗	✗	44.3
	✔	✔	✔	✗	46.7
	✔	✔	✔	✔	**47.9**

Table 6. Performance comparison of adding a single or more GC modules to different stages of ResNet-50 on Something-Something V1 dataset.

C. Visualization

We provide some visualization examples of class activation maps using TSN with ECal-G/T/S/L and GC to clearly show the vital parts they learn. The sampled visualization results are shown in Figure 8. In the implementation, we use 8-frame center crops as input and the Grad-CAM [35] technique to obtain the heatmaps. These videos are selected from Something-Something V1 dataset. The categories in Something-Something dataset emphasize not only the short-term interactions between objects (e.g., "*Wiping something off something*", "*Closing something*" and "*Bending something*") but also the long-range dependencies (e.g., "*Pretending to do*", "*Failing to do*" and "*Doing something so that it is to be*"). Based on the visualization results, the GC-TSN, which aggregates the four context calibrators ECal-G/T/S/L in parallel, indeed yields more reasonable class activation maps than the original TSN and its variants ECal-TSNs with single-context.

D. Gating weight distribution

We calculate the mean of channel weights for each ECal. Fig.9 shows the distributions of weights for 17 Something-Something V1 categories involving space-time dynamics and 17 Kinetics-400 categories with less motion variations. The results shows that GC can *inherently* learn the importance of ECals by assigning higher weights for ECal-L/T on Something-Something V1 and ECal-S on Kinetics-400.

E. Results on other datasets

In the Appendix, we additionally provide the results on EGTEA Gaze+ [25], Diving48 [24] and Basketball-8&Soccer-10 [13] datasets. Particularly, the **EGTEA Gaze+** dataset offers first-person videos, containing 106 non-scripted daily activities occurred in the kitchen. The **Diving48** dataset consists of 48 unambiguous dive sequence, which requires modeling long-term temporal dynamics. The **Basketball-8&Soccer-10** datasets are composed of two datasets for sport classification: Basketball-8 with 8 group activities and Soccer-10 with 10 group activities.

EGTEA Gaze+. Table 7 shows the results of different methods on the first-vision EGTEA Gaze+ dataset. Similar observations as on other datasets, GC-Nets contribute significant improvements to their backbones when dealing with the short-term kitchen activities. Specifically, we observe 0.4%-6.2% performance increase among the results on the three train/validation splits. This demonstrates that our proposed GC module is generic for short-term temporal modeling.

Diving48. This dataset is also a "temporally-hevay" dataset. Since this newly released dataset version has been thoroughly revised for wrong labels, we re-run the backbones of TSN, TSM, GST and TDN using 16 frames (center crop) as input. Table 8 shows the performance comparison. Our GC-TDN achieves the highest of 87.6% Top-1 accuracy, which increases its backbone TDN by 3.0 percent-

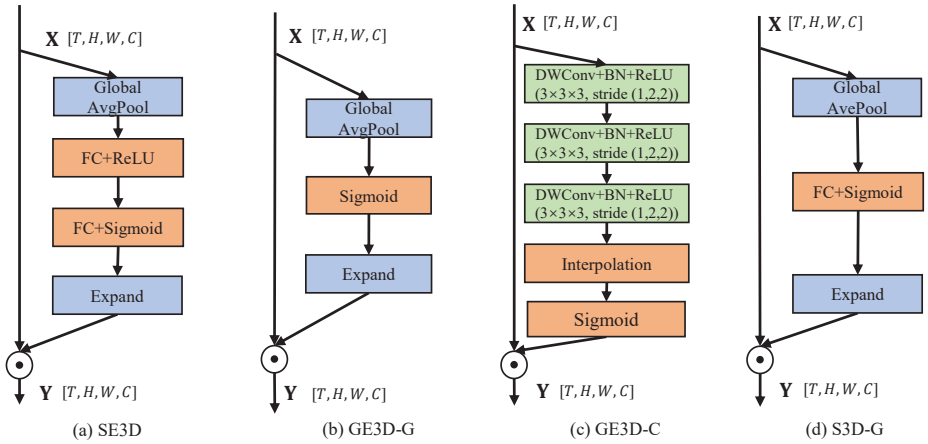

Figure 7. Illustration of architectures of SE3D, GE3D-G/C and S3D-G.

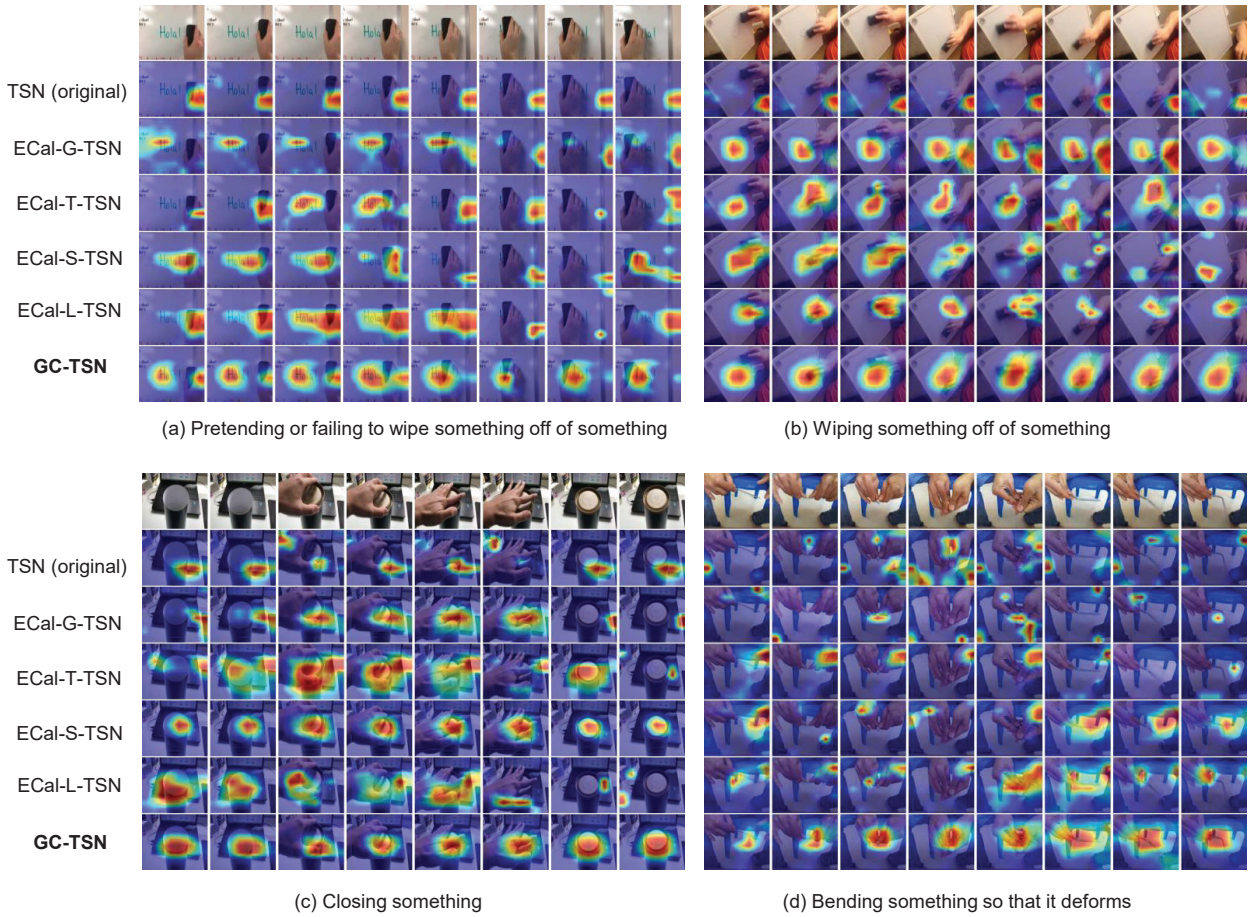

Figure 8. Visualization of class activation maps on sample video clips from the Something-Something V1 dataset. The first row presents original frames and each of the other rows presents the visualization results of a model.

age and is even better than the Transformer-based VIMPAC (85.5%).

Basketball-8&Soccer-10. The fine-grained sport highlights [13] take place with various local interactions among offensive players, defensive players and other objects, and

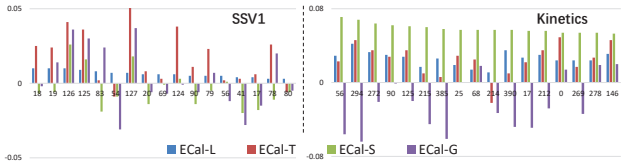

Figure 9. Means of gating weights (before Sigmoid) of ECal-L/T/S/G (GC-TSN) on Something-Something V1 and Kinetics-400. X-axis: category index.

Method	Backbone	#Frame	Split1	Split2	Split3
I3D-2stream [25]	ResNet34	24	55.8	53.1	53.6
R34-2stream [38]	ResNet34	25	62.2	61.5	58.6
SAP [48]	ResNet50	64	64.1	62.1	62.0
TSN (our impl.)	ResNet50	8	61.6	58.5	55.2
GST (our impl.)	ResNet50	8	63.3	61.2	59.2
TSM (our impl.)	ResNet50	8	63.5	62.8	59.5
TDN (our impl.)	ResNet50	8	63.9	60.8	60.2
GC-TSN	ResNet50	8	66.4	64.6	61.4
GC-GST	ResNet50	8	65.5	61.6	60.6
GC-TSM	ResNet50	8	**66.5**	**66.1**	**62.6**
GC-TDN	ResNet50	8	65.0	61.8	61.0

Table 7. Performance (Top-1 accuracy %) comparison on EGTEA Gaze+ dataset using the official train/validation split 1/2/3.

Method	**Backbone**	**#Frame**	**Top-1**
SlowFast,16×8 from [2]	ResNet101	64+16	77.6
TimeSformer-HR [2]	Transformer	16	78.0
TimeSformer-L [2]	Transformer	96	81.0
VIMPAC [39]	Transformer	32	85.5
TSN (our impl.)	ResNet50	16	79.0
GST (our impl.)	ResNet50	16	78.9
TSM (our impl.)	ResNet50	16	83.2
TDN (our impl.)	ResNet50	16	84.6
GC-TSN	ResNet50	16	86.8
GC-GST	ResNet50	16	82.5
GC-TSM	ResNet50	16	87.2
GC-TDN	ResNet50	16	**87.6**

Table 8. Performance (Top-1 accuracy %) comparison on the updated Diving48 dataset using the train/validation split v2.

Model	Pretrain	Basketball-8		Soccer-10	
		Validation	Test	Validation	Test
I3D [3]	ImageNet	—	75.4	—	88.3
Nonlocal-I3D [47]	ImageNet	—	77.2	—	88.3
GST [30] (our impl.)	ImageNet	78.8	75.8	87.9	87.6
GC-GST	ImageNet	81.8	78.4	88.3	88.5
TSN [56]	ImageNet	71.9	68.5	86.2	83.7
GC-TSN	ImageNet	81.8	78.8	89.5	88.9
TSM [26]	Kinetics	77.6	73.3	88.7	87.9
+CBA-QSA [13]	Kinetics	—	78.5	—	89.3
TSM-NLN [26]	Kinetics	—	76.2	—	88.2
+CBA-QSA [13]	Kinetics	—	79.5	—	88.7
GC-TSM	Kinetics	**83.8**	**80.2**	**90.3**	**89.4**
TDN [45]	ImageNet	80.3	78.4	86.9	86.1
GC-TDN	ImageNet	83.0	79.7	87.7	87.1

Table 9. Comparison of performance (Top1 accuracy %) of diffent methods with 8 frames × 1 clip input on Sport Highlights datasets. The results of I3D, Nonlocal-I3D and TSM-NLN are cited from [13].

the local interactions could be either short-term (e.g., the "Blocked shot" highlight in Basketball) or long-term (e.g., the "Layup" highlight in Basketball and the "Shooting and goalkeeping" highlight in Soccer). This indicates that both the global and local axial contexts can benefit the sport activity recognition. No surprisingly, as shown in Table 9, all GC-Nets, i.e., GC-TSN, GC-GST, GC-TSM and GC-TDN, consistently boost their base networks, e.g., 68.5%→78.8% for TSN, 75.8%→78.4% for GST, 73.3%→80.2% for TSM and 78.4%→79.7% for GC-TDN. Compared to the similar feature calibration works [47] and [13], our GC module performs best with the same backbone TSM, obtaining the highest Top1 accuracy 80.2%/89.4% on Basketball-8/Soccer-10.

F. Video annotations

Table 10 lists the selected 28 activity categories from Something-Something V1 dataset used in Figure 6.

ID	Name
label-1	Approaching something with your camera
label-15	Folding something
label-26	Lifting a surface with something on it but not enough for it to slide down
label-33	Moving away from something with your camera
label-37	Moving something and something away from each other
label-38	Moving something and something closer to each other
label-44	Moving something down
label-46	Moving something up
label-51	Plugging something into something but pulling it right out as you remove your hand
label-58	Poking something so that it falls over
label-61	Pouring something into something until it overflows
label-70	Pretending to poke something
label-73	Pretending to put something into something
label-77	Pretending to put something underneath something
label-78	Pretending to scoop something up with something
label-81	Pretending to squeeze something
label-84	Pretending to throw something
label-88	Pulling something from right to left
label-91	Pulling two ends of something but nothing happens
label-106	Putting something in front of something
label-112	Putting something onto a slanted surface but it doesn't glide down
label-118	Putting something that cannot actually stand upright upright on the table, so it falls on its side
label-132	Something being deflected from something
label-152	Throwing something
label-156	Throwing something onto a surface
label-167	Turning the camera left while filming something
label-168	Turning the camera right while filming something
label-173	Unfolding something

Table 10. Categories of the selected activities in Figure 6.