Supplementary Material of
Prune Spatio-temporal Tokens by Semantic-aware Temporal Accumulation

Shuangrui Ding[1,3] Peisen Zhao[2] Xiaopeng Zhang[2] Rui Qian[3] Hongkai Xiong[1] Qi Tian[2]
[1]Shanghai Jiao Tong University [2]Huawei Cloud [3]The Chinese University of Hong Kong
{dsr1212, xionghongkai}@sjtu.edu.cn qr021@ie.cuhk.edu.hk
{pszhao93, zxphistory}@gmail.com tian.qi1@huawei.com

A. More Experimental Results

Training details. We also train the ViT-B with our proposed STA. We adopt dense sampling [5, 2] on K400. We sample 16 consecutive frames with the stride of 4. The resolution is 224×224. We perform RandAug augmentation (9, 0.5) [1], label smoothing (0.1) [4], mixup (0.8) [7], cutmix (1.0) [6], and random horizontal flip (0.5). In addition, we adopt the repeated augmentation [3]. With DeepSpeed [1], We use the linearly scale scheme to ensure effective parameter updates across different batch sizes during training, *i.e.,* lr = base learning rate $\times$ batch size / 256. Specifically, we use the AdamW optimizer with a base learning rate of $1e{-}3$ and weight decay of 0.05. Beside, using a cosine decay learning rate scheduler and 5 epochs of linear warm-up, we finetune the model for 100 epochs with a total batch size of 128 on 4 nodes of 8 Tesla V100 GPUs.

Training results. Besides speeding up the inference of off-the-shelf backbones, our algorithm also has the potential to expedite training. We report the training hours for ViT-Base in Table 1. STA cuts the training time in half. Without modifying the training recipe, the trained model only drops 0.6 % in Top-1 accuracy. We believe that STA would be more effective to maintain the performance when training deeper backbones. We leave it as the future work.

Number of views. To analyze the impact of the number of test clips on our method, we conduct an experiment by varying the number of clips and comparing the results with the baseline ViT-L model. In Table. 2, we show that the relative performance drop remains constant at approximately 0.1% regardless of the number of views, when the drop number is set to $r_1 = 64$. Furthermore, when using a lower value of $r_1 = 48$, there is no significant decrease in performance compared to the baseline.

[1] https://github.com/microsoft/DeepSpeed

Model	clips/s	Training time	Top-1
ViT-B	53	28 hrs	81.2
STA^{48}-ViT-B	96	15 hrs	80.6

Table 1. Comparison on training time on Kinetics-400. We measure training time on 4 nodes of 8 V100.

Views	Drop Number r_1			
	0	48	64	80
2×3	83.36	83.21	83.09	82.56
4×3	85.10	85.00	84.85	84.35
6×3	85.05	85.07	84.84	84.59
8×3	84.91	84.93	84.80	84.43
16×3	84.91	84.97	84.89	84.48

Table 2. Ablation on the temporal views of test clips.

# of STA	GFLOPs	Top-1	Location	GFLOPs	Top-1
2	302	84.5	1,9,17	308	85.0
3	308	85.0	3,11,19	339	85.0
4	305	84.8	5,13,21	370	85.1

Table 3. Ablation on the number of STA blocks and insert location.

Number of STA blocks and insert location We devise two extra ablation studies shown in Table 3. Our experiments demonstrate that incorporating 3 progressive blocks at the very first beginning achieves an optimal trade-off. This approach allows for preferable computation while delivering maximal performance.

B. More Visualization

We provide more visualization for our STA on K400 in Figure 1 and SSV2 in Figure 2, which display image patches that correspond to the tokens retained after three stages of pruning. We observe that the pruning results align well with our objective of preserving detail-rich tokens and resisting temporal redundancy. Specifically, upon examining the guitar-playing sequence in Figure 1, STA accurately preserves two partially visible guitars on the wall. Addi-

tionally, the dropped tokens shown in Figure 2 at different timestamps are distributed unevenly, preserving the diversity of the video content.

References

[1] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 1

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1

[3] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020. 1

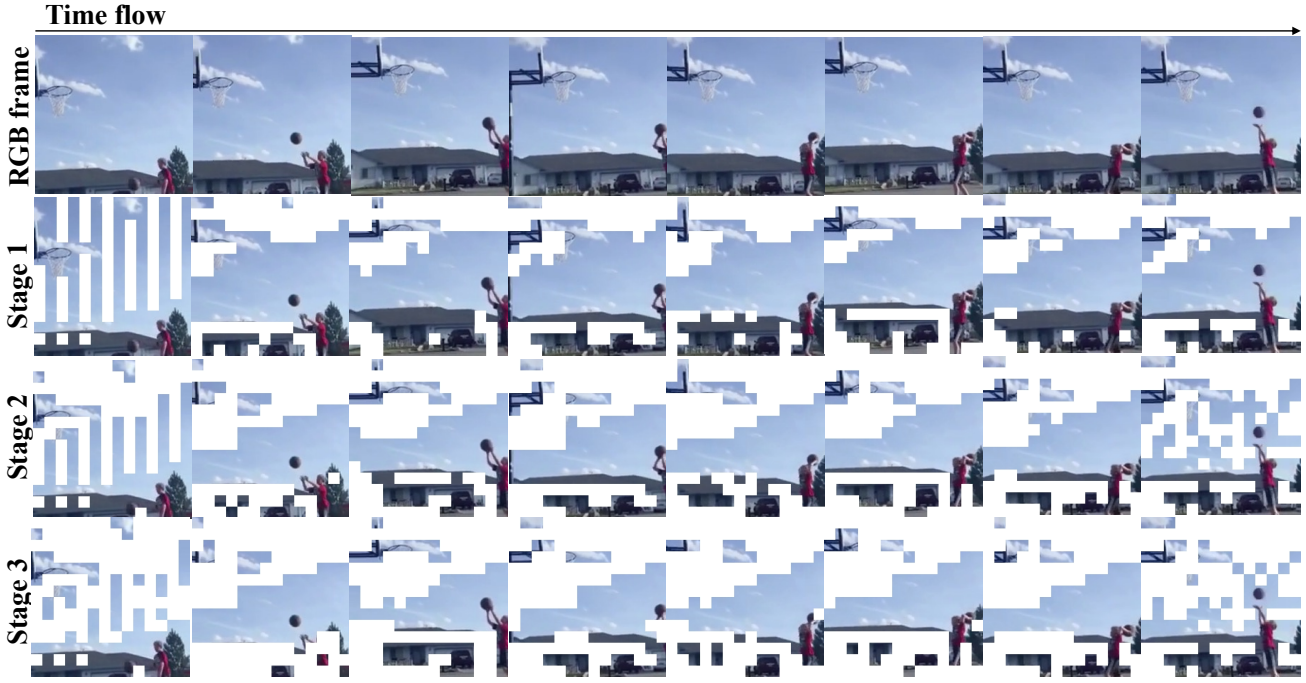
[4] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 1

[5] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 1

[6] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 1

[7] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 1


Time flow
RGB frame
Stage 1
Stage 2
Stage 3


Time flow
RGB frame
Stage 1
Stage 2
Stage 3

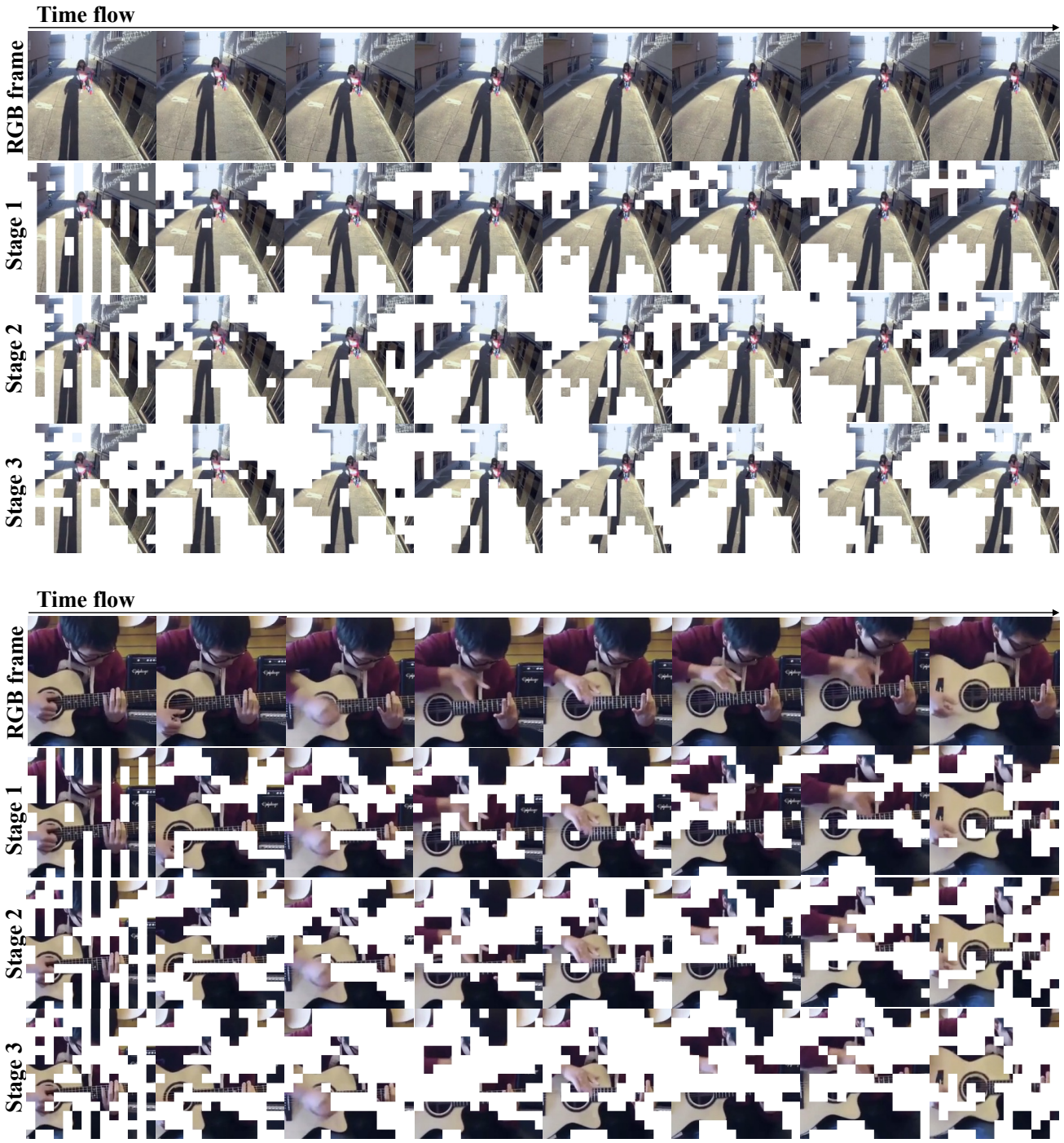
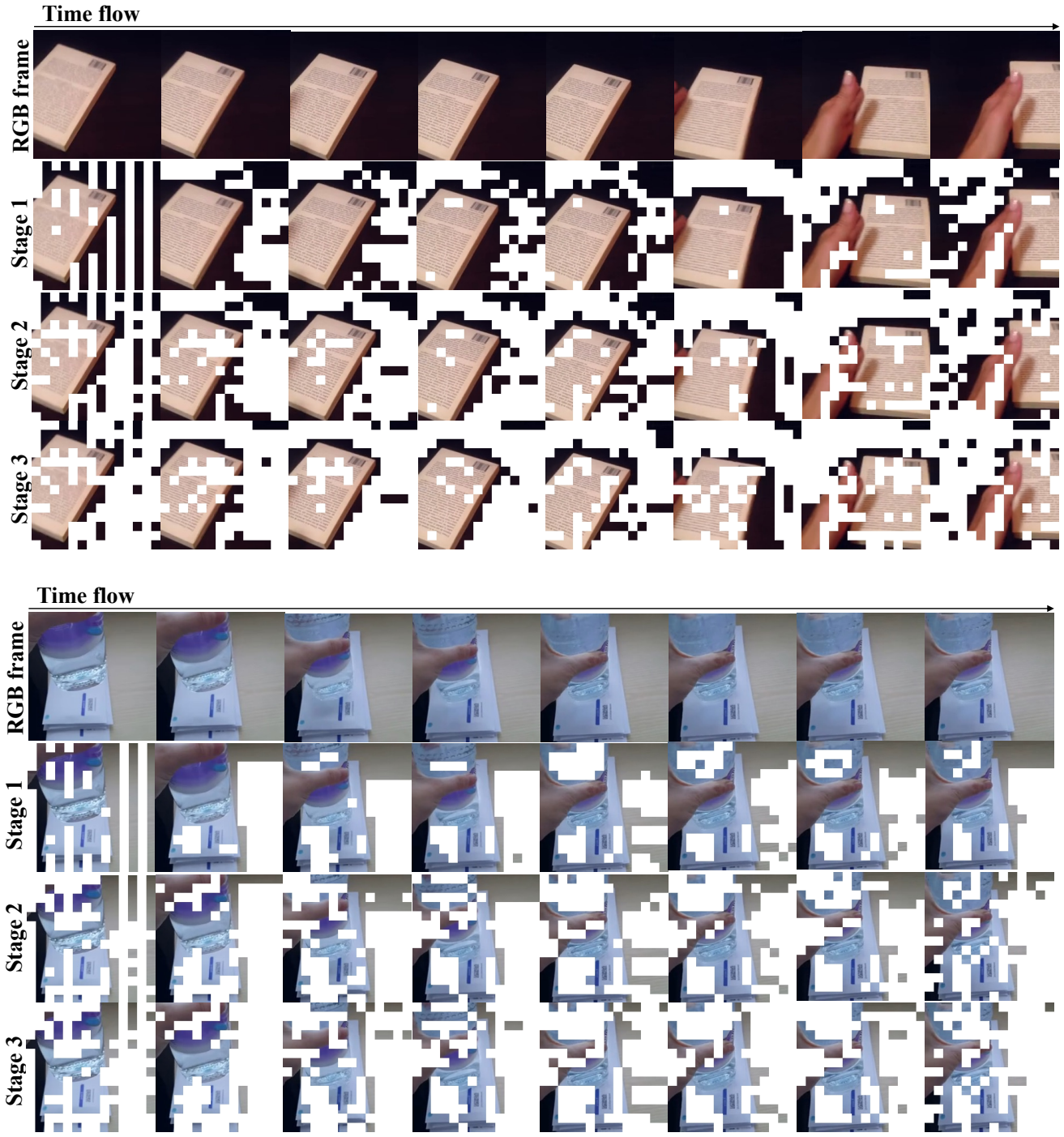


Figure 1. Visualization of our STA strategy on K400.


Time flow
RGB frame
Stage 1
Stage 2
Stage 3
Time flow
RGB frame
Stage 1
Stage 2
Stage 3

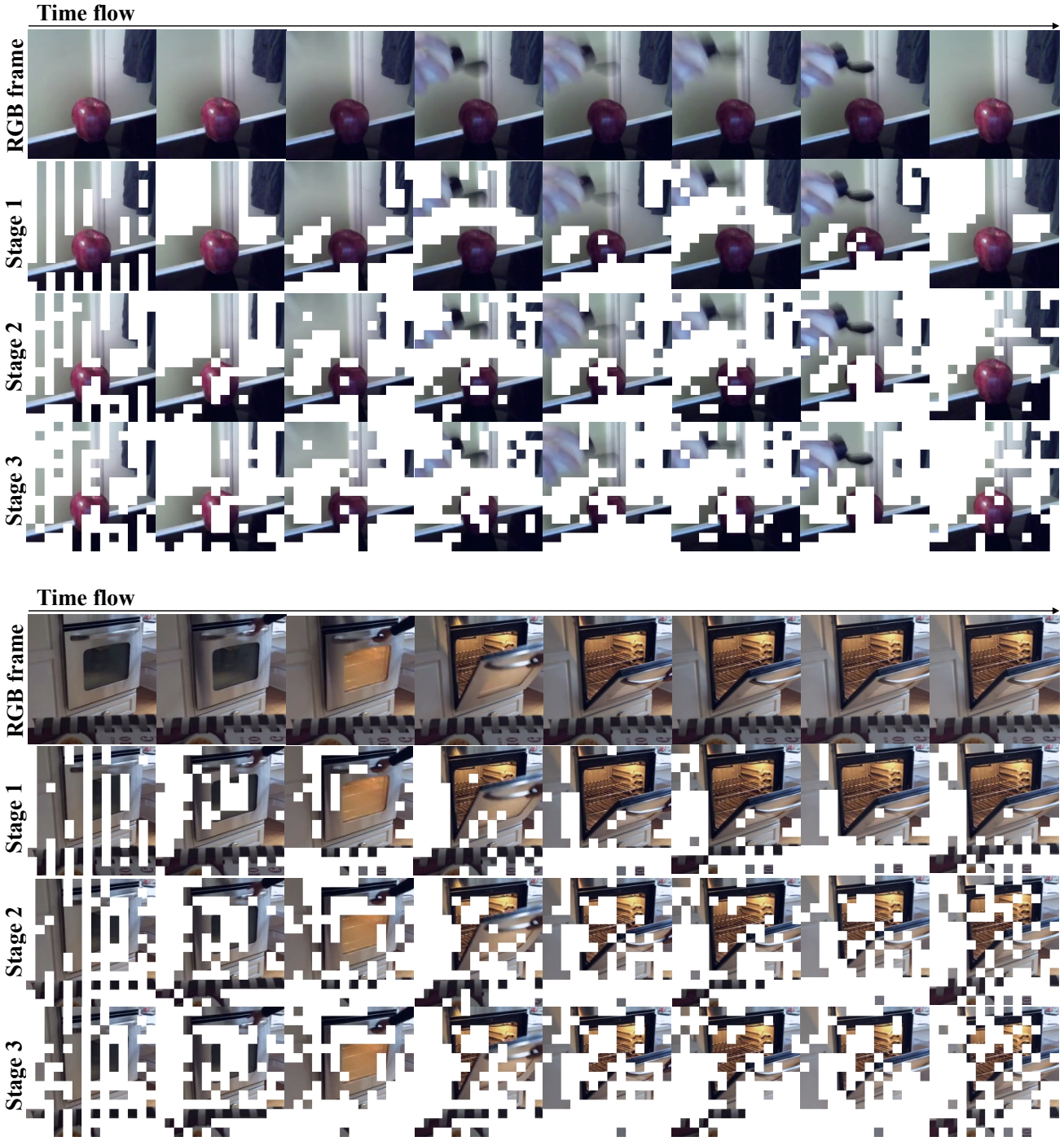


Figure 2. Visualization of our STA strategy on SSV2.