A Experimental Details

In this section, we present implementation details for the proposed greedy K-center sampling and the frame-based sampling used in our experiments.

A.1 Pre-processing Video Clips

For a fair comparison, we carefully extract a clip out of raw video prior to performing K-centered sampling in order to set the physical time range covered by our method and the frame-based sampling method to be equal. Since the precise procedures for pre-processing clips from raw videos in Kinetics and Something-Something datasets are different, we describe them separately.

Kinetics dataset. We transcode all raw videos to play at 30 frames-per-second refresh rates beforehand of any sampling, then perform either the frame-based sampling or the K-centered search executed as follows. For the frame-based sampling, we take the equally-spaced F frames at the sampling period of S frames (*i.e.*, the grid-frame-sampling strategy). On the other hand, for the K-centered search, we perform the greedy K-centered sampling given a clip composed of an equally-spaced set of $R \cdot F$ frames at the sampling period of $\frac{S}{R}$ frames. That is, we pick a set of frames lying in the equal time range covered by the frame-based sampling at R times denser rates, then apply the K-centered search within those $R \cdot F$ frames.

Something-Something dataset. We transcode all raw videos to play at 12 frames-per-second refresh rates beforehand of any sampling, then perform either the frame-based sampling or the K-centered search executed as follows. For the frame-based sampling, we split a video into equally-sized, non-overlapping F segments and then select a frame from each segment; we randomly select a frame in training time and the frame at the center of a segment in test time. For the K-centered search, we similarly split a video into equally-sized non-overlapping $R \cdot F$ segments, select one frame per segment, then perform the K-centered search within the selected $R \cdot F$ frames.

Table 7. Choices for the pre-processing clips used for our K-centered sampling with ViT, TimeSformer, and XViT under Something-Something v2 (SSv2) and Kinetics datasets.

Model	SSv2			Kinetics		
	S	F	R	S	F	R
ViT [11]	-	8	4	16	8	4
TimeSformer [3]	-	8	4	32	8	4
XViT [4]	-	16	2	16	8	4

Table 8. Comparison of our K-centered sampling and frame-based sampling methods in Kinetics-200 (K200) dataset. We report Top-1 and Top-5 action classification accuracies (%) on the validation set of the dataset. Computational budgets for forwarding a video sample are equal between the same model. For a fair comparison, we unify all models to use the same ImageNet-1k pre-training and evaluation protocols. The bold denotes the best result.

Model	Sampling	K200	
		Top-1	Top-5
ViT [11]	Frame-based	81.72	**95.44**
	K-centered (ours)	**82.90**	95.42
TimeSformer [3]	Frame-based	84.12	**96.36**
	K-centered (ours)	**84.24**	96.10
Motionformer [35]	Frame-based	78.40	92.86
	K-centered (ours)	**80.26**	**94.46**
XViT [4]	Frame-based	79.02	94.32
	K-centered (ours)	**80.06**	**94.44**

Table 9. Comparison of our K-centered sampling applied to long-range 16-frames (L) variants of Motionformer and TimeSformer experimented in Kinetics-200 (K200) dataset. We report Top-1 and Top-5 classification accuracies (%).

Model	Sampling	K200	
		Top-1	Top-5
ViT (L) [11]	Frame-based	84.92	96.94
	K-centered (ours)	**85.54**	**96.98**
TimeSformer (L) [3]	Frame-based	84.94	96.04
	K-centered (ours)	**85.32**	**96.32**

A.2 Hyperparmeter Choice

Depending on combinations of datasets (Kinetics and Something-Something) and models (ViT [11], TimeSformer [3], and XViT [4]), we select hyperparameters S, F, and R so that all models evaluated in a dataset perform similarly in terms of their classification accuracies. The detailed choices for the hyperparameters are reported in Table 7. Note that for Motionformer [35], the original frame-based baseline utilizes spatiotemporal voxels, instead of patches, as inputs where it is non-trivial to derive an equivalent set of the hyperparameters for our K-centered sampling. Therefore, we simply choose to follow the model's default setting of sampling 16 frames at period 4 for the baseline and $\{S = 16, F = 8, R = 4\}$ for our K-centered sampling model.

Table 10. Effect of hybrid sampling, *i.e.*, sampling both frames and patches, on ViT under Something-Something v2 dataset. Note that the total sampled areas are equal. We report Top-1 and Top-5 classification accuracies (%). The bold indicates the best result.

Metric	Hybrid Sampling							
	0	1	2	3	4	5	6	7
Top-1	63.38	**64.73**	63.95	63.83	64.03	63.88	63.62	63.66
Top-5	88.37	88.95	88.46	88.89	**89.03**	88.83	88.95	88.75

Table 11. Effect of hybrid sampling on (a) TimeSformer and (b) Motionformer under Kinetics-200 dataset. Note that the total sampled areas are equal. We report Top-1 and Top-5 classification accuracies (%). The bold indicates the best result.

Metric	Hybrid Sampling							
	0	1	2	3	4	5	6	7
Top-1	83.56	83.50	83.84	84.04	**84.24**	83.52	83.84	83.76
Top-5	95.92	96.32	96.04	96.08	**96.10**	95.92	96.06	95.92

(a) TimeSformer

Metric	Hybrid Sampling							
	0	1	2	3	4	5	6	7
Top-1	**80.26**	79.50	80.08	79.6	80.16	79.54	79.84	79.74
Top-5	94.46	**94.66**	94.28	93.9	93.46	93.98	94.46	94.16

(b) Motionformer

B Additional Experiments

B.1 Kinetics-200 dataset

We present the experimental results in Table 8 for comparing our *K*-centered sampling with the conventional frame-based sampling strategy in Kinetics-200 dataset. Overall, trends are consistent with those on Kinetics-400 in Table 2. The results are expected, as Kinetics-200 is a randomly-sampled subset of Kinetics-400.

B.2 Long-range model variants

We present the experimental results for the long-range 16-frames variants (L) of ViT [11] and TimeSformer [3] in Kinetics-200 dataset in Table 9. Compared to the baselines that process 8 frames in 224×224 resolution, the long-range models process 16 frames in the same resolution; thus, the number of tokens is doubled. Overall, trends are consistent with the models in Table 2. This experiment reveals that our method can be scaled up to a longer time span set-up.

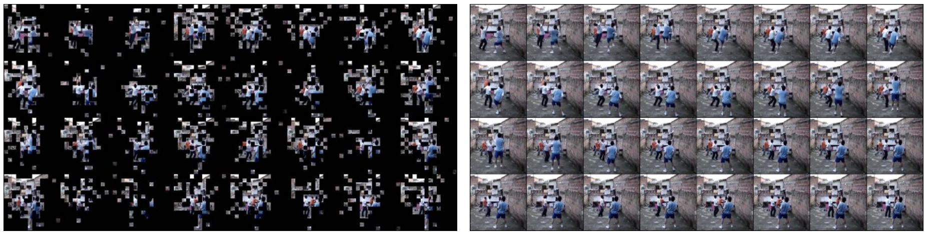


Fig. 7. An example of K-centered patch sampling in a complex background sample of Kinetics-400 dataset . The highlighted region in the left video denotes the sampled patches, and the right video indicates the original video. Our method frequently samples the patch with moving objects while less frequently samples the redundant patches such as backgrounds.

B.3 Effect of the hybrid sampling

We present additional experimental results for the effect of hybrid sampling tested with ViT [11] model under Something-Something v2 dataset in Table 10 and those tested with video transformers under Kinetics-200 dataset in Table 11. Specifically, we present the results for TimeSformer [3] in Table 11a and for Motionformer [35] in Table 11b.

C The FLOPs and Time Cost for K-centered Sampling

Executing the structure-aware K-center search (Sec. 3.3) to sample 1,568 patches from the source of 6,272 patches demands 0.94 GFLOPs, while GFLOPs for ViT [11], TimeSformer [3], Motionformer [35] and XViT [4] in Table 2 are 180, 196, 369 and 142, respectively. As for the time cost, The average (1,000 trials) extra latency of our Python implementation is 7.03ms, within the total forward time of 64.4ms for Motionformer [35]. Note that the latency can be more optimized (e.g., implementing in C).

D Discussions on the Complex Background

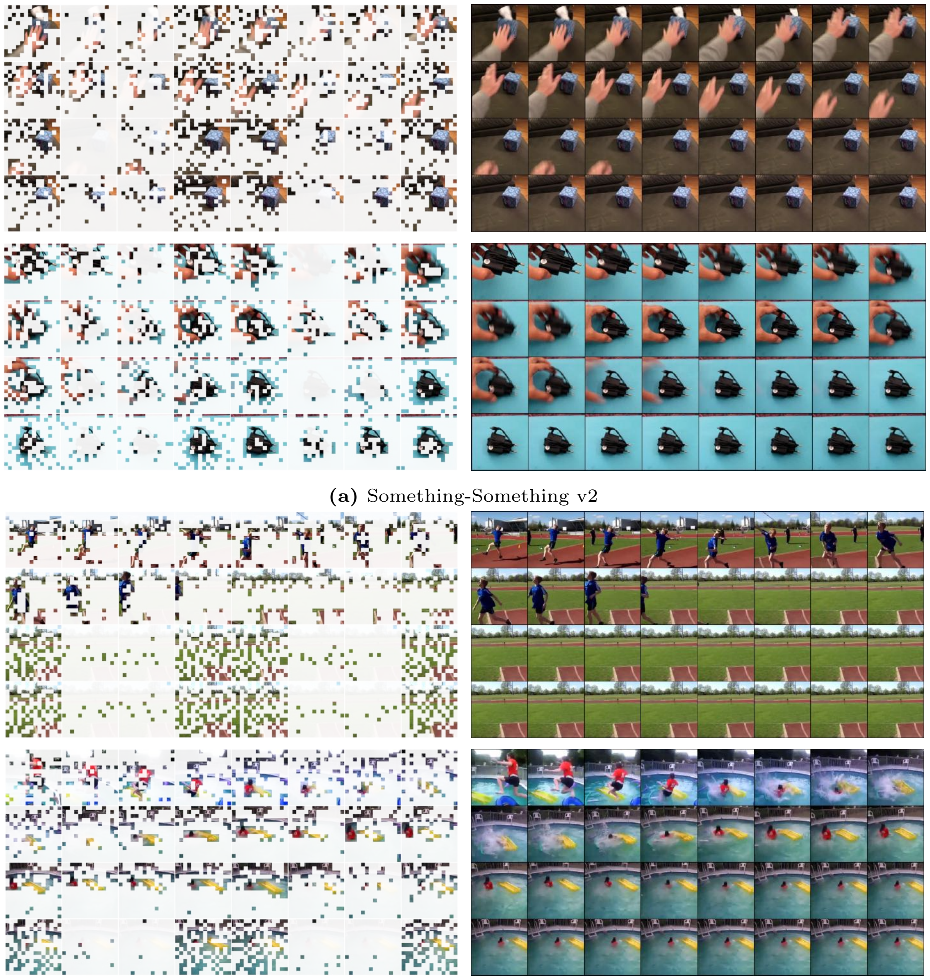
Our method samples non-redundant patches with respect to both spatial and temporal axes. Since (even complex) background is likely to be redundant over time, our method tends to sample patches around moving objects rather than fixed complex backgrounds, as depicted in Fig. 4 and Fig. 7.

E Additional Visualizations for Structure-awareness Parameters

We visualize various selections of the structure-awareness parameters: the division parameters H', W' and T' and the hybrid sampling parameters. Specifically, in

Fig. 8, Fig. 9, Fig. 10 and Fig. 11 we visualize the $H' = W' = 1$, $T' = 8$ under different hybrid samplings. This setting is considered with Motionformer [35] models in our experiments.

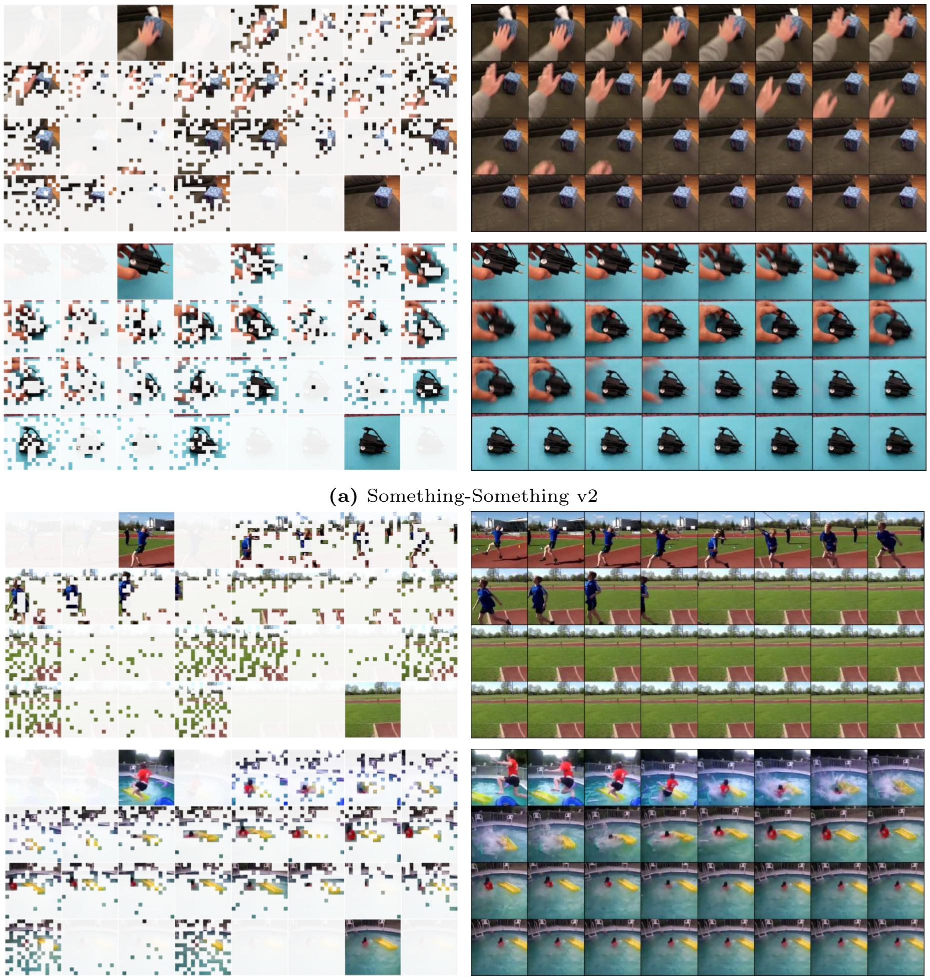
We also visualize the $H' = W' = 14$, $T' = 8$ under different hybrid samplings in Fig. 12, Fig. 13, Fig. 14 and Fig. 15, which are considered with TimeSformer [3] and XViT [4] models in our experiments.



(a) Something-Something v2

(b) Kinetics-400

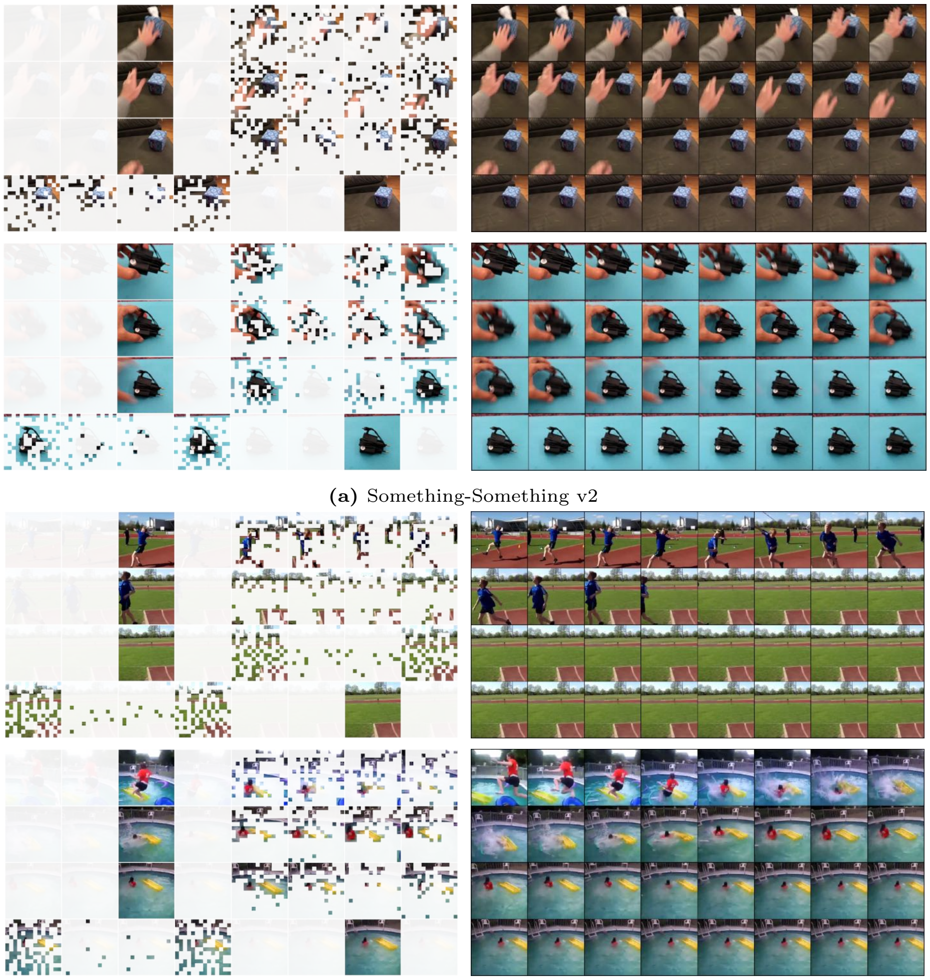
Fig. 8. Visualization examples of our K-centered patch sampling on Something-Something v2 and Kinetics-400 dataset with $H'(W') = 1, T' = 8$ and Hybrid-0. The highlighted region in the left video denotes the sampled patches, and the right video indicates the original video. The distance coefficients are uniformly set $\omega_s = \omega_t = 1$.



(a) Something-Something v2

(b) Kinetics-400

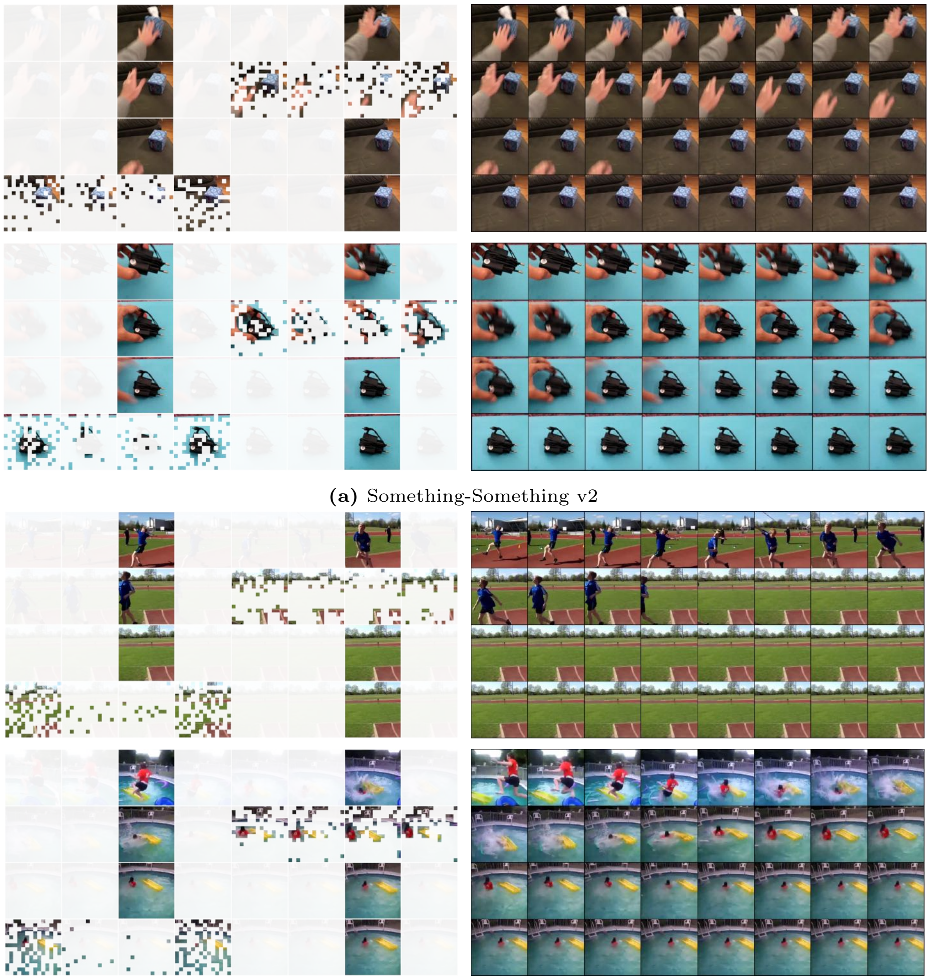
Fig. 9. Visualization examples of our$K$-centered patch sampling on Something-Something v2 and Kinetics-400 dataset with $H'(W') = 1, T' = 8$ and Hybrid-2. The highlighted region in the left video denotes the sampled patches, and the right video indicates the original video. The distance coefficients are uniformly set $\omega_s = \omega_t = 1$.



(a) Something-Something v2

(b) Kinetics-400

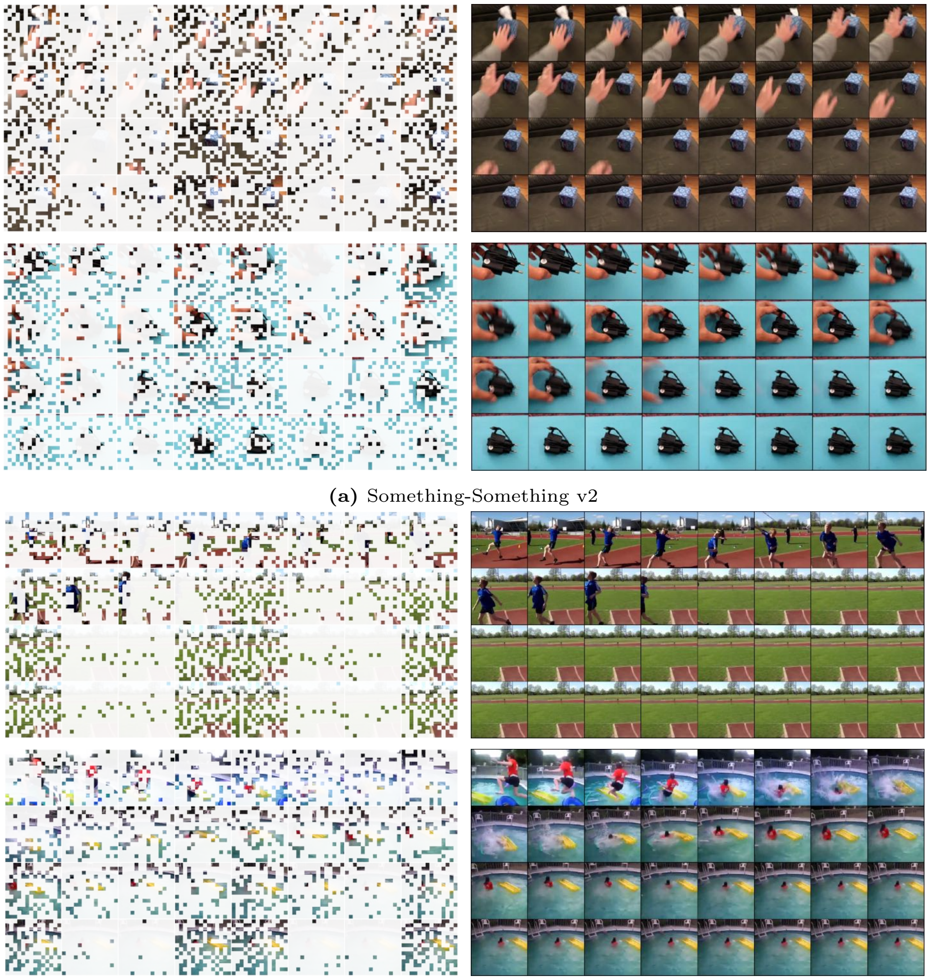
Fig. 10. Visualization examples of our K-centered patch sampling on Something-Something v2 and Kinetics-400 dataset with $H'(W') = 1, T' = 8$ and Hybrid-4. The highlighted region in the left video denotes the sampled patches, and the right video indicates the original video. The distance coefficients are uniformly set $\omega_s = \omega_t = 1$.



(a) Something-Something v2

(b) Kinetics-400

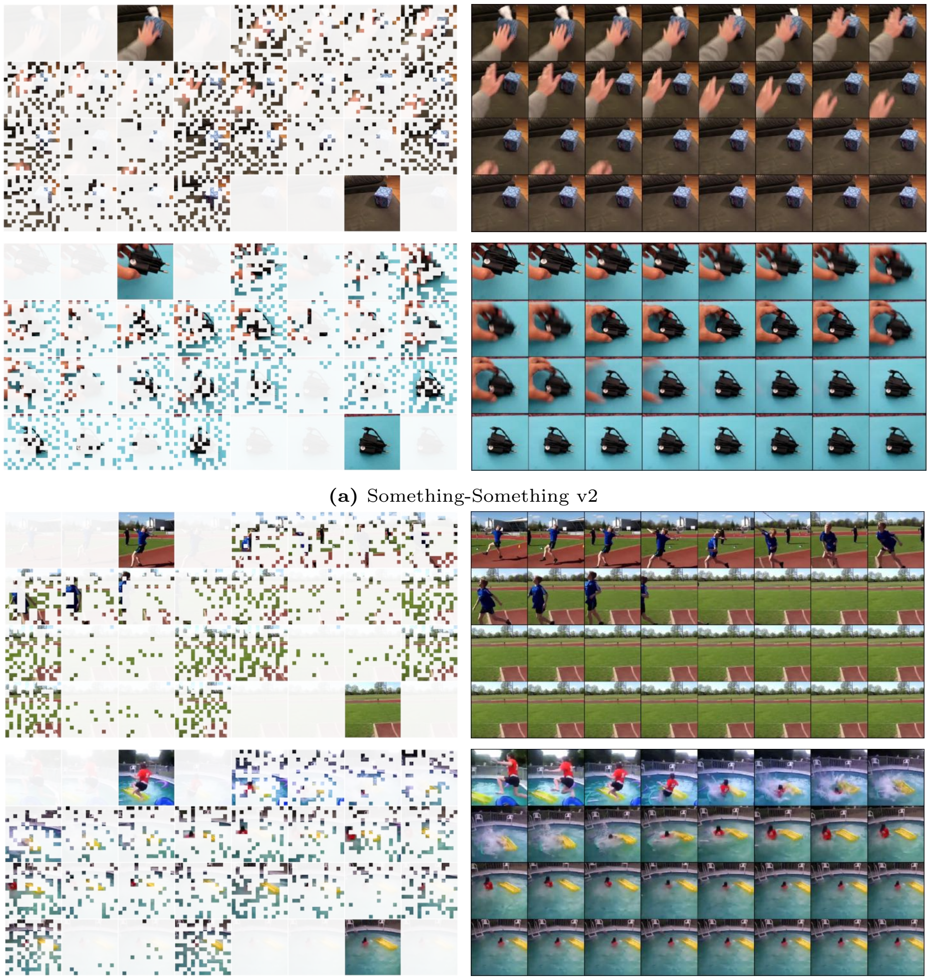
Fig. 11. Visualization examples of our K-centered patch sampling on Something-Something v2 and Kinetics-400 dataset with $H'(W') = 1, T' = 8$ and Hybrid-6. The highlighted region in the left video denotes the sampled patches, and the right video indicates the original video. The distance coefficients are uniformly set $\omega_s = \omega_t = 1$.



(a) Something-Something v2

(b) Kinetics-400

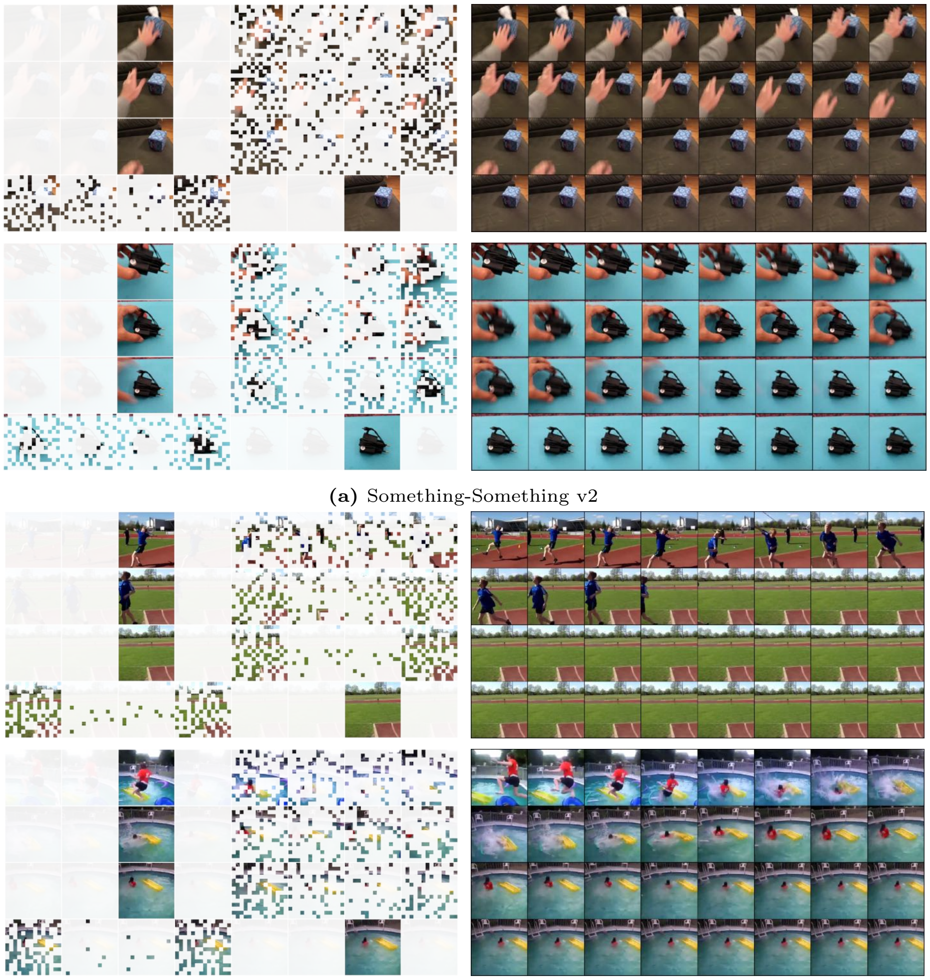
Fig. 12. Visualization examples of our K-centered patch sampling on Something-Something v2 and Kinetics-400 dataset with $H'(W') = 1, T' = 14$ and Hybrid-0. The highlighted region in the left video denotes the sampled patches, and the right video indicates the original video. The distance coefficients are uniformly set $\omega_s = \omega_t = 1$.



(a) Something-Something v2

(b) Kinetics-400

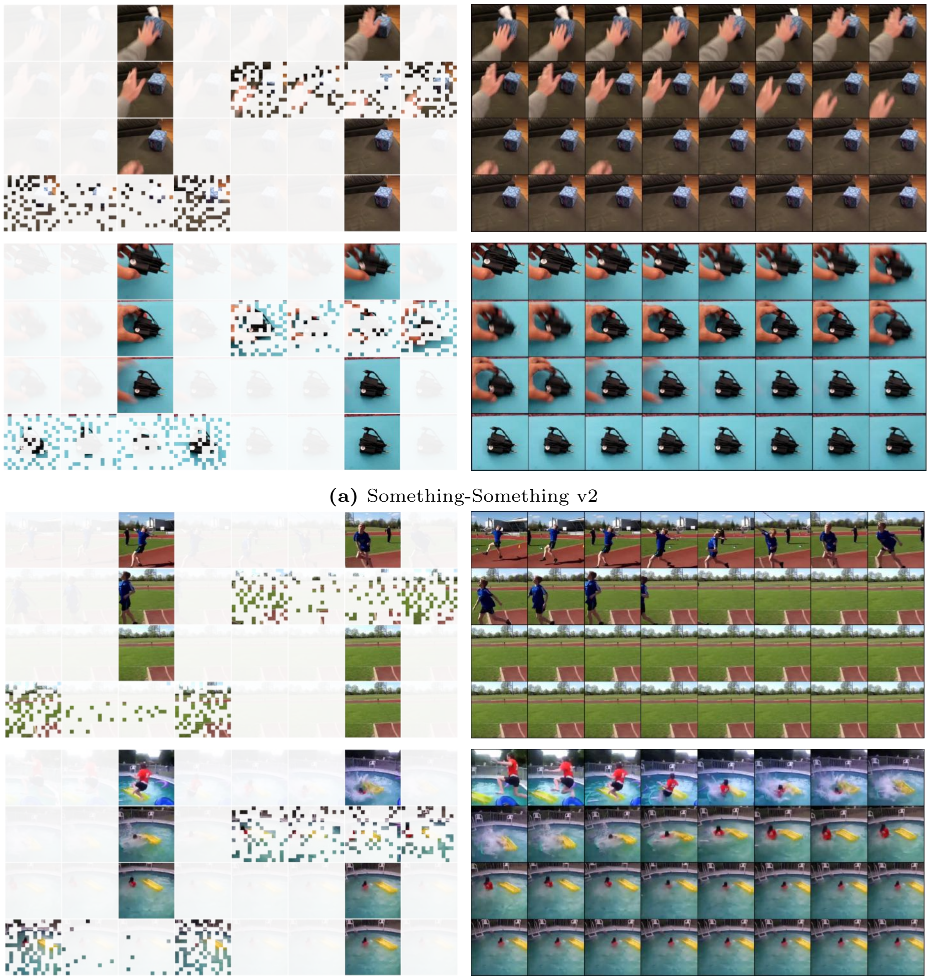
Fig. 13. Visualization examples of our K-centered patch sampling on Something-Something v2 and Kinetics-400 dataset with $H'(W') = 14, T' = 8$ and Hybrid-2. The highlighted region in the left video denotes the sampled patches, and the right video indicates the original video. The distance coefficients are uniformly set $\omega_s = \omega_t = 1$.



(a) Something-Something v2

(b) Kinetics-400

Fig. 14. Visualization examples of our K-centered patch sampling on Something-Something v2 and Kinetics-400 dataset with $H'(W') = 14, T' = 8$ and Hybrid-4. The highlighted region in the left video denotes the sampled patches, and the right video indicates the original video. The distance coefficients are uniformly set $\omega_s = \omega_t = 1$.



(a) Something-Something v2

(b) Kinetics-400

Fig. 15. Visualization examples of our K-centered patch sampling on Something-Something v2 and Kinetics-400 dataset with $H'(W') = 14, T' = 8$ and Hybrid-6. The highlighted region in the left video denotes the sampled patches, and the right video indicates the original video. The distance coefficients are uniformly set $\omega_s = \omega_t = 1$.