	Kinetics-400			Something-Something-v2		
Model	Acc	FT Time	Speedup	Acc	FT Time	Speedup
ViT-B	80.1	14.4h	1.0×	70.3	10.1h	1.0×
ToMe$_{r_{64}}$	80.0	13.4h	1.1×	69.7	9.4h	1.1×
STA$_{r_{64}}$	80.0	13.4h	1.1×	69.4	9.4h	1.1×
Random (0.7)	79.2	10.2h	1.4×	69.3	7.2h	1.4×
RLT (Ours)	**80.1**	**10.2h**	**1.4×**	**70.2**	**7.2h**	**1.4×**
ViT-L	84.8	21.6h	1.0x	74.3	15.2h	1.0×
ToMe$_{r_{64}}$	84.4	18.3h	1.2×	74.3	12.9h	1.2×
STA$_{r_{64}}$	82.2	18.1h	1.2×	73.8	12.7h	1.2×
Random (0.7)	83.1	15.4h	1.4×	74.3	10.8h	1.4×
RLT (Ours)	**84.7**	**15.4h**	**1.4×**	**74.4**	**10.8h**	**1.4×**

Table 1: **Training results on action recognition.** RLT significantly reduces fine-tuning time with comparable performance to the baseline on both Kinetics-400 and Something-Something-v2.

Model	Acc	FT Time
ViT-B	80.1	14.4h
RLT (no length)	80.1	10.2h
RLT	80.1	10.2h
RLT (no length, w/random)	79.3	8.1h
RLT (w/random)	79.8	8.1h
ViT-L	84.8	21.6h
RLT (no length)	84.6	15.4h
RLT	84.6	15.4h
RLT (no length, w/random)	84.2	11.3h
RLT (w/random)	83.3	11.3h

Table 2: **Effect of length encoding.** When fine-tuning with RLT only, length encoding has minimal effect, but helps significantly when combined with random masking.

	Kinetics-400				Something-Something-v2			
Model	Acc	GFLOPS	Clips/s	Speedup	Acc	GFLOPS	Clips/s	Speedup
ViT-B	80.5	180	31.4	1.0×	70.8	180	31.4	1.0×
ToMe$_{r_{64}}$	80.4	131	34.4	1.09×	69.1	131	34.4	1.09×
STA$_{r_{64}}$	80.4	131	34.4	1.09×	69.1	131	34.4	1.09×
Random	80.1	120	**53.0**	**1.68×**	69.3	120	**53.0**	**1.68×**
RLT (Ours)	**80.6**	**120**	52.6	1.67×	**69.8**	**120**	52.6	1.67×
ViT-L	84.8	598	11.5	1.0×	74.3	598	11.5	1.0×
STA$_{r_{64}}$	80.4	308	34.4	1.09×	69.1	308	34.4	1.09×
ToMe$_{r_{64}}$	84.3	**285**	**19.3**	**1.68×**	73.6	**285**	**19.3**	**1.68×**
Random	84.1	405	18.8	1.63×	73.3	405	18.8	1.63×
RLT (Ours)	**84.6**	405	18.71	1.62×	**74.1**	405	18.71	1.62×
ViT-H	86.8	1192	6.65	1.0×	-	-	-	-
ToMe$_{r_{32}}$	86.1	766	8.51	1.27×	-	-	-	-
STA$_{r_{64}}$	80.4	**611**	34.4	1.09×	-	-	-	-
Random	85.1	816	9.66	1.45×	-	-	-	-
RLT (Ours)	**86.3**	816	**9.66**	**1.45×**	-	-	-	-

Table 3: **Inference-only results on action recognition.** With batch size 1, RLT with $\tau = 0.1$ consistently achieves the closest performance to the baseline, comparable or faster than Token Merging or random masking. We omit ViT-H results on Something-Something-v2 due to lack of existing pre-trained checkpoints.

Dataset	Accuracy	Time
ViT-B	80.1	14.4h
RLT (samples)	80.1	**10.2h**
RLT (tokens)	**80.9**	14.4h
ViT-L	84.8	21.6h
RLT (samples)	84.6	**15.4h**
RLT (tokens)	**85.1**	21.6h

Table 4: **Number of Samples vs. Number of Tokens.** Training with RLT on the same number of tokens leads to improved performance over the baseline, since it can train for more epochs in the same amount of time. Training with the same number of samples is much faster.

Dataset	FPS	Seq Length	RLT Seq Length
K400	7.5	1568	1113 (-29%)
K400	15	3136	2007 (-36%)
K400	30	6272	3450 (-45%)
SSv2	7.5	1568	1082 (-31%)
SSv2	15	3136	1942 (-38%)
SSv2	30	6272	3261 (-48%)
EK-100	3.5	1568	1004(-36%)
COIN	30	6272	1819 (-71%)
Breakfast	15	6272	1317 (-79%)

Table 5: **Per-Dataset Sequence Length Reduction.** RLT reduces the average sequence length across multiple datasets. All frames are resized to 224×224.

Dataset	Threshold(τ)	Top-1 Accuracy	Throughput (clips/s)
K400	0(base)	80.1	11.5
K400	0.05	80.3	13.8
K400	0.1	80.2	18.4
K400	0.15	80.0	18.9
K400	0.2	79.2	20.2
SSv2	0(base)	74.3	11.5
SSv2	0.05	74.5	13.9
SSv2	0.1	74.3	18.7
SSv2	0.15	74.1	20.3
SSv2	0.11	73.8	22.1

Table 6: **Threshold effect across datasets** We measure the effect of varying the threshold for ViT-L on both Kinetics and SSv2.