Table S.1: **Ablation study on the impact of the number of steps r in soft grouping.**We conducted experiments with modular training using the DeiT-S model. The decoupled embedding module was trained for 10 epochs. We observed that the decoupled embedding module, when trained with a high reduction rate r, generalizes well to lower rates.

reduction rate r at train time	test-time reduction rate r			
	16	14	12	10
16	78.92	79.33	79.42	79.60
14	78.80	79.23	79.43	79.60
12	78.77	79.22	79.41	79.61
10	78.67	79.19	79.38	79.60

Table S.2: **Ablation study on the impact of temperature scaling.** We experimented with a modular training using the DeiT-S model. We trained the decoupled embedding module for 10 epochs. we observe that values within the range of 0.1 to 0.3 consistently provide gains with an accuracy difference of 0.1%.

Temperature scale	Acc@1 (-50%)	Acc@1 (-35%)
0.05	78.41	79.19
0.1	78.87	79.51
0.2	78.91	79.50
0.3	78.92	79.57
0.5	78.83	79.54
1	78.59	79.34

Table S.3: **Image classification results with AugReg ViT-S pretrained on 384×384 resolution.** The result shows that our method can adapt to settings with an increased number of tokens, achieving performance gains.

Method	Reduction	Acc@1	GFLOPs	im/s
ViT-S (384)	-	83.8	15.7	394
ToMe	51.5% ($r = 47$)	82.1	7.60	728
DTEM	52.0% ($r = 48$)	82.3	7.54	733

Table S.4: **Rank correlation coefficient changed through training.** We monitor changes in the Kendall rank correlation between token similarities derived from two different features: self-attention keys (as in ToMe) and decoupled embeddings. The result shows a decreased correlation as learning progresses, indicating that the decoupled embedding seeks a different measure of similarity for merging.

	1 to 4 Blocks	5 to 8 Blocks	9 to 12 Blocks
Kendall's τ	$0.517 \rightarrow 0.401$	$0.457 \rightarrow 0.402$	$0.591 \rightarrow 0.519$

Table S.5: **Comparison with alternative design choices for soft grouping.** In ToMe + GS, we applied the Gumbel Softmax with the top-1 operation from ToMe. We also tested DynamicViT applied in modular way (off-the-shelf). We also test removing effective size m and subsequent proportional attention. The results shows that our proposed design for soft grouping performs the best.

Method	-50% reduction	-35% reduction
ToMe + GS + soft merging	78.14	79.2
DynamicViT (off-the-shelf)	75.53	78.96
DTEM	78.99	79.44
- wo prop attn	77.86	79.16

Table S.6: **Ablation study by varying the decoupled embedding dimension on captioning and segmentation**. In the main experimental results, we use a dimension of 64 for the decoupled embedding module. The results demonstrate that this module directly impacts the quality of token merging.

	dimension	32	48	64*	128
Captioning (base, r=13)	CIDEr	106.5	108.8	110.4	113.1
	FLOPs	10.46	10.48	10.5	10.57
Segmentation (r=0.4)	mIoU	40.77	41.6	42.64	-
	FLOPs	22.07	22.17	22.27	-

Table S.7: **Extended image captioning evaluation results when token merging is applied.** The result shows that our method is particularly effective in challenging, more resource-constrained settings with higher reduction rates. caption. We note that for reduction rates over 41% and 49% for GIT-B and GIT-L respectively, there was a significant decrease in captioning quality.

	Reduction	B@4	M	C	S	#		Reduction	B@4	M	C	S	#
GIT-B	-	38.8	30.1	127.6	23.6	197	GIT-L	-	40.7	29.6	134	23.8	197
ToMe	12%	**37.9**	**28.6**	**123.7**	**22.4**	149	ToMe	-	-	-	-	-	-
	25%	**35.4**	**27.1**	**115.9**	**21**	101		18%	40.1	28.9	131.1	23	125
	27%	**35.7**	**26.9**	**115.3**	**20.9**	89		24%	39.4	28.8	128.7	22.7	101
	32%	34.6	26.4	113.1	20.3	77		31%	36.9	27.3	122.1	21.5	77
	35%	33.5	25.8	109.3	19.8	65		37%	36.4	27.1	120.1	21.5	53
	38%	33.3	25.5	107.9	19.5	53		43%	34.0	25.8	112.2	20.2	29
	41%	31.9	24.8	104.3	19.0	41		49%	31.7	24.8	105.1	19.3	7
DTEM	12%	38	28.6	124.2	22.3	149	DTEM	-	-	-	-	-	-
	25%	36	27.3	118.9	21.4	101		18%	40.1	29.1	131.5	23.2	125
	27%	36.4	27.4	119.3	21.4	89		24%	39.4	28.9	129.5	23	101
	31%	36.2	27.1	118.1	20.8	77		31%	37.9	27.8	124.4	21.9	77
	34%	34.5	26.5	114.2	20.5	65		37%	37.0	27.5	122.9	21.7	53
	37%	34.3	26.2	112.9	20.1	53		43%	35.7	26.6	117.6	20.9	29
	41%	33.3	25.7	110.4	19.9	41		49%	33.3	25.7	111.1	20.1	7