Adaptive Token Sampling for Efficient Vision Transformers

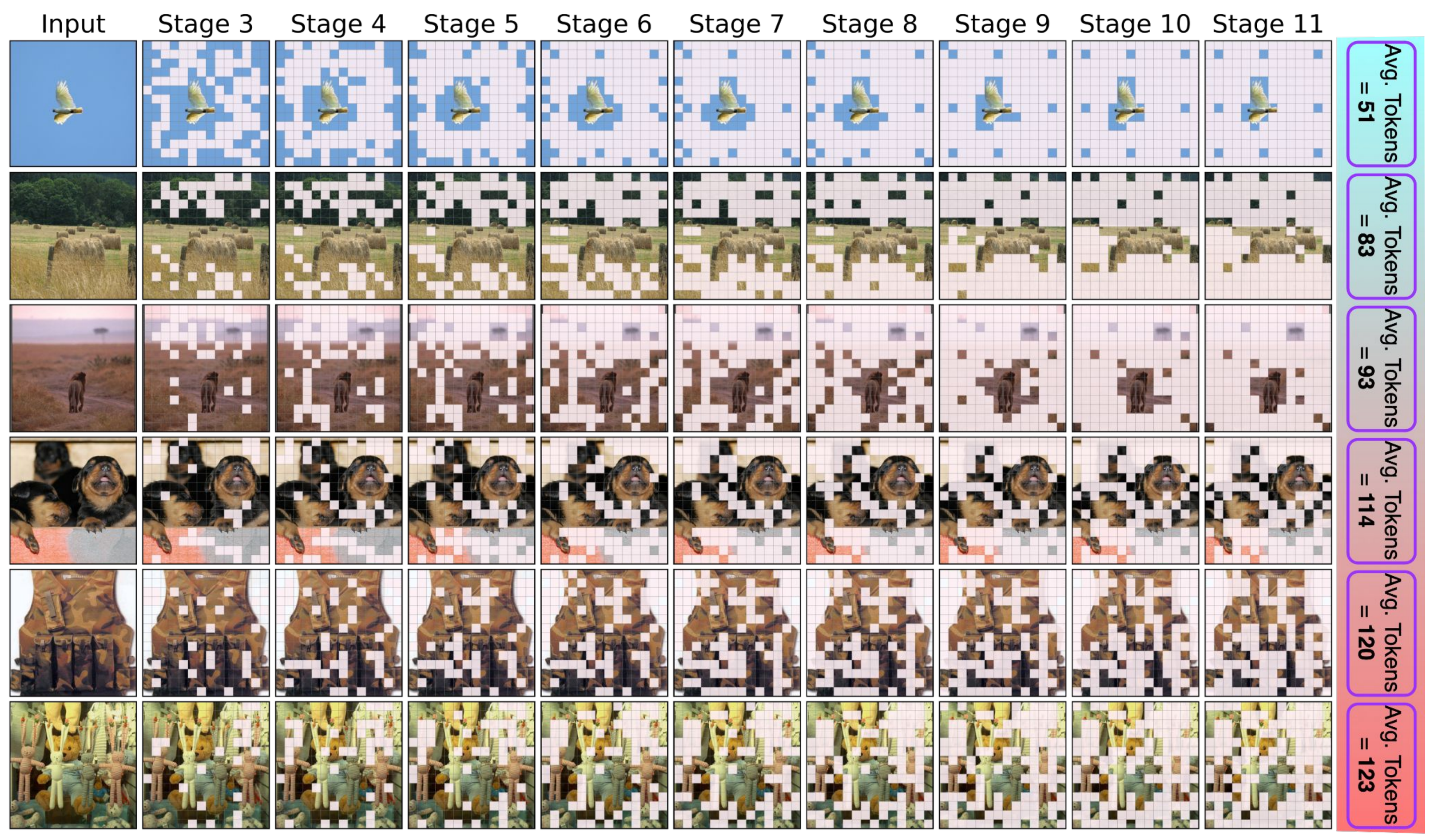
Mohsen Fayyaz*, Soroush Abbasi Koohpayegani*, Farnoush Rezaei Jafari*,
Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, Jürgen Gall

Microsoft, UC Davis (University of California), TU Berlin, BIFOLD, ECCV, Universität Bonn

Redundancy in Vision Transformers

Problem

- In conventional neural networks, the amount of computation used is proportional to the size of the input, instead of the complexity of the content of the data being processed.
- Typical input data for neural architectures have an inherent complexity that is independent of the input size.
- Static tokens resolution in vision transformers leads to unnecessary computational overhead.



✅ **Our Solution:** Adaptively sample significant tokens based on the input content!

Adaptive Token Sampling

ATS is an ***adaptive differentiable parameter-free*** module, which can be plugged into existing vision transformers and make them more ***efficient***.

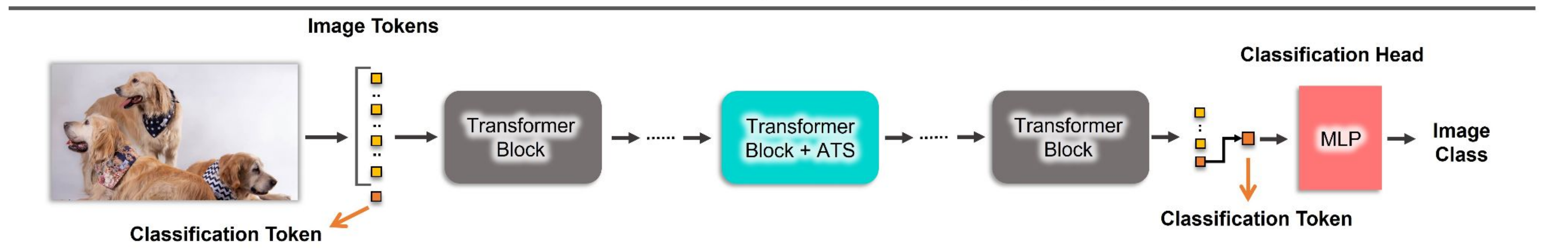
✔**Adaptive** ⟶ ATS adapts the number of tokens based on the complexity of the input image/video.

✔**Differentiable** ⟶ A vision transformer equipped with ATS can be trained/fine-tuned.

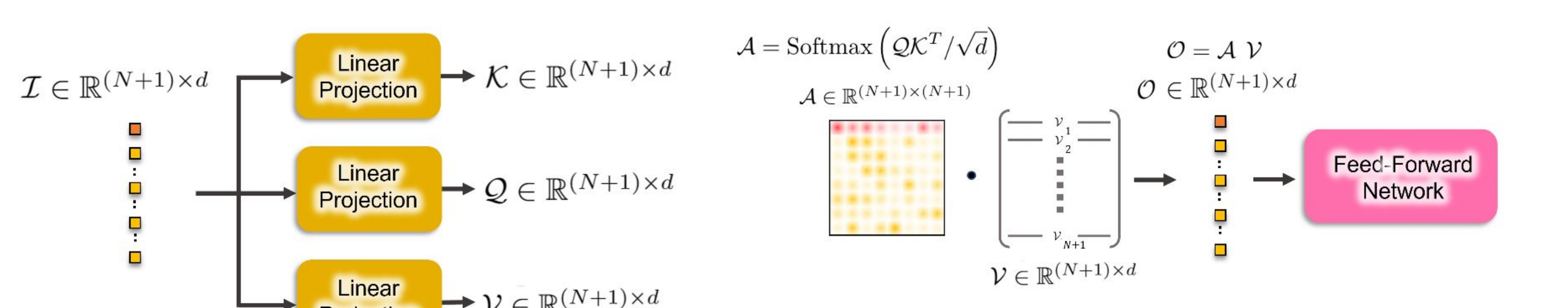
✔**Parameter-free** ⟶ ATS can also be added to the existing off-the-shelf pre-trained vision transformers without any further training.

✔**Efficient** ⟶ ATS improves the SOTA by reducing their computational costs (GFLOPs) by 2X.

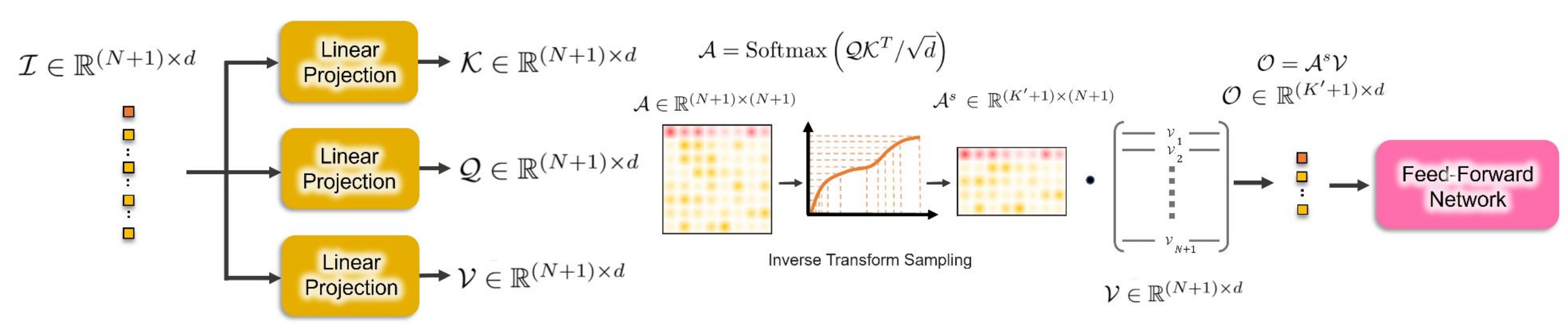
Vision Transformer + Adaptive Token Sampler (ATS)



Transformer Block

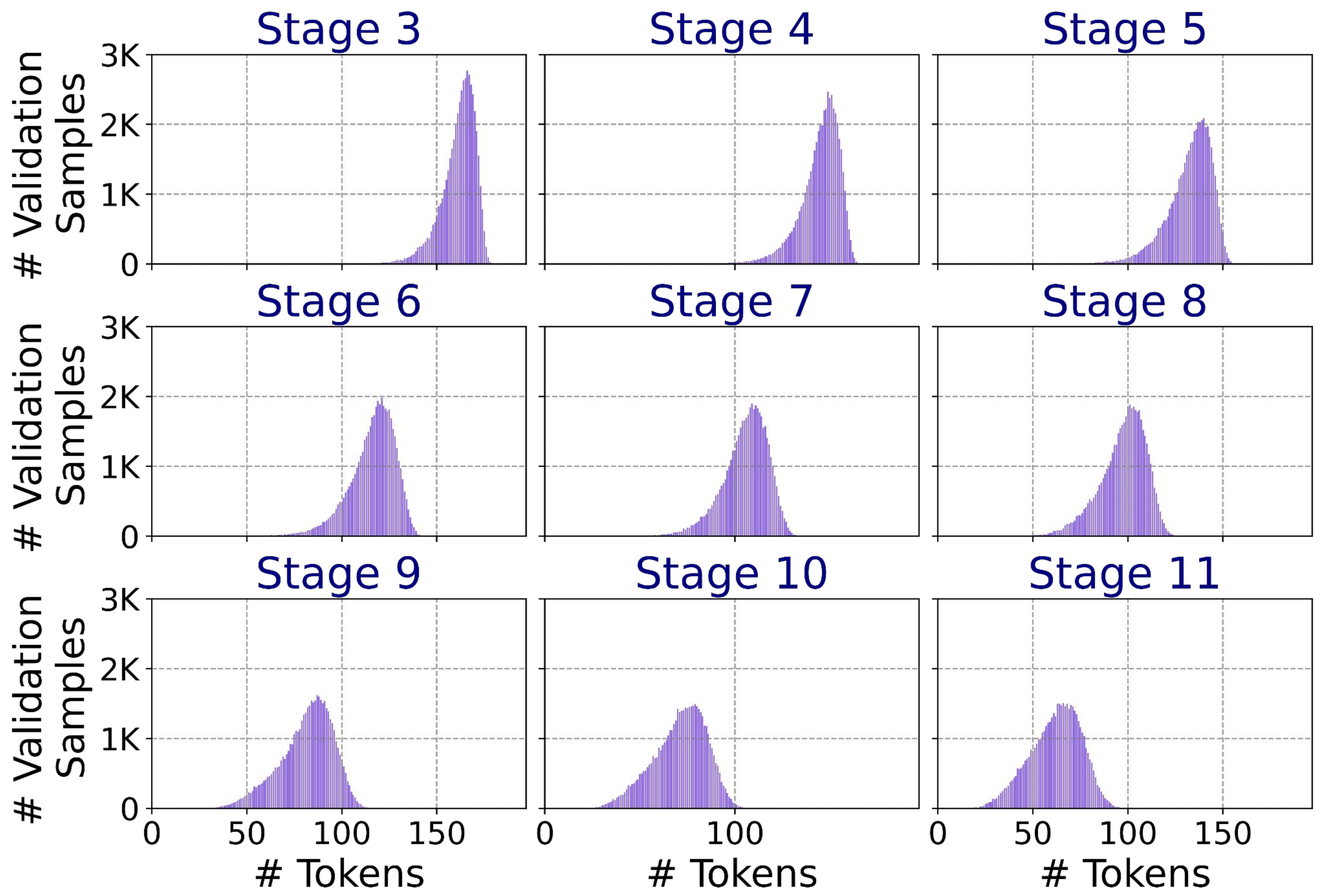


Transformer Block + Adaptive Token Sampler (ATS)



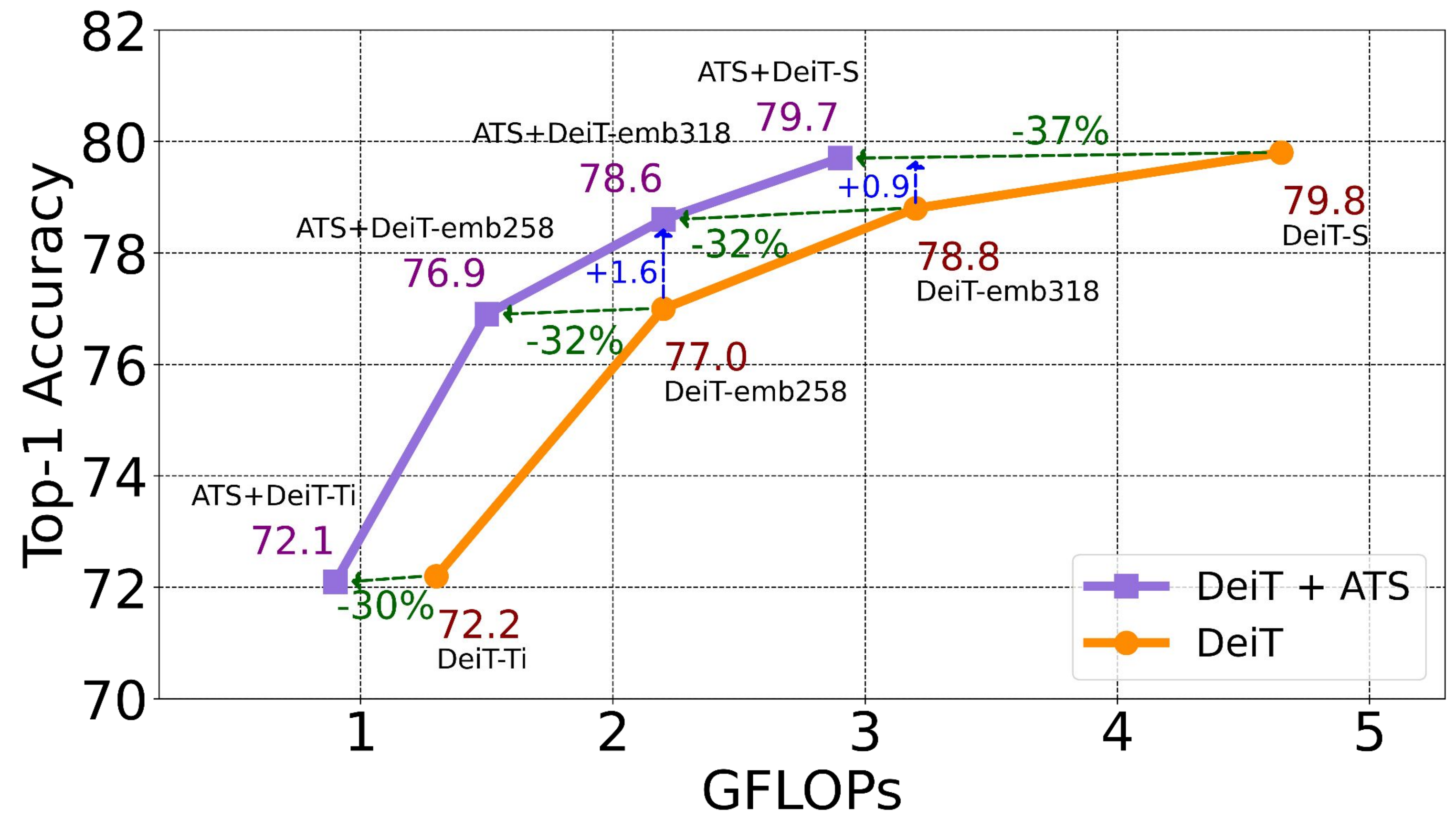
Adaptive Sampling

- Histogram of the number of sampled tokens at each ATS stage of our multi-stage DeiT-S+ATS model on the ImageNet validation set.
- The y-axis corresponds to the number of images.
- The x-axis corresponds to the number of sampled tokens.



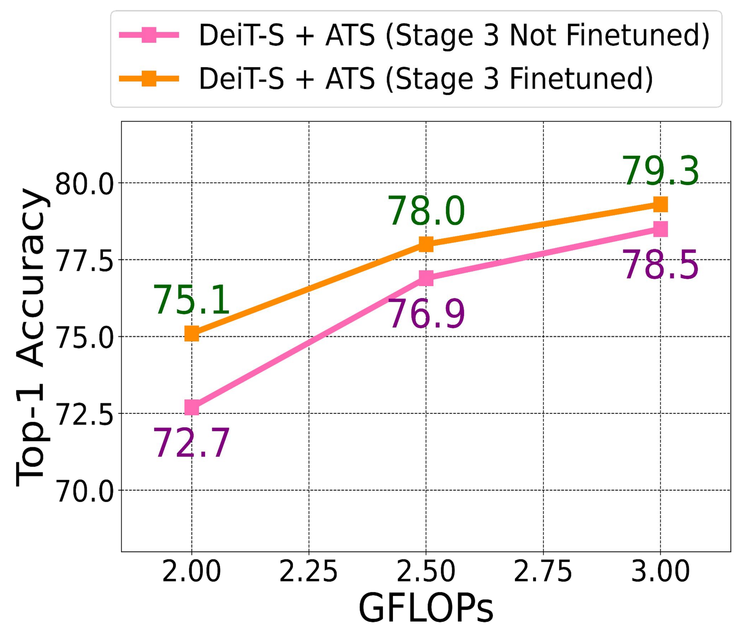
Model Scaling

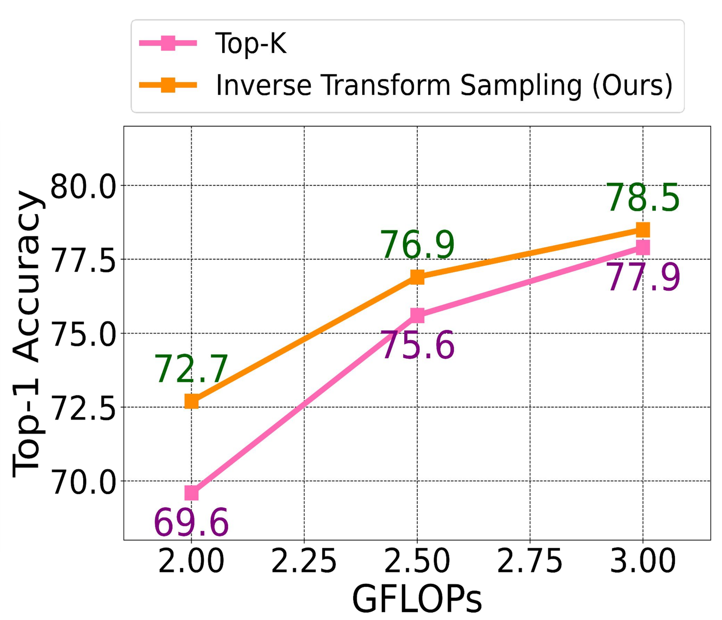
- Performance comparison on the ImageNet validation set.
- Our proposed adaptive token sampling method achieves a state-of-the-art trade-off between accuracy and GFLOPs.
- We can reduce the GFLOPs of the DeiT-S model by 37% while almost maintaining its accuracy.

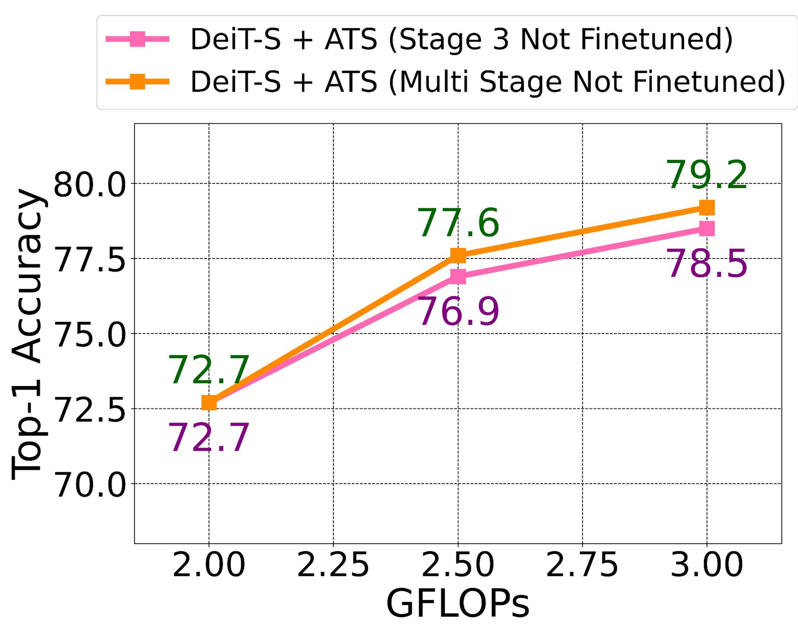


Ablations

All results are on ImageNet-1K val. set







SOTA Results

Action Recognition Results on Kinetics-600

Model	Top-1	Top-5	Views	GFLOPs
X3D-XL [17]	81.9	95.5	10×3	1,452
X3D-XL+ATFR [16]	82.1	95.6	10×3	768
TimeSformer-HR [1]	82.4	96	1×3	5,110
TimeSformer-HR+ATS (Ours)	82.2	96	1×3	**3,103**
ViViT-L/16x2 [1]	82.5	95.6	4×3	17,352
Swin-B [39]	84.0	96.5	4×3	3,384
MViT-B-24, 32×3 [14]	84.1	96.5	1×5	7,080
TokenLearner 16at12(L/16) [49]	84.4	96.0	4×3	9,192
X-ViT (16×) [2]	84.5	96.3	1×3	850
X-ViT+ATS (16×) (Ours)	84.4	96.2	1×3	**521**

Image Classification Results on ImageNet-1K

Model	Params (M)	GFLOPs	Resolution	Top-1
ViT-Base/16 [13]	86.6	17.6	224	77.9
HVT-S-1 [42]	22.09	2.4	224	78.0
IA-RED^2 [41]	-	2.9	224	78.6
DynamicViT-DeiT-S (30 Epochs) [46]	22.77	2.9	224	79.3
EViT-DeiT-S (30 epochs) [36]	22.1	3.0	224	79.5
DeiT-S+ATS (Ours)	**22.05**	**2.9**	224	79.7
DeiT-S [53]	22.05	4.6	224	79.8
PS-ViT-B/18 [68]	21.3	8.8	224	82.3
PS-ViT-B/18+ATS (Ours)	21.3	**5.6**	224	82.2
CvT-21$_{384}$ [63]	32.0	24.9	384	83.3
CvT-21$_{384}$+ATS (Ours)	32.0	**17.4**	384	83.1