Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Please see section 4.3.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Please see section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] Our work does not contain theoretical results.
   (b) Did you include complete proofs of all theoretical results? [N/A] Our work does not contain theoretical results.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We used publicly available data in all of our experiments. Meanwhile we either provide the detailed implementations or cite the papers of them following the authors instructions (See Section 4). All of our codes are provided in https://github.com/VITA-Group/SViTE.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We provides all the training details in Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We did not report the error bars since running vision transformer on ImageNet are extremely resource-consuming. For example, each reported number takes around 960 V100 GPU hours. We will continue running the experiments and report the confidence intervals in future versions.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We describe the details of computation resources in Section A1 of the supplement.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We used publicly available data, i.e., ImageNet, in our experiments. We cited the corresponding papers published by the creators in Section 4.
   (b) Did you mention the license of the assets? [No] The license of ImageNet is included in the paper that we have cited.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The ImageNet we used are publicly available. And all our codes are included in https://github.com/VITA-Group/SViTE.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We did not collect/curate new data.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] All ImageNet datasets are already publicly available and broadly adopted. I do not think there are any issues of personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A1 More Implementation Details

Computing resources. All experiments use Tesla V100-SXM2-32GB GPUs as computing resources. Specifically, each experiment is ran with 8 V100s for $4 \sim 5$ days.

Why do we choose 600 training epochs for SViTE experiments? Choosing 600 training epochs for sparse training is to maintain similar training FLOPs compared to dense ViT training. Specifically, if training a dense DeiT model for 300 epochs needs 1x FLOPs, training SViTE-DeiT at 40% sparsity for 600 epochs needs ~ 0.95x FLOPs. In summary, we compare our SViTE (600 epochs) and DeiT baselines (300 epochs) based on a similar training budget. Such comparison fashion is widely adopted in sparse training literature like [35, 34] (see Table 2's caption in [35] and Figure 2 in [34] for details). Meanwhile, note that the reported running time is per epoch saving (i.e., total running time / total epoch), which would not be affected by the number of training epochs.

Baseline models with longer epochs. Actually, the performance of DeiT training without distillation saturates after $300 \sim 400$ epochs, as stated in [2]. We also conduct longer epoch (600 epochs) training for DeiT-Small and -Base models. Our results collected in the table A7 align with the original DeiT paper [2]. It suggests that our proposed SViTE is still able to achieve better accuracy with fewer parameters and fewer training&inference computations. Specifically, at 40% structured sparsity, our sparsified DeiT-Base can achieve 0.21% accuracy gain, at $\sim 51\%$ training FLOPs, 33.13% inference FLOPs, and 24.70% running time savings, compared to its dense counterpart with 600 epochs.

Table A7: Performance of DeiT-Small/-Base with longer training epochs on ImageNet-1K.

Metrics	DeiT-Small 300 Epochs	DeiT-Small 600 Epochs	DeiT-Base 300 Epochs	DeiT-Base 600 Epochs
Accuracy (%)	79.90	80.02 (+0.12)	81.80	82.01 (+0.21)

A2 More Experimental Results

Sparse topology of SViTE-Base with unstructured sparsity. As shown in Figure A6, we observe that from the initial random mask to explored mask in SViTE, plenty of structural patterns emerge (i.e., the darker "vertical" lines mean completely pruned neurons in the MLPs). It is supervising that unstructured sparse exploration can lead to structured patterns, which implies the great potential to be accelerated in real-world hardware devices.

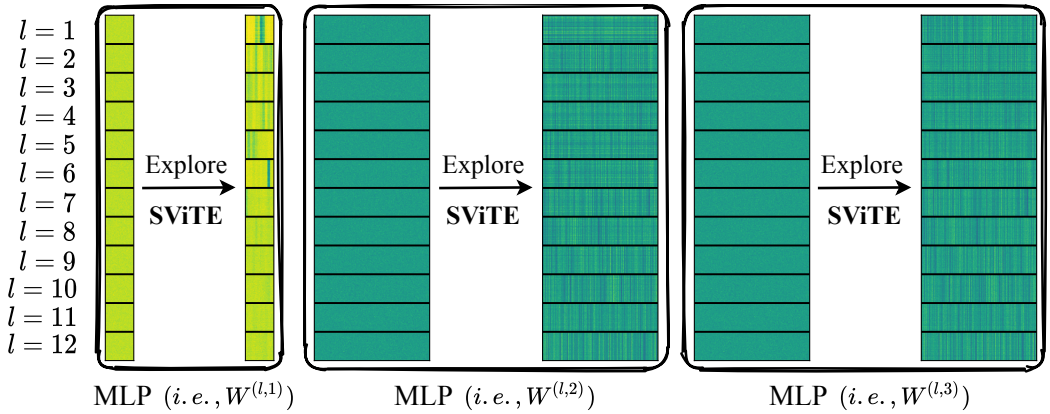

Figure A6: Binary mask visualizations of SViTE-Base at 50% unstructured sparsity. Within each box, *left* is the initial random mask; *right* is the explored mask from SViTE.

Sparse topology of S^{2}ViTE-Base with structured sparsity. Figure A7 shows mask visualizations of pruned multi-attention heads and MLPs in vision transformers. It shows that S^{2}ViTE indeed explores totally different connectivity patterns, compared to the initial topology.

Ablation of only applying our learnable token selector. We compare these three setup: (a) DeiT-Small (79.90 test accuracy); (b) DeiT-Small + Token selector with 10% data sparsity (78.67 test accuracy); (c) DeiT-Small + Token selector with 10% data sparsity + SViTE with 50% unstructured sparsity (79.91 test accuracy). It demonstrates that simultaneously enforcing data and architecture sparsity brings more performance gains.

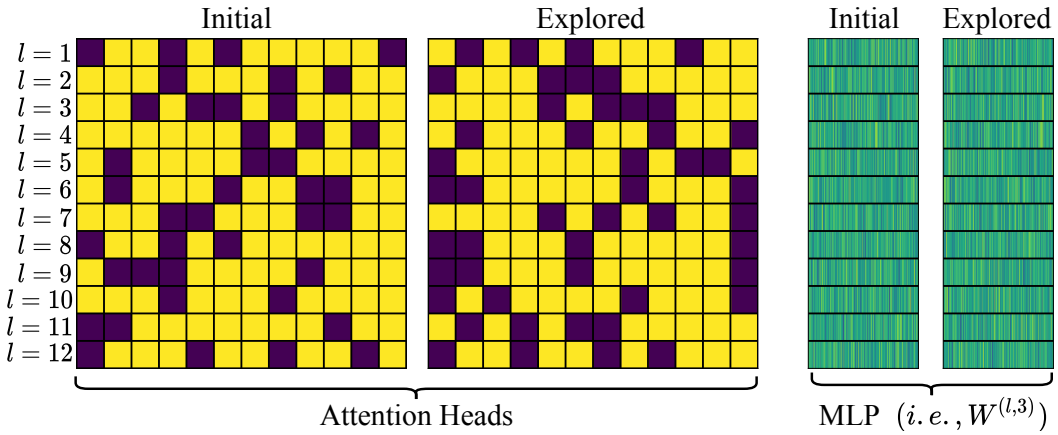

Figure A7: (*Left*) The "survivors" summary of existing attention heads in sparse vision transformers from S^{2}ViTE. Dark entry is the pruned attention head; bright entry means the remaining attention head. (*Right*) Binary masks of all $W^{(l,3)}$ MLPs in S^{2}ViTE-Base. *Initial* denotes the random connectivity in the beginning, and *Explored* is the explored typologies at the end. Visualized S^{2}ViTE-Base has 40% structural sparsity.

Ablation of the layerwise sparsity of attention maps. As shown in Table A8, we investigate the layerwise sparsity of attention maps. Dense DeiT-Small and SViTE+-Small with 10% data sparsity and 50% model sparsity are adopted for experiments. We calculate the percentage of elements in attention maps whose magnitude is smaller than 10^{-4}. We observe that the bottom layers' attention maps of SViTE+ are denser than the ones in dense ViT, while it is opposite for the top layers.

Table A8: Layerwise sparsity of attention maps.

Layer Index	1	2	3	4	5	6	7	8	9	10	11	12
Dense-Small	28.44	44.70	37.04	13.87	13.62	13.67	12.98	5.99	5.08	3.46	2.74	15.05
SViTE+-Small	24.37	33.24	28.05	19.51	9.03	13.79	13.24	8.88	6.78	5.65	2.88	35.22