Exploring the Generalization, Efficiency, and Inductive Bias of SHViT Across Visual Domains

Christo Mathew Priyal Garg Vishal Venkataramani
Rutgers University, New Brunswick
{cm1788, pg501, vv382}@scarletmail.rutgers.edu

Abstract

This project investigates the generalization behavior, inductive bias, and data efficiency of the Single-Head Vision Transformer (SHViT) across three distinct visual domains: natural object classification (CIFAR-100), remote sensing (EuroSAT), and medical imagery (MedMNIST). Contrary to the prevailing literature that Vision Transformers require massive datasets to compete with CNNs, our experiments reveal that SHViT significantly outperforms both MobileNetV2 and DeiT-Tiny in low-data regimes, achieving **5–20%** higher accuracy on 10% training splits. Furthermore, we uncover a novel property of Single-Head Attention: it acts as an architectural regularizer that improves robustness to corruption. Our results demonstrate that SHViT's hybrid design–combining a convolutional stem with single-head global attention–provides a "best of both worlds" inductive bias, offering superior data efficiency and domain transferability than either pure CNNs or standard ViTs.

1 Introduction

Vision Transformers (ViTs) (Dosovitskiy, 2020) have recently achieved state-of-the-art performance on large-scale benchmarks but require substantial computational resources due to dense multi-head attention and fine-grained patch embeddings. Although various efficient designs exist, many focus on reducing computational cost without deeply examining whether architectural simplifications influence inductive bias, robustness, and domain transfer.

SHViT (Yun and Ro, 2024) proposes a simplified transformer that replaces multi-head attention with single-head attention while introducing partial-channel operations and large-stride patchification. These decisions are motivated by the hypothesis that many attention heads are redundant and that early spatial processing can be handled more efficiently by convolutional stems.

However, prior work has primarily evaluated SHViT on large-scale benchmarks such as ImageNet, leaving open questions regarding how the architecture behaves under:

- domain shift (remote sensing or medical imagery),
- small-data training conditions,
- perturbations and corruptions,
- inductive bias differences vs. standard ViTs.

Our goal is to study SHViT beyond the original benchmarks and understand whether its efficiency mechanisms offer practical benefits beyond computation and memory reductions.

2 Background

2.1 Efficient Vision Transformers

Recent work on efficient ViTs attempts to reduce redundancy through:

- token pruning,
- sparse attention,
- convolutional hybrids,
- lightweight patch embeddings.

SHViT differs by applying simplification uniformly across all architectural stages.

2.2 SHViT Architecture

SHViT introduces:

1. **Large-stride patch embedding** to lower token count.
2. **Single-head attention** to avoid redundant computation.
3. **Partial-channel feature routing** that processes subsets of channels.

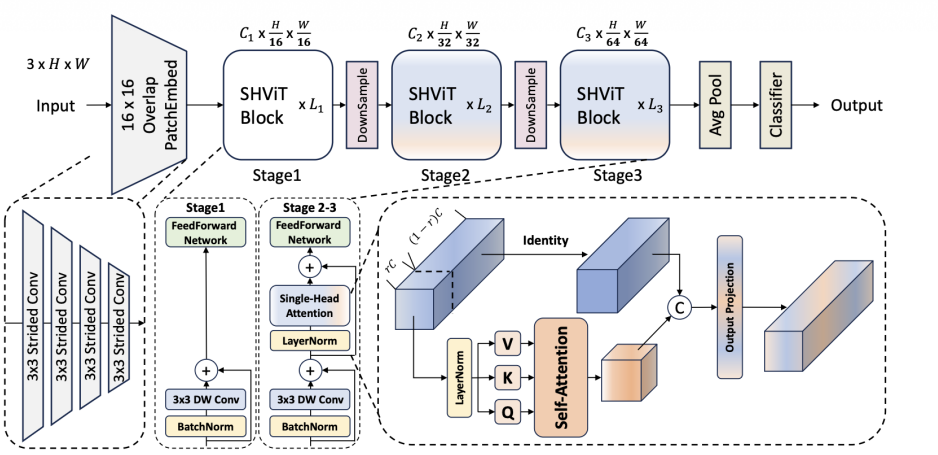


Figure 1: Overview of the SHViT macro-architecture showing convolutional stem, single-head attention, and partial-channel routing.

4. **Convolutional stem** to capture local structure.

3 Research Motivation

Natural, medical, and satellite images differ significantly in their underlying statistics. Natural images are object-centric, while remote sensing and medical images are often texture-centric and scale-invariant. Standard ViTs, with their flexible but data-hungry nature, often fail to generalize to these domains without massive pre-training. Since SHViT deliberately modifies the architectural inductive bias (via a 16×16 convolutional stem and restricted attention heads), it presents an ideal test case for analyzing whether architectural constraints can actually improve generalization. By benchmarking against MobileNet (high inductive bias) and DeiT (low inductive bias) trained from scratch, we aim to isolate the specific contributions of SHViT's hybrid design.

4 Methodology

4.1 Experimental Setup

To isolate the impact of architectural inductive biases, we benchmark **SHViT variants (S1, S2)** against two distinct baselines: **MobileNetV2** (Mehta and Rastegari, 2021), representing a convolutional architecture with strong local priors, and **DeiT-Tiny** (Touvron et al., 2021), representing standard multi-head self-attention with weaker inductive bias[cite: 89, 90, 91].

Training Protocol: Models are trained from scratch on CIFAR-100 (Krizhevsky et al., 2009) and EuroSAT (Helber et al., 2019) to evaluate learning efficiency from random initialization. MedMNIST (PathMNIST) (Yang et al., 2023) is evaluated under constrained conditions to address medical class imbalances. To ensure fairness, all architectures employ a unified training recipe with comparable optimization settings–including consistent learning rate schedules, weight decay, and heavy data augmentation–thereby attributing performance differences to architecture rather than hyperparameter tuning.

4.2 Evaluation Dimensions

We analyze SHViT behavior along four dimensions:

- **Data size:** 10%, 32.5%, 55%, 77.5%, 100% training splits.
- **Domain:** natural (CIFAR), remote sensing (EuroSAT), and medical (PathMNIST) domains.
- **Corruptions and transforms:** Gaussian noise, rotation, resizing, and grayscale, at multiple severity levels following the common corruption benchmark (Hendrycks and Dietterich, 2019).(Figures in Appendix).
- **Patch stride:** varying stride inside SHViT to study efficiency-performance trade-offs and effects on domain generalization.

This protocol treats SHViT as a probe for inductive bias and robustness across domains, rather than focusing on a single benchmark scenario.

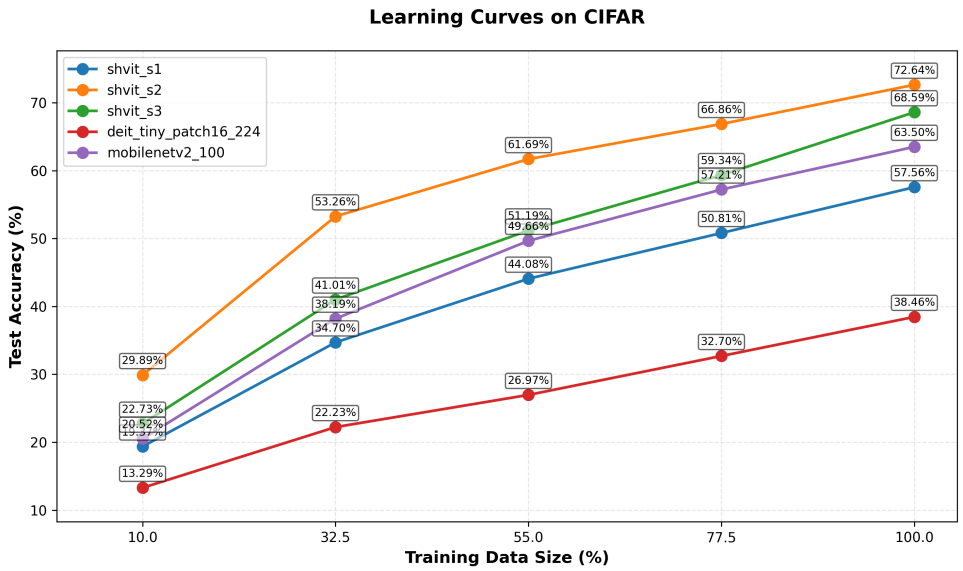


Figure 2: Top-1 Accuracy vs. training fraction on CIFAR-100 for MobileNet, DeiT-Tiny, and SHViT variants.

5 Results

5.1 Low-Data Learning and Accuracy

Contrary to standard expectations, SHViT demonstrates superior data efficiency compared to CNNs. As shown in Figure 2, SHViT-S2 outperforms MobileNetV2 by a significant margin in the ultra-low data regime (10% training data). On CIFAR-100, SHViT-S2 achieves 29% accuracy compared to MobileNet's 20% and DeiT's 13%. Similarly, on EuroSAT (10% split), SHViT reaches 72% accuracy while MobileNet lags at 52%. This suggests that the 16×16 convolutional stem provides sufficient early stability to mimic a CNN's learning curve, while the global attention mechanism allows it to surpass the CNN's capacity even with limited samples.

Learning curves for EuroSAT and MedMNIST follow similar trends and are provided in Appendix A.

5.2 Domain Generalization and Transfer

Figure 3 reveals that SHViT-S2 adapts more effectively than both baselines, achieving 97% accuracy compared to MobileNetV2 (95%) and DeiT-Tiny (92%). This "Rapid Domain Adaptation" indicates that the features learned by SHViT are more transferable. While MobileNet captures the local textures of satellite imagery well, SHViT's single-head attention likely captures global geometric structures (e.g., highway grids, river paths) that are missed by the CNN's limited receptive field, resulting in superior fine-tuning performance.

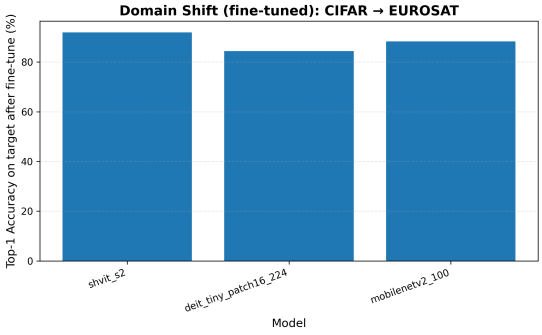


Figure 3: Domain shift results from CIFAR→EuroSAT and CIFAR→MedMNIST.

5.3 Robustness to Corruptions and Transformations

Across Gaussian noise, rotation, resizing, and grayscale conversion, SHViT shows smoother degradation than DeiT-Tiny, especially in low-data regimes. MobileNet remains strong under mild corruptions due to convolutional locality.

Due to space constraints, full quantitative robustness curves (severity 1, 3, and 5), rotation, and color robustness are deferred to Appendix B and Appendix C.

5.4 Patch Stride Effects

We observe a domain-dependent sensitivity to patch stride. On object-centric tasks like CIFAR, increasing the patch stride from 8×8 to 32×32 causes a massive 15.2% drop in accuracy, as high spatial resolution is required to resolve small objects. However, on texture-centric tasks like MedM-

NIST and EuroSAT, performance is remarkably stable, dropping only 3.2%. This indicates that SHViT can be aggressively optimized for speed (using large strides) in medical and remote sensing applications without compromising accuracy, a flexibility not present in standard fixed-resolution CNNs.

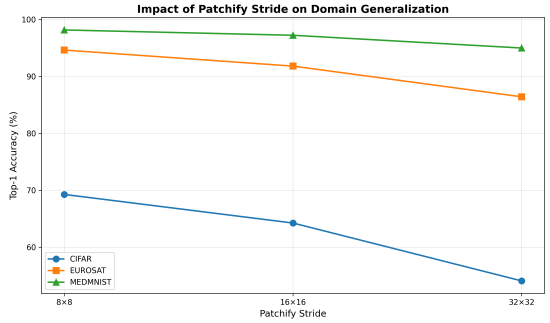


Figure 4: Impact of patch stride on domain generalization from CIFAR-100, EuroSAT and MedMNIST for SHViT variants.

Additional stride plots (accuracy details, efficiency curves, and sensitivity) are provided in Appendix D.

5.5 Qualitative Robustness: Saliency Under Noise

To better understand how architectural inductive biases influence robustness, we examine *where* each model attends when the input is corrupted with Gaussian noise. We generate saliency maps using Grad-CAM–style visualizations (Selvaraju et al., 2017) applied to the final feature representations of SHViT and DeiT-Tiny. These visualizations highlight the spatial regions most responsible for the model's predictions.

Qualitatively, SHViT exhibits a more stable and object-centric attention pattern under corruption. In many CIFAR-100 examples, SHViT-S2 continues to focus on the foreground object even when significant noise is added, enabling it to maintain correct predictions. In contrast, DeiT-Tiny frequently exhibits fragmented or background-dominated saliency, suggesting that its multi-head attention distribution becomes unstable under noise.

Figure 5 shows a representative case where SHViT remains robust and correctly identifies the object, while DeiT-Tiny's attention shifts toward noisy background patches, leading to misclassification. An additional example where the opposite occurs–SHViT fails while DeiT-Tiny remains stable–is provided in Appendix E, highlighting that the two architectures exhibit complementary inductive biases under corruption.

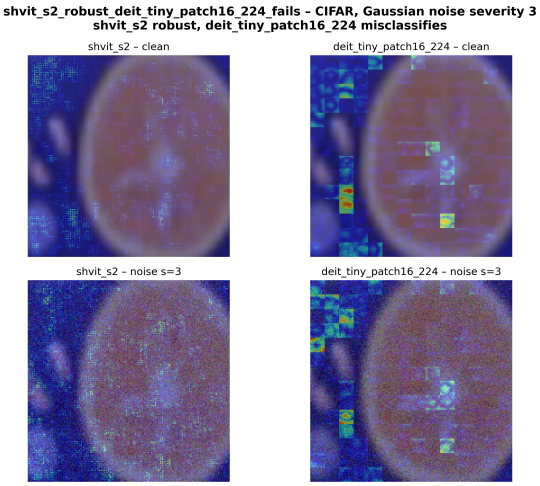


Figure 5: Saliency visualization under Gaussian noise using Grad-CAM. Example saliency visualization where SHViT remains object-centric and correct under noise while DeiT-Tiny drifts and fails.

6 Discussion

The success of SHViT in low-data and noisy regimes challenges the "scale is all you need" dogma. Our results suggest that the "Hybrid Sweet Spot"–combining a convolutional stem for early feature stability with a simplified global attention mechanism–creates a learner that is more robust than either of its parents. The Single-Head mechanism does not just save compute; it enforces a sparsity constraint that prevents the model from memorizing noise or over-fitting to texture, which explains its superior performance on EuroSAT and MedMNIST.

7 Conclusion

Our study shows that SHViT offers more than computational efficiency: its combination of a convolutional stem with single-head global attention yields architectural properties that improve robustness, sample efficiency, and cross-domain generalization. These findings suggest that simplified attention acts as an implicit regularizer, promoting stable spatial representations even under limited data or corruption. Taken together, SHViT occupies a promising middle ground between CNNs and standard ViTs, pointing toward hybrid architectures that achieve strong generalization without large-scale pretraining.

References

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Alex Krizhevsky, Geoffrey Hinton, and 1 others. 2009. Learning multiple layers of features from tiny images.

Sachin Mehta and Mohammad Rastegari. 2021. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41.

Seokju Yun and Youngmin Ro. 2024. Shvit: Single-head vision transformer with memory efficient macro design. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5756–5767.

Appendix

A Additional Learning Curves

This appendix includes learning curves for all three datasets. These complement the CIFAR-100 results presented in the main text. Overall, the trends are consistent across datasets, with SHViT maintaining competitive accuracy in low-data regimes.

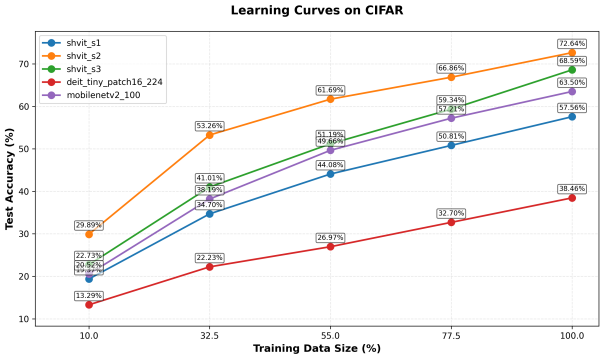


Figure 6: Learning curve on CIFAR-100 for MobileNet, DeiT-Tiny and SHViT variants.

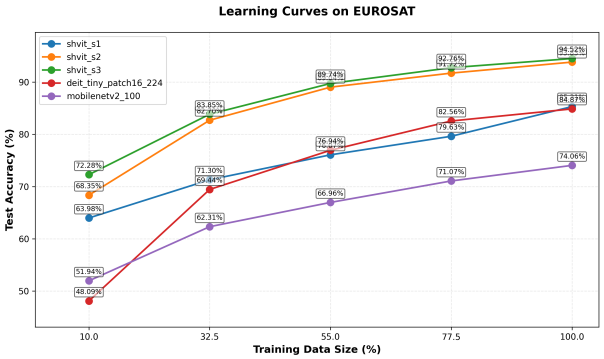


Figure 7: Learning curve on EuroSAT for MobileNet, DeiT-Tiny and SHViT variants.

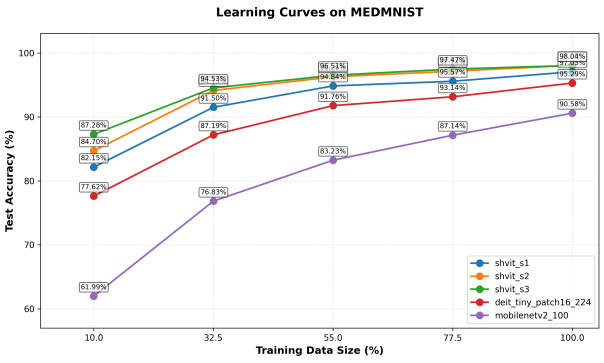


Figure 8: Learning curve on MedMNIST (PathMNIST) for MobileNet, DeiT-Tiny and SHViT variants.

B Corruption Severity Results

Figure 9 shows robustness under corruption severity levels 1, 3, and 5 respectively on CIFAR-100. These complement the single saliency example in the main text and demonstrate that SHViT often exhibits a smoother degradation pattern than DeiT-Tiny in many corruption categories.

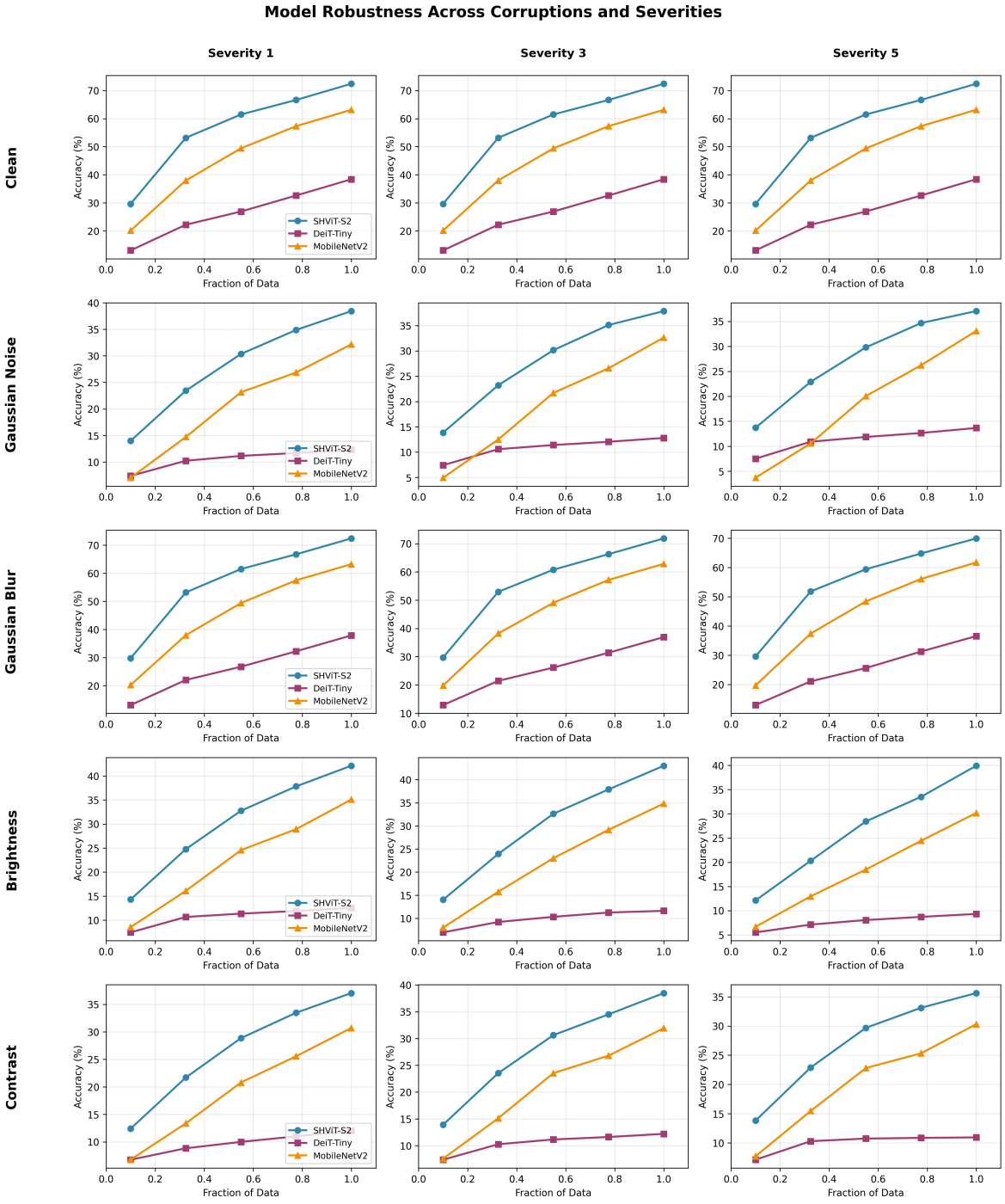
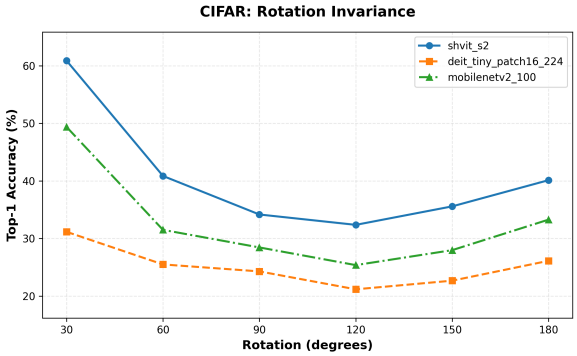


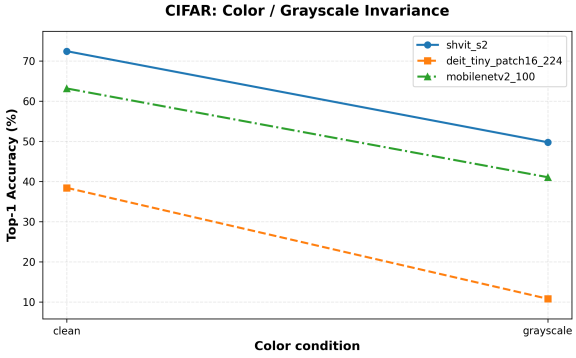
Figure 9: Comparison of robustness under various corruptions with severities 1,3,5 on CIFAR-100.

C Geometric and Color Robustness Results

In addition to Gaussian noise corruptions, we evaluate robustness to geometric and color transformations on CIFAR-100. These experiments probe how well the models handle changes in viewpoint and appearance that do not alter the underlying semantic content.



(a) Accuracy on CIFAR-100 under different rotation angles for MobileNet, DeiT-Tiny, and SHViT variants.



(b) Accuracy on CIFAR-100 under color perturbations (e.g., grayscale conversion or strong color jitter) for MobileNet, DeiT-Tiny, and SHViT variants.

Figure 10: Robustness analysis on CIFAR-100: (a) rotation robustness; (b) color perturbation robustness.

D Additional Patch Stride Results

Figure 4 in the main section shows stride effects on domain generalization. Here we include additional stride experiments including accuracy breakdowns and efficiency relationships.

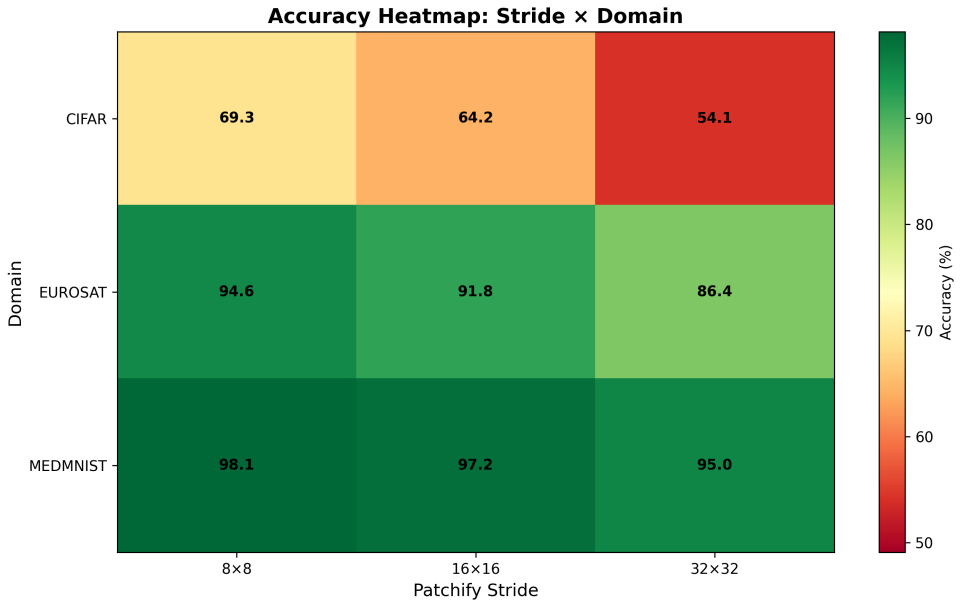
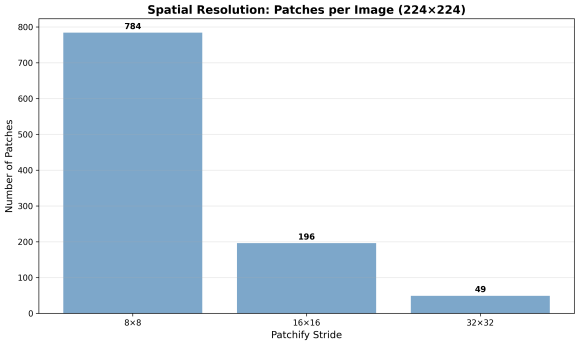
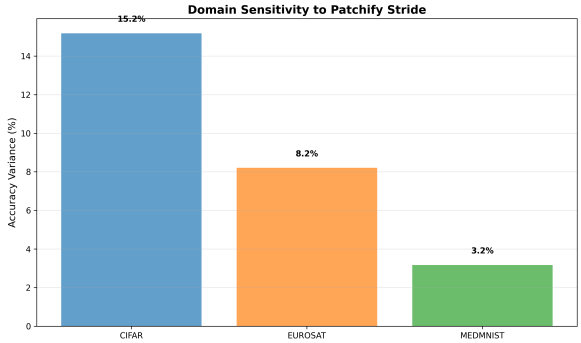


Figure 11: Additional accuracy trends for increasing patch stride.



(a) Inference efficiency under varying patch stride.



(b) Sensitivity plots showing the effect of stride on model behavior.

Figure 12: Analysis of patch stride effects: (a) inference efficiency measured through computational metrics; (b) model sensitivity showing behavioral changes across stride variations.

E Additional Saliency Visualizations

Finally, we include a complementary saliency example illustrating a failure case for SHViT-S2 under noise, where DeiT-Tiny maintains a more localized focus.

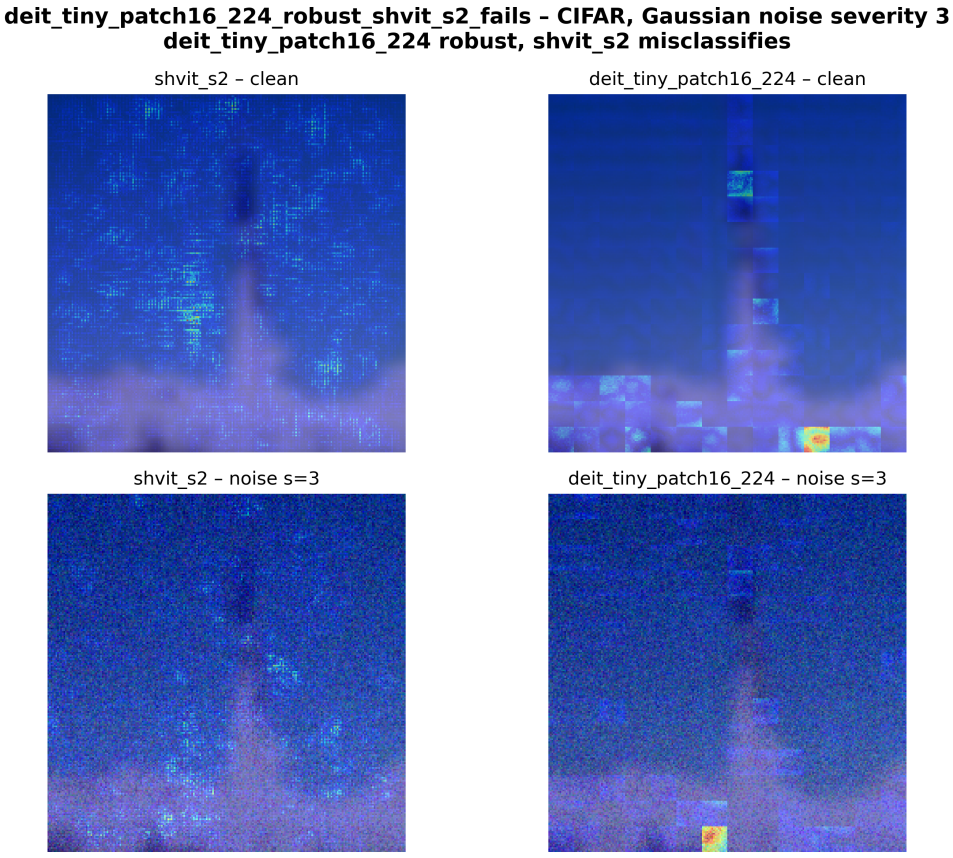


Figure 13: Example saliency visualization under noise where SHViT-S2 fails to maintain object-centric attention, while DeiT-Tiny exhibits a more stable focus and predicts correctly.