

A Computational Redundancy in ViTs

We investigate the computational redundancy in Vision Transformers (ViTs) from two complementary perspectives: *data-level* and *model-level* redundancy.

Data-level redundancy. We begin by evaluating the robustness of ViTs under partial observations of the input. Specifically, we conduct two types of perturbations: (1) randomly dropping a proportion of patch tokens before entering the transformer, and (2) randomly zeroing out elements in the attention weight matrices during self-attention computation. In both cases, the [CLS] token is retained. As shown in Figure 7, the top-1 accuracy on ImageNet remains remarkably stable even after removing up to 50% of tokens or attention weights. This indicates that ViTs possess strong resilience to incomplete or noisy visual evidence, likely due to the high degree of representational redundancy inherent in dense token embeddings and global attention.

Model-level redundancy. We further explore the internal redundancy of ViTs by ablating key components of the architecture at inference time. We consider: (1) randomly disabling a subset of attention heads in each layer, and (2) randomly dropping a proportion of hidden units in the intermediate layers of the feedforward network (FFN). As seen in Figure 7, both forms of perturbation lead to graceful degradation in performance. Even with 30–50% of heads or FFN neurons removed, the models still maintain high accuracy. This reinforces the observation that ViTs are significantly overparameterized, and many internal computations can be suppressed without compromising the final output.

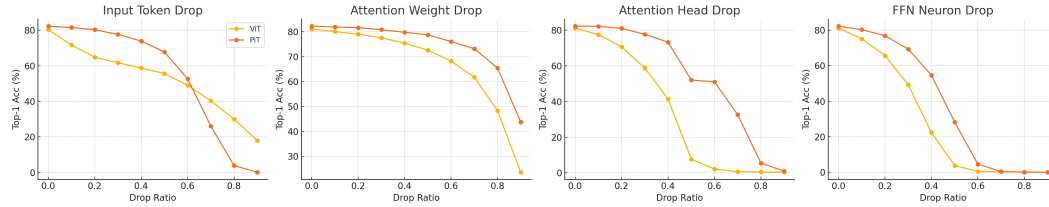


Figure 7: Top-1 accuracy of ViT under various types of token and structural drop perturbations. ViTs exhibit strong robustness to both input-level and architecture-level degradation, suggesting substantial redundancy in both data representation and model computation.

Related works on studying the computational redundancy of transformers and ViTs. There have been many works that systematically study and leverage the redundancy within the ViT’s architecture. For example, Bolya et al. [2022], Yin et al. [2022], Shang et al. [2024], Arif et al. [2025] find that dropping unimportant visual tokens or merging similar tokens will accelerate the inference of ViTs without harming the model performance. Jin et al. [2024], Fu et al. [2024], He et al. [2024] find that there exists similarity to some degree between different attention heads. Some works leverage the computational redundancy to enhance the performance of model, *e.g.*, the use of MoE [Lin et al., 2024, Chen et al., 2024a].