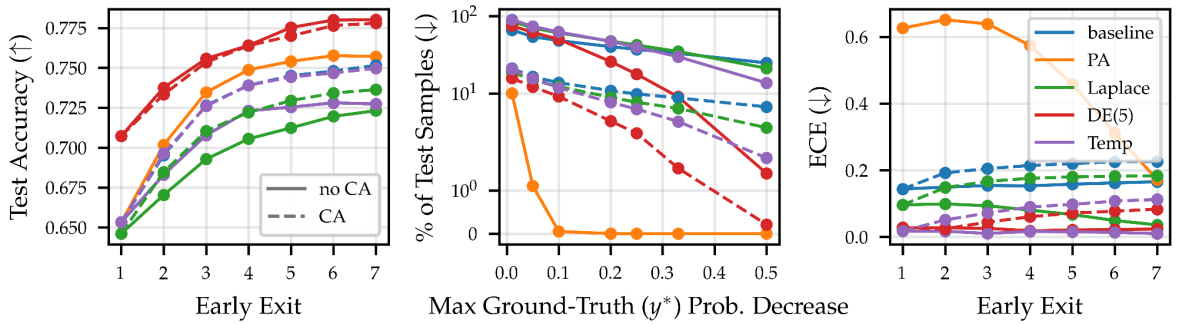

Fig R.1 (6rYN, kJKM, G2Sm, ne1V): Comparison of our PA with calibrated baselines using MSDNet as a backbone on CIFAR-100 dataset. For calibration, we use last-layer Laplace, deep ensembles with M=5, and temperature-scaling. Additionally, we consider both caching (dashed) and non-caching (solid) versions of calibrated baselines. None of the considered calibrated baselines outperforms our PA in terms of monotonicity (middle), despite being more complex.

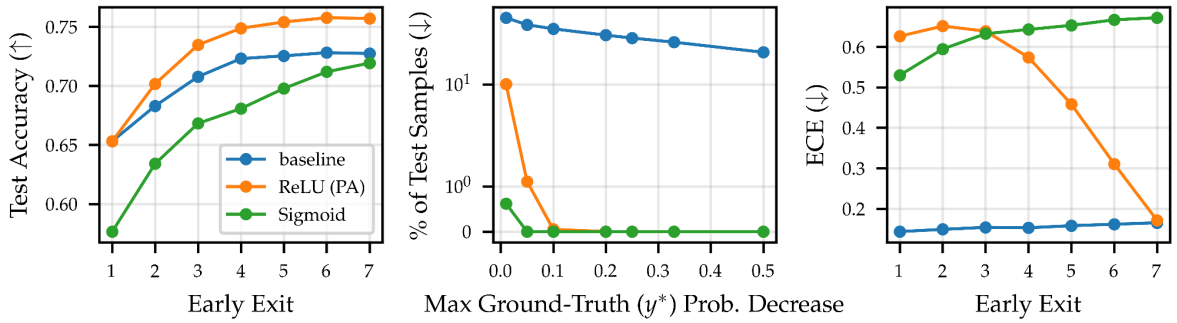

Fig R.2 (G2Sm): Results for sigmoid activation on CIFAR-100 dataset using MSDNet as backbone.

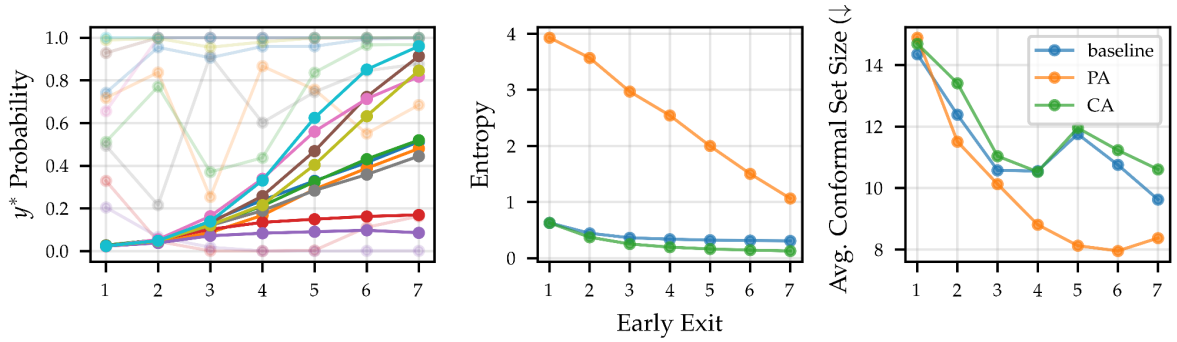

Fig R.3 (ne1V): We reproduce Figure 4 using IMTA model on CIFAR-100 dataset. Our findings from Section 6.2 remain consistent.

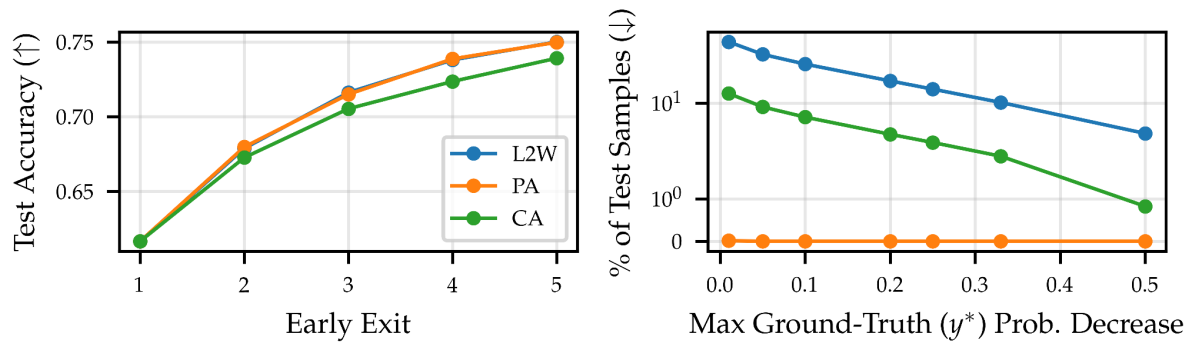

Fig R.4 (NWJy): Results for ImageNet using L2W model as a backbone. Our findings remain consistent, L2W without PA is significantly less monotone than with our PA.