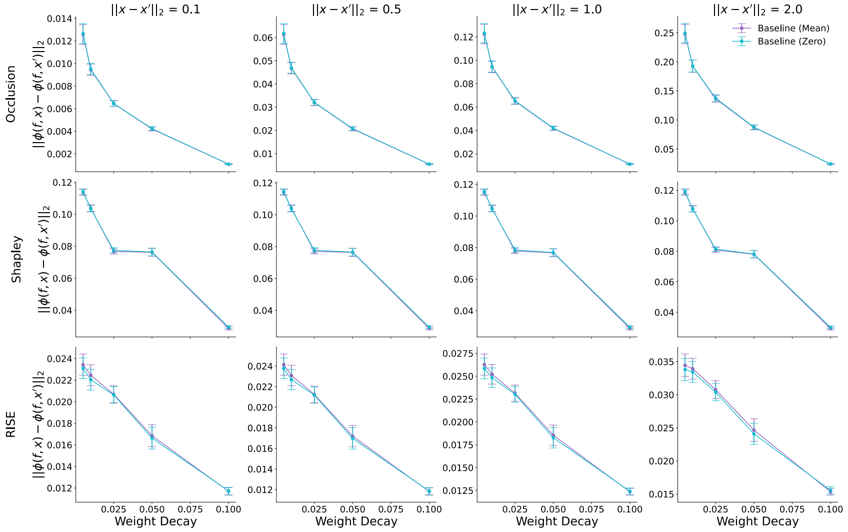

Figure R1: Attribution differences for ResNet-18 trained on CIFAR-10 with increasing weight decay. We show the means and 95% confidence intervals across explicand-perturbation pairs.

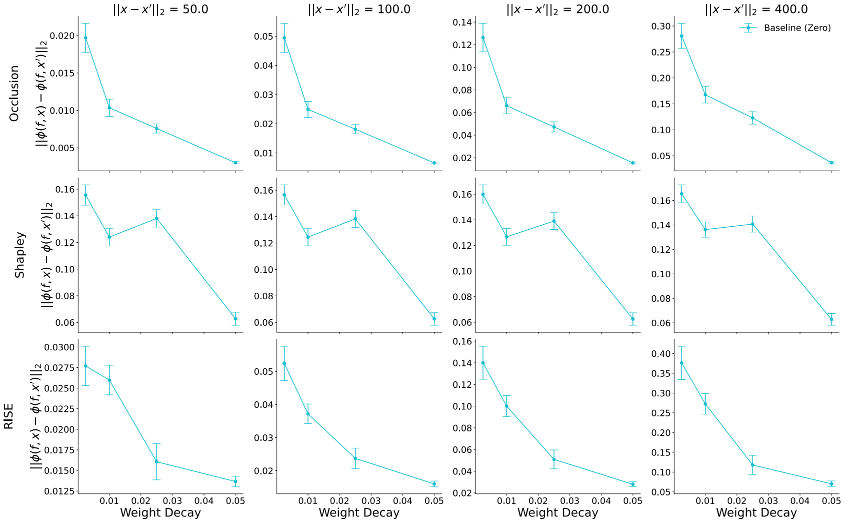

Figure R2: Attribution differences for ResNet-50 trained on 10 classes of ImageNet with increasing weight decay. We show the means and 95% confidence intervals across explicand-perturbation pairs.

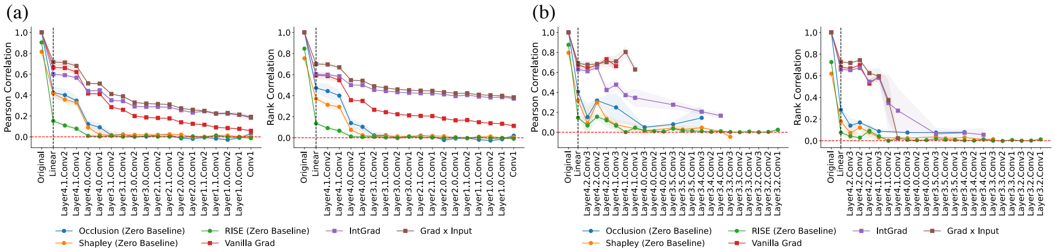

Figure R3: Sanity checks for attributions using cascading randomization for (a) ResNet-18 trained on CIFAR-10; and (b) ResNet-50 trained on 10 classes of ImageNet. Features are removed by replacing them with zeros. Missing points correspond to undefined correlations due to attributions with all zero values. We show the means and 95% confidence intervals across 5 random seeds.