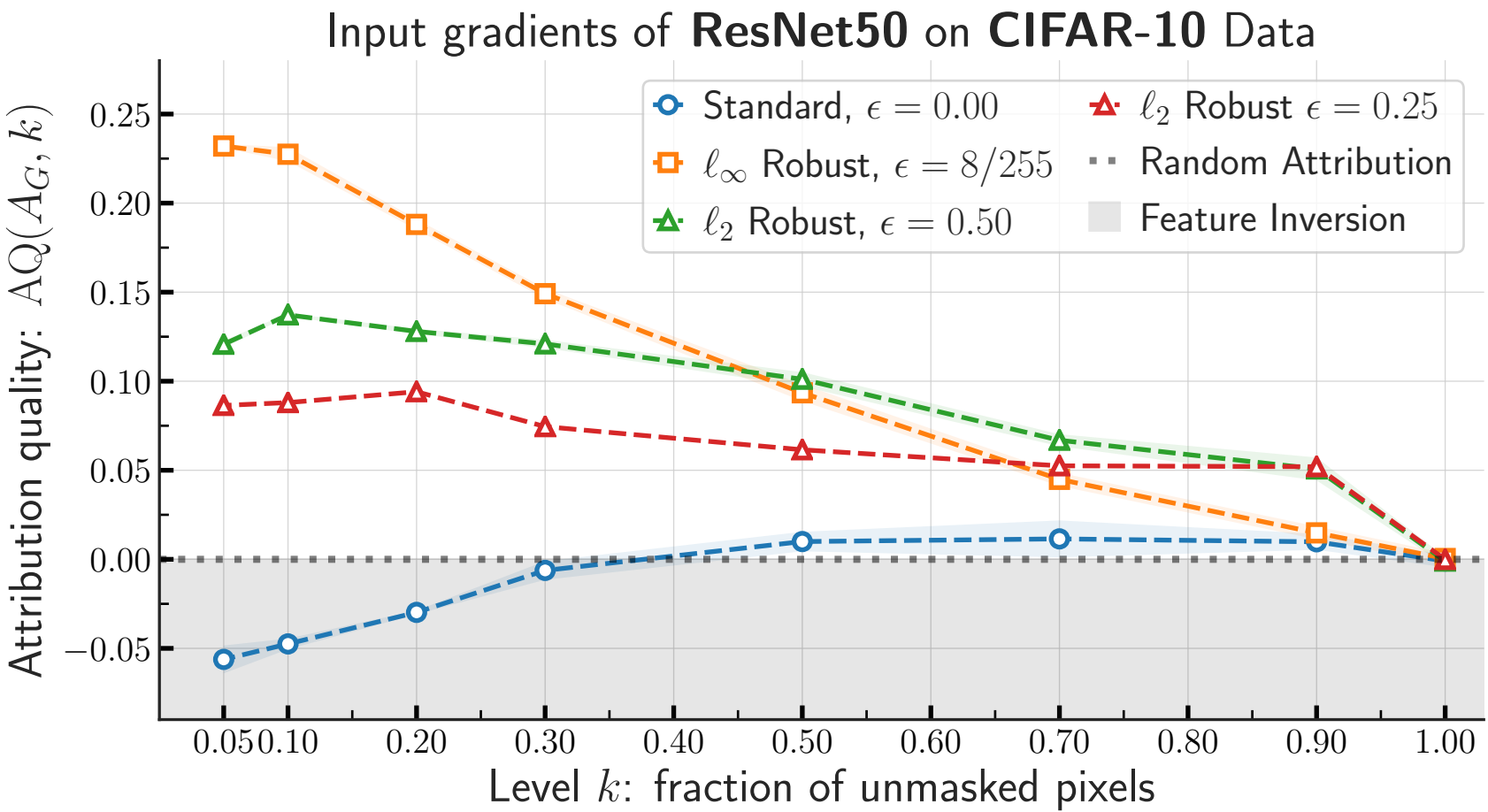
Input gradients of **ResNet50** on **CIFAR-10** Data
Standard, $\epsilon = 0.00$
ℓ_∞ Robust, $\epsilon = 8/255$
ℓ_2 Robust, $\epsilon = 0.50$
ℓ_2 Robust $\epsilon = 0.25$
Random Attribution
Feature Inversion
Attribution quality: $\mathrm{AQ}(A_G, k)$
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
0.05
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Level k: fraction of unmasked pixels