

Figure 6.2: Falsification Rates Across Attribution Methods
(n=100 pairs, K=100 counterfactuals, 95% CI)

