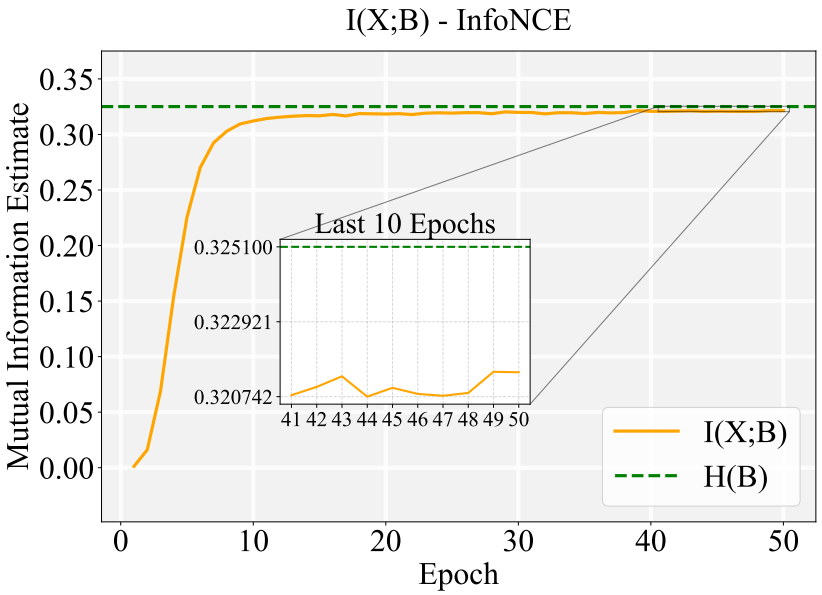

Figure R 1: Comparison of MI estimation between an upper bound $H(B)$ ($I(X;B) = H(B) - H(B|X) \leq H(B) = -0.1\ln(0.1) - 0.9\ln(0.9) \approx 0.3251$ nats) and $I(X;B) = 0.3208$ nats calculated by InfoNCE. The results implicates the image effectively encodes trigger information in X, with only $\sim 1.3\%$ estimation error.

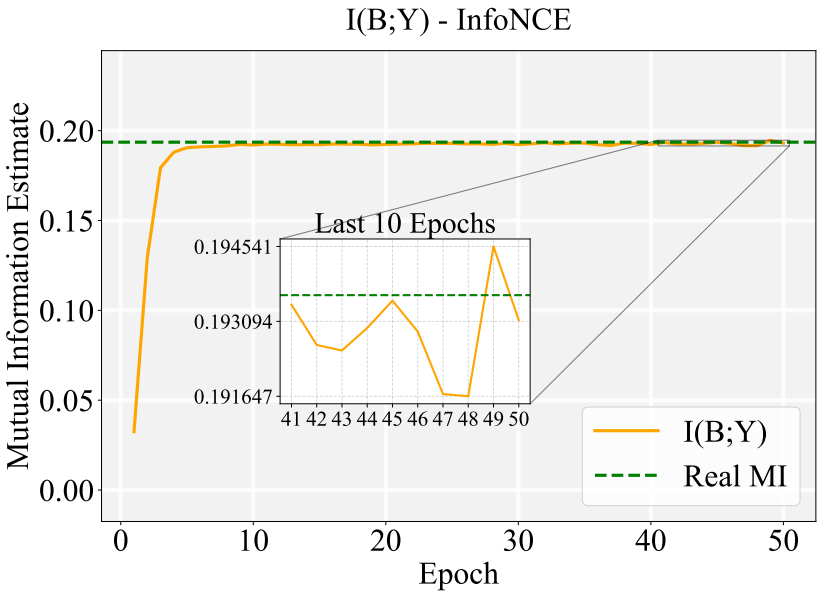

Figure R 2: Comparison of MI estimation between the real MI and $I(B;Y) = 0.1929$ nats calculated by InfoNCE. Since the distribution of B and Y is known, the real MI of $I(B;Y)$ can be directly obtained based on the formula of MI $I(B;Y) = \sum_{b,y} P(b,y) \ln \frac{P(b,y)}{P(b)P(y)} \approx 0.1936$ nats. The estimation error is $\approx 0.36\%$.

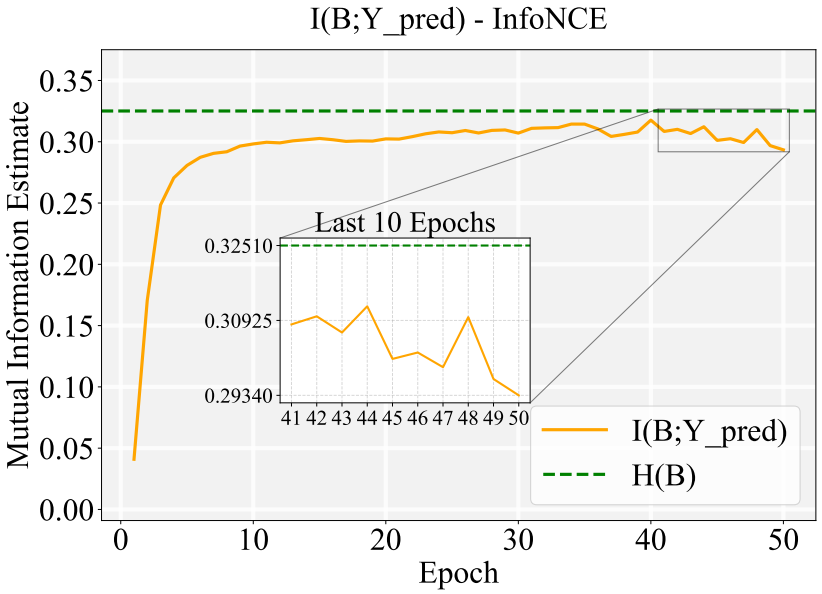

Figure R 3: Comparison of MI estimation between an upper bound $H(B)$ ($I(B;Y_{pred}) = H(B) - H(B|Y_{pred}) \leq H(B) \approx 0.3251$ nats) and $I(B;Y_{pred}) = 0.3025$ nats calculated by InfoNCE. The results implicates predictions strongly correlate with trigger presence. The gap reflects minor noise or partial reliance on semantic features.

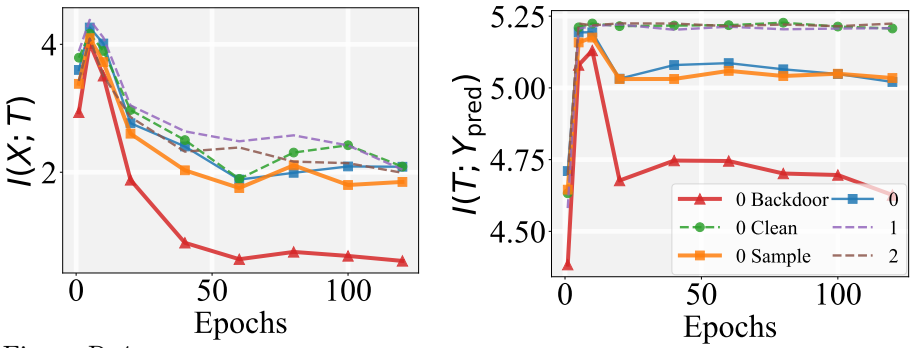

Figure R 4: MI dynamics under Blend attack on the CIFAR-10 dataset using a ResNet-18 model with a 10% poisoning ratio and $\gamma = 0.1$. The subfigures show $I(X;T)$ (left) and $I(T;Y_{\text{pred}})$ (right) across training phases. Global perturbations disrupt semantic features, leading to consistently lower $I(X;T)$ and $I(T;Y_{pred})$ for backdoor samples. While I(X;T) of backdoor samples still compresses significantly with lower noise (below 0.5 nats), it remains slightly higher than with $\gamma = 0.4$. Importantly, $I(T;Y_{pred})$ increases across all classes with lower noise. The result shows that while overall MI values increase with reduced noise (as expected), the pattern of compression for Blend backdoor samples persists, confirming this is an intrinsic property of how the network encodes and processes these global perturbations rather than an estimation issue or noise-driven effect.