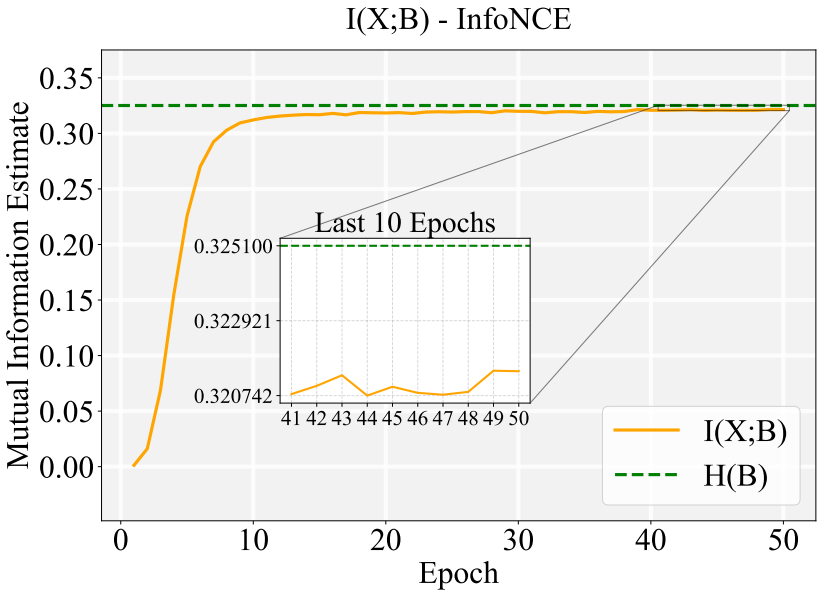

Figure R 1: Comparison of MI estimation between an upper bound $H(B)$ ($I(X;B) = H(B) - H(B|X) \leq H(B) = -0.1\ln(0.1) - 0.9\ln(0.9) \approx 0.3251$ nats) and $I(X;B)$ calculated by InfoNCE. The results implicates the image effectively encodes trigger information in X, with only $\sim 1.3\%$ estimation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

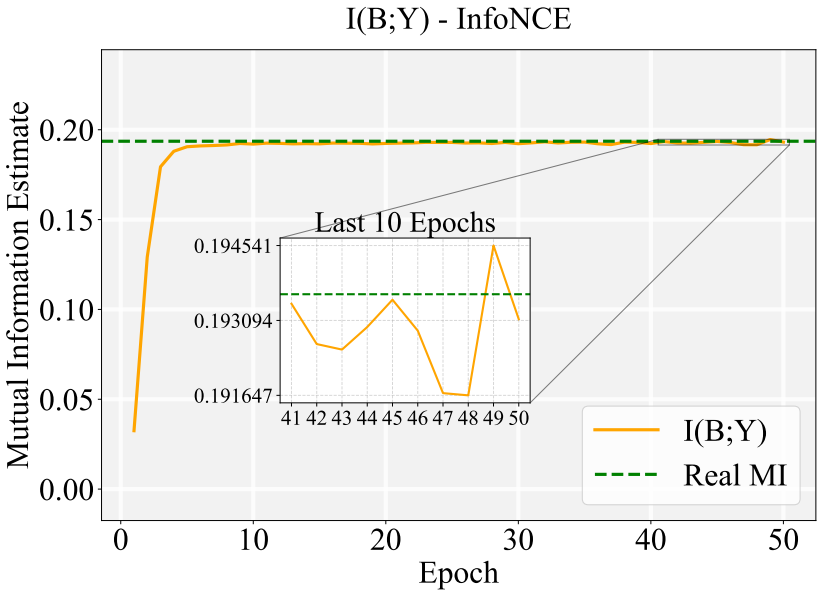

Figure R 2: Comparison of MI estimation between the real MI and $I(B;Y)$ calculated by InfoNCE. Since the distribution of B and Y is known, the real MI of $I(B;Y)$ can be directly obtained based on the formula of MI $I(B;Y) = \sum_{b,y} P(b,y) \ln \frac{P(b,y)}{P(b)P(y)} \approx 0.1936$ nats.. The estimation error is under $\sim 1\%$.

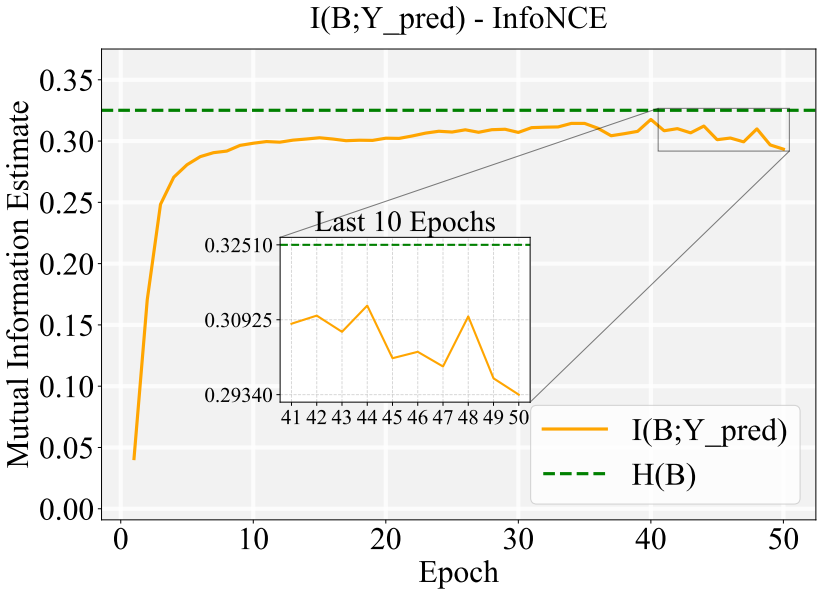

Figure R 3: Comparison of MI estimation between an upper bound $H(B)$ ($I(B;Y_{pred}) = H(B) - H(B|Y_{pred}) \leq H(B) \approx 0.3251$ nats) and $I(B;Y_{pred})$ calculated by InfoNCE. The results implicates predictions strongly correlate with trigger presence. The gap reflects minor noise or partial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

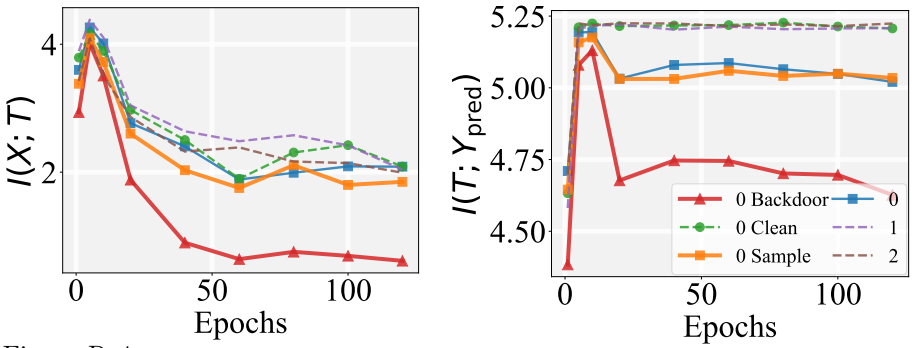

Figure R 4: MI dynamics under Blend attack on the CIFAR-10 dataset using a ResNet-18 model with a 10% poisoning ratio and $\gamma = 0.1$. The subfigures show $I(X;T)$ (left) and $I(T;Y_{\text{pred}})$ (right) across training phases. $I(X;T)$ exhibits the distinct two-phase behavior of backdoor samples and higher MI after the compression phased. $I(T;Y_{\text{pred}})$ of backdoor samples closely matches that of clean samples. This occurs because the presence of regularization samples forces the model to rely on more than just the trigger for predictions, requiring it to learn more diverse features to distinguish between regularized samples and backdoor samples that change the label to the target class. Consequently, similar to BadNets, the model learns both the trigger and the original semantic features from the backdoor samples' source classes.