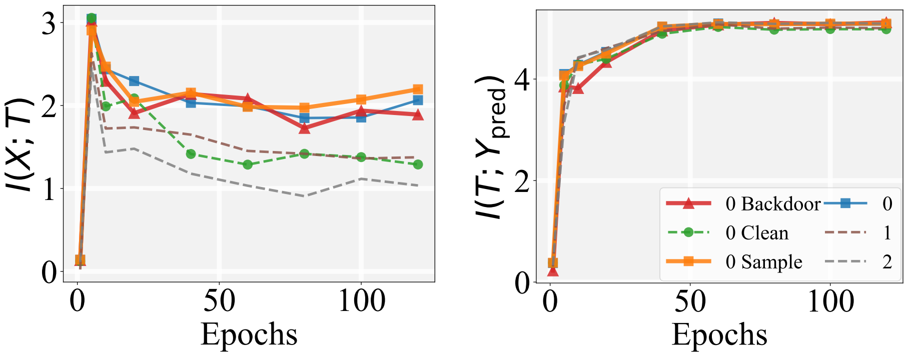
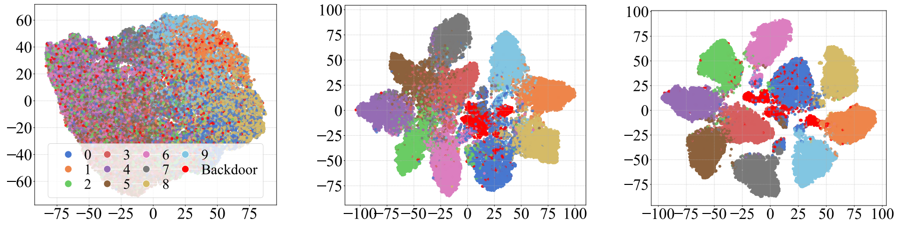


Figure R 1: MI dynamics under Adaptive Blend attack on the CIFAR-10 dataset using a ResNet-18 model with a 10% poisoning ratio and $\gamma = 0.4$. The subfigures show $I(X;T)$ (left) and $I(T;Y_{\text{pred}})$ (right) across training phases. $I(X;T)$ exhibits the distinct two-phase behavior of backdoor 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.



(a) Early learning phase (b) Compression phase (c) Convergence phase

Figure R 2: t-SNE visualization of the last hidden layer representations T under Adapive Blend attacks (CIFAR-10, 10% poisoning ratio). Backdoor samples form a distinct cluster with visible sub-clusters reflecting the model's dual representation of trigger and semantic features. Some backdoor samples merge with the target class while regularization samples cause clean classes to move closer to the backdoor samples in feature space.