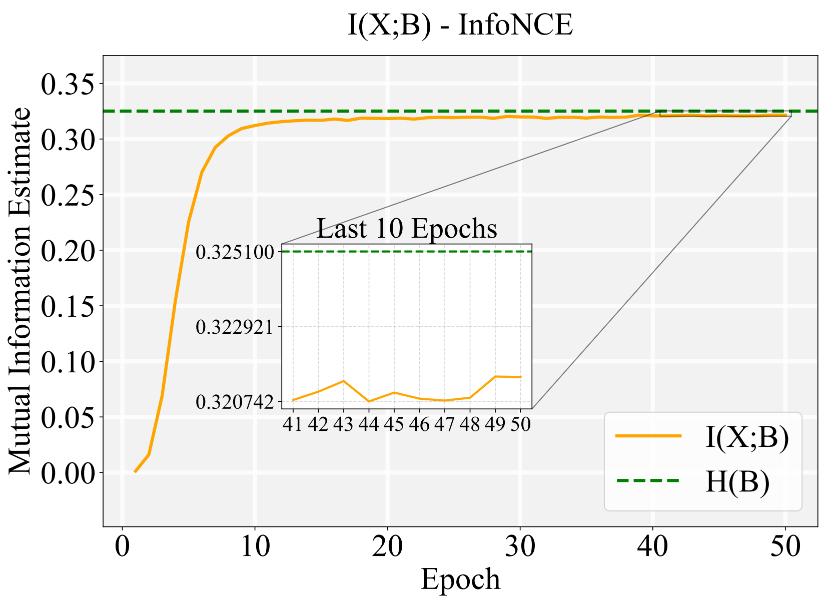

Figure R 1: Comparison of MI estimation between an upper bound $H(B)$ ($I(X;B) = H(B) - H(B|X) \leq H(B) = -0.1\ln(0.1) - 0.9\ln(0.9) \approx 0.3251\,\text{nats}$) and $I(X;B)$ calculated by InfoNCE. The results implicate the image effectively encodes trigger information in X, with only $\sim 1.3\%$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

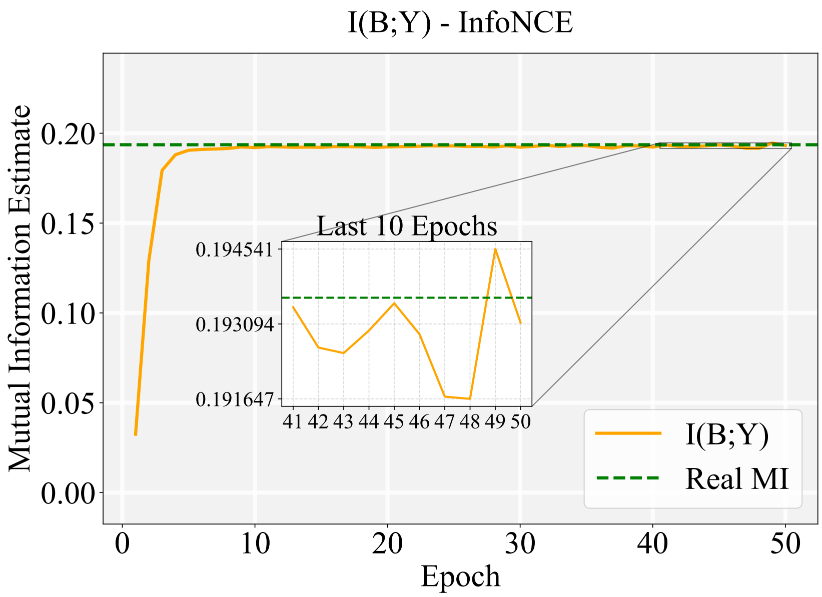

Figure R 2: Comparison of MI estimation between the real MI and $I(B;Y)$ calculated by InfoNCE. Since the distribution of B and Y is known, the real MI of $I(B;Y)$ can be directly obtained based on the formula of MI $I(B;Y) = \sum_{b,y} P(b,y) \ln \frac{P(b,y)}{P(b)P(y)} \approx 0.1936\,\text{nats}.$.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

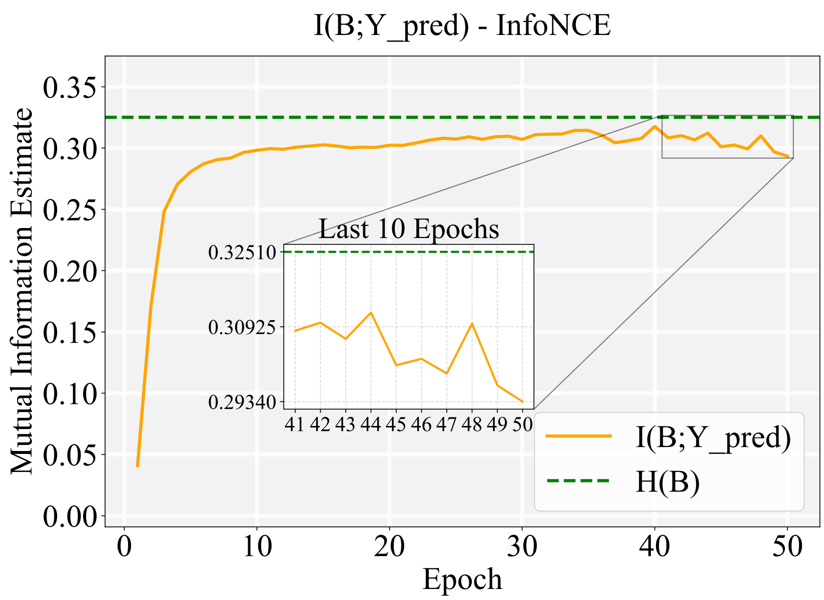

Figure R 3: Comparison of MI estimation between an upper bound $H(B)$ ($I(B;Y_{pred}) = H(B) - H(B|Y_{pred}) \leq H(B) \approx 0.3251\,\text{nats}$) and $I(B;Y_{pred})$ calculated by InfoNCE. The results implicate predictions strongly correlate with trigger presence. The gap reflects minor noise or partial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