



Figure 1: Conflict of attributions on six different attribution methods on different DNNs: BERT-large and LLaMA trained on the SST-2 dataset, and VGG-11 trained on CIFAR-10 and MNIST datasets. The ratio  $P(S) = \frac{|\phi(S) - \sum_{i \in S} \phi(i)|}{|\sum_{i \in S} \phi(i)| + |\phi(S)|}$  represents the conflict significance of the conflict of the coalition  $S$ . For each DNN, the figure shows the histogram of  $P(S)$  values of 2000 different coalitions extracted from 200 samples. All attribution methods exhibit highly conflicted attributions and it verifies that the attribution conflict is an intrinsic property of the attribution.

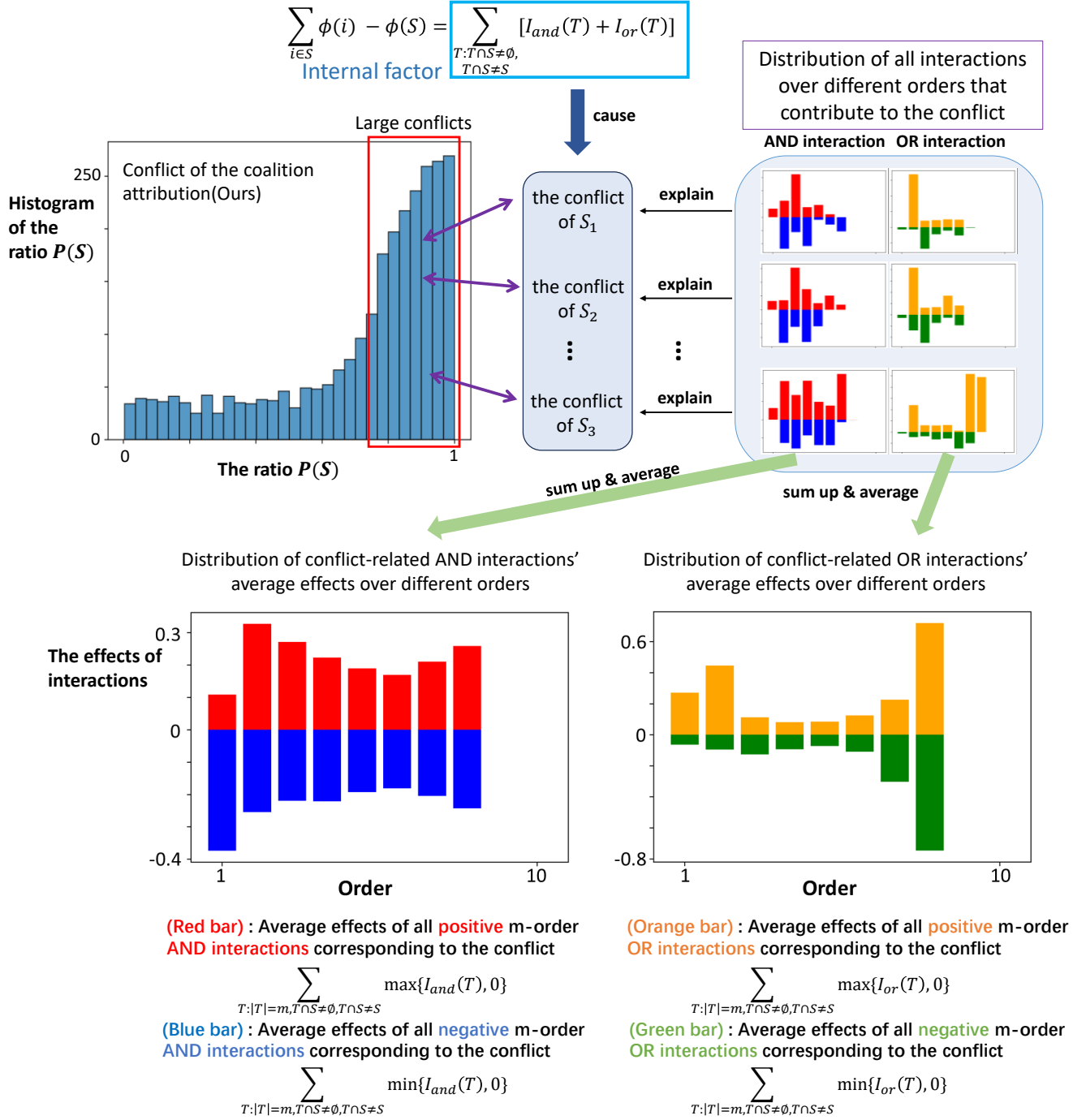


Figure 2: Internal factor that determines the attribution conflict. The conflict between individual variables' attributions and the attribution of the coalition  $S$  comes from numerical effects of all interactions  $T$  that contain just partial but not all variables in  $S$ , subject to  $\emptyset \neq T \cap S \neq S$ . The lower figure shows the distribution of average effects of all interactions over different orders that contribute to the conflict.