S1. Supplementary Material

In this section we provide supplementary materials.

S1.1. More visualizations

This section we demonstrate more visualizations of the input feature sensitivities. The parameter sensitivity visualization are not shown further due to the abstraction and complexity.

MNIST. The visualization of the input feature sensitivities for the MNIST handwritten dataset is illustrated in Fig. S4. It can be observed that the sensitivities of the explainability methods, except for GradCAM, are highly consistent with the prediction. This is consistent with the conclusions from the quantitative evaluation of this dataset in Sec. 4.2.

GTSRB. The sensitivity visualization for the GTSRB dataset is shown in Fig. S5. The sensitivity consistency of VB drops dramatically for this relatively more complex dataset and model, where its sensitivity map encompasses almost all input features. Similarly, the sensitivity maps of LRP and GradCAM fail to match the predictions, where the vast majority of the sensitive features of LRP are focused on the background, while GradCAM provides almost no feedback on feature perturbations. GB,IG and DeepLift maintain a relatively stable performance, with their sensitive areas mainly concentrated on the arrow symbols of the signboard, which is consistent with the sensitivity area of the prediction.

S1.2. Presentation of more experimental results

This section we show more experimental results, including the Top-k evaluation of the data SenC (Fig. S1). Also, we show the Top-1 and Top-3 agreement results for ImageNet in Fig. S2.

S1.3. Masking rate selections

The mask coverage is an adjustable hyperparameter when generating masks. Lower masking rates represent that the vast majority of the values on the matrix are retained, and conversely, the majority are perturbed. We find that the appropriate masking rate depends on the complexity of the model and data. As shown in Fig. S3, when experimenting on the simply structured MNIST dataset, we found that the optimal masking rate is between [0.2,0.6]. This is because a too high masking rate for MNIST may completely obscure the numbers and prevent the model from making predictions, while a too low masking rate makes it almost difficult to obscure the numbers, resulting in the model making almost constant predictions. However, the situation changes on ResNet18 and MobileNetV3, mainly due to the increased image complexity. It can be observed that as the complexity of the dataset increases, the performance of low masking gets significantly better and even exceeds that of high or medium masking. This is because the basis on which the model makes decisions grows richer, and masking some of the features may still be ineffective in interfering with the model predictions. Therefore, we recommend choosing an appropriate masking rate based on the complexity of the data and model.

S1.4. Layer-wise parameter SenC evaluation

This section we show the layer-wise parameter SenC for various models on different datasets.

MNIST. The layer- wise parameter SenC of MNIST can be seen in Fig S6. Since ModelCNN is simple in structure and contains only 4 layers, we exhibit the evaluation results for all layers. The sensitivity consistency of the first two convolutional layers is relatively high for all explainability methods, with IG having the best performance and GradCAM the lowest. For the last two fully connected layers, the explainability approaches are still slightly more consistent than the randomly masked baseline, despite the obvious gap with the convolutional layers. We believe that on the one hand fully connected layers cannot extract local features like convolutional layers such that the channels are not well-defined, and on the other hand fully connected layers are more tightly connected between channels and are more vulnerable to hard perturbations.

CIFAR-10. The layer-wise parameter SenC of ResNet18 trained on CIFAR-10 is displayed in Fig. S7. The structure of ResNet18 is complicated, therefore we only present the first convolutional layer, the last fully-connected layer and the intermediate layer belonging to "$layer1$". It can be observed that LRP and GradCAM almost collapse, with their consistency not differing significantly from the randomly masked baseline. In contrast, IG and DeepLift still maintain outstanding consistency, especially on layer conv1 and layer1.0.conv1, remarkably outperforming randomized baselines and other explainability methods.

GTSRB. The layer-wise parameter SenC of ResNet18 trained on CIFAR-10 is displayed in Fig. S8. We select the two convolutional layers in the first feature module and the last two fully connected layers of the final classifier for evaluation. The results are analogous to CIFAR-10, with IG and DeepLift outperforming the rest of the explainability methods, especially DeepLift, which still achieves an average SenC of more than 0.25 for the first two layers. Again, VB, LRP, and GradCAM exhibit almost no superior consistency over randomized explanations, increasing concerns that the identical parameter plays different roles in explanations and predictions.

S1.5. Processing time analysis

The processing times of SenC on each model are displayed in Table S1. The experiments are performed on an NVIDIA A100 GPU with the same number of perturbation

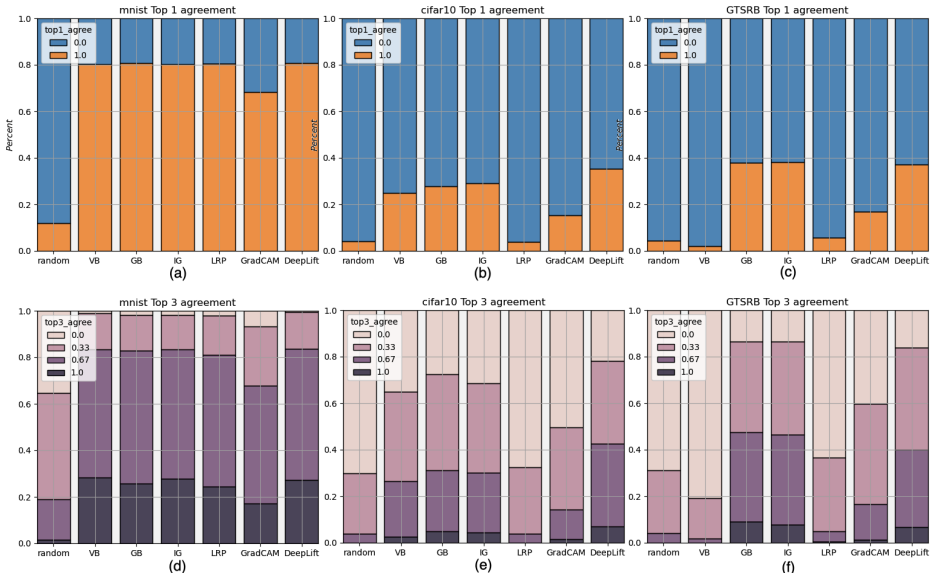

Figure S1. From up to bottom are Top-1 and Top-3 agreement, respectively. The x-axis in all plots represents different explainability methods. The y-axis in agreement indicates percentages. In Top-1 agreement, larger proportion of 1.0 (orange) fractions signify a higher percentage of agreement on the most sensitive features (better). In Top-3 agreement, higher percentage of darker color sections indicates better agreement.

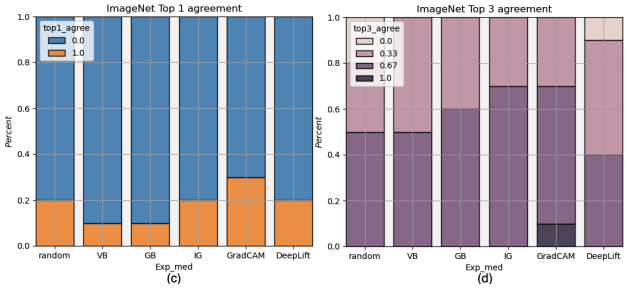

Figure S2. Top-1 and Top-3 evaluation results for ImageNet.

masks for both data and parameters of 5000 and 10000. We admit that SenC requires a significant amount of time to compute, especially for models with more complex structures such as ModelNetV3. However, at this stage, perturbations are the only way to analyze the relationship between explanations and predictions, since both models and explainability methods are themselves somewhat agnostic.

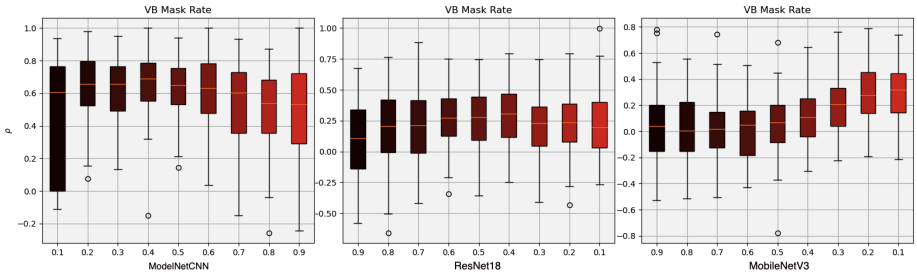

Figure S3. Impact of the masking rate on the evaluation results of different models. We show VB as examples, masking 10% to 90% of the parameters as perturbations for the same set of inputs, respectively, and recording the results of the final evaluation.

	ModelNetCNN	ResNet18	MobileNetV3
Data SenC	12.21 ± 0.37	46.14 ± 1.82	99.80 ± 1.91
Parameter SenC (per layer)	36.05 ± 0.33	186.95 ± 2.74	503.35 ± 9.59

Table S1. Processing time (second) of data and parameter SenC on ModelNetCNN,ResNet18 and MobileNetV3 respectively.

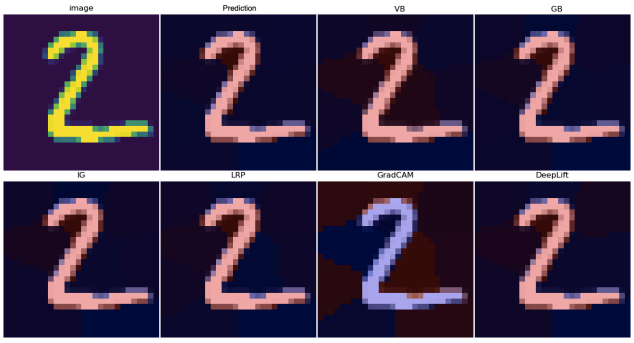

Figure S4. Visualization of input feature sensitivity for MNIST datasets.

Figure S5. Visualization of input feature sensitivity for GTSRB datasets.

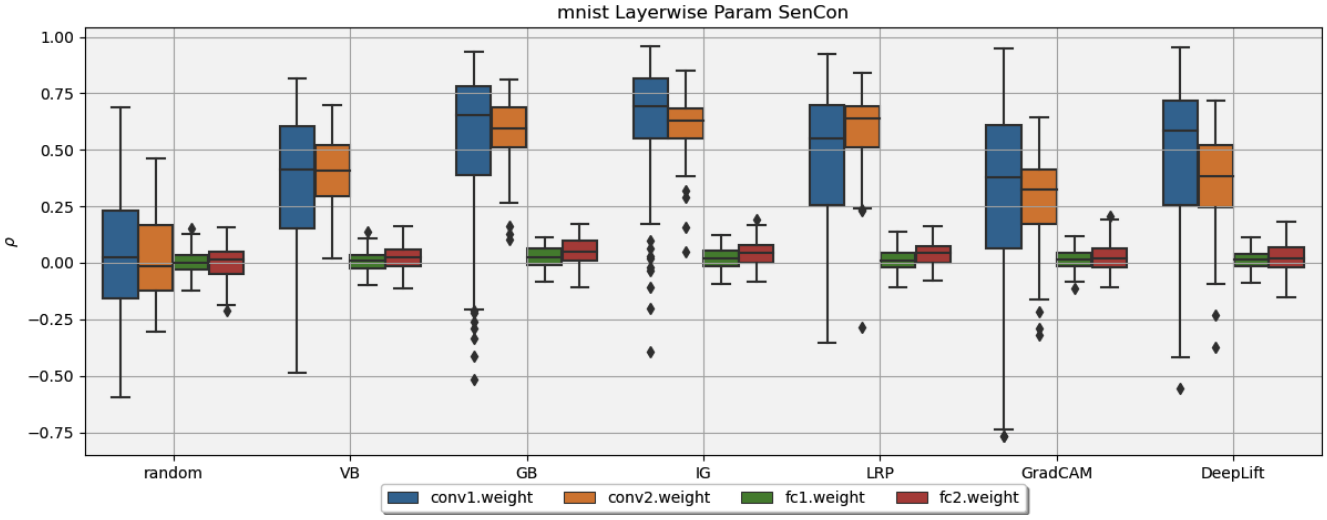

Figure S6. Layer-wise parameter sensitivity consistency assessment of ModelCNN trained on MNIST dataset. The x-coordinates are the different explainability methods, the y-coordinates are the Spearman correlation coefficients for the parameter sensitivities, and each box in the figure represents a specific layer.

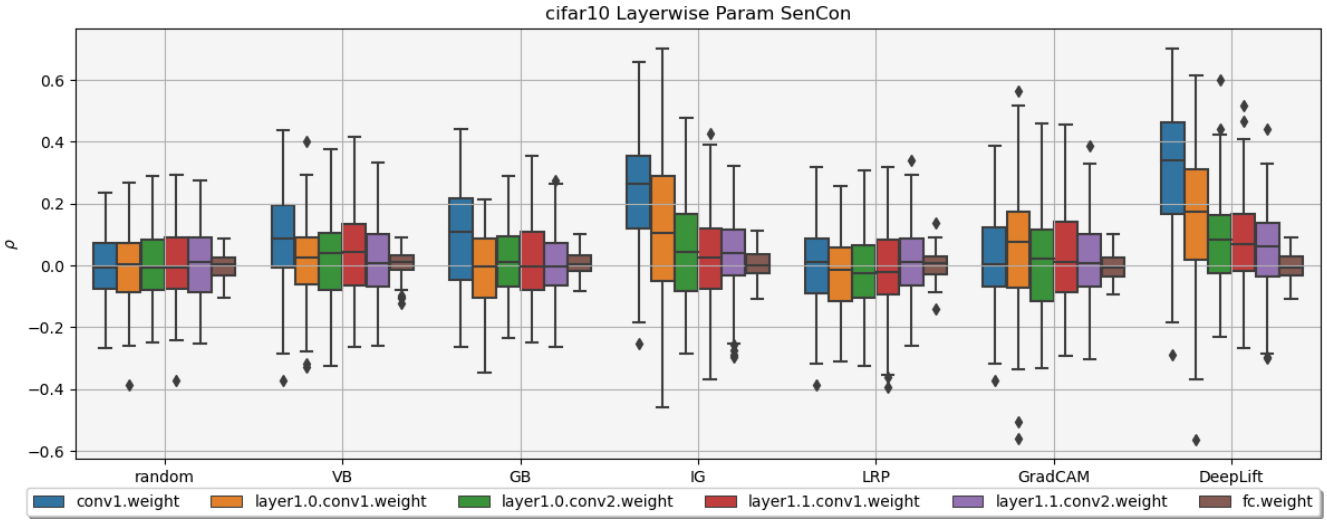

Figure S7. Layer-wise parameter sensitivity consistency assessment of ResNet18 trained on CIFAR-10 dataset.

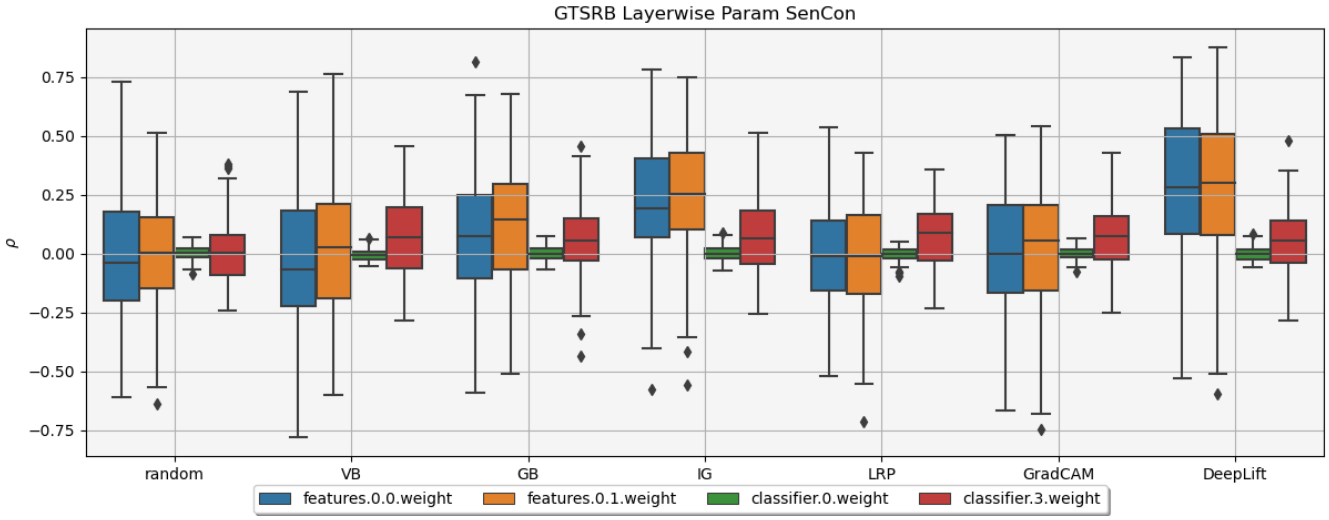

Figure S8. Layer-wise parameter sensitivity consistency assessment of MobileNetV3 trained on GTSRB dataset.