Supplementary material

1.1. Implementation details

We used two pretrained models in the experiments, specifically the ResNet-18 from Torchvision[2] [8] and ViT [4] from timm[3]. Both models were end-to-end fine-tuned with Stochastic gradient descent (SGD) by 50 epochs using a learning rate equal to 0.001 and batch size 32. In addition, we ablate the weight contribution of the Right answer factor (i.e., $\lambda 2$ in equation 1) for all non-standard methods and present the best result in the paper. Table S1 presents the $\lambda 2$ ranges used in the experiments.

Table S1. **Regularizer rate values used during training for the right for the right reasons methods.**

Method	$\lambda 2$
ActDiff	$[2^0, 2^3]$
GradMask	$[10^{-5}, 5 * 10^{-3}]$
RRR	$[10^1, 10^3]$

1.2. Interpretability is fragile

Section 3.6 present a discussion about interpretability fairness. It shows that models robust to the background do not necessarily imply attributing high importance to signal features. However, as the main text of the paper only presents the results for the Saliency interpretability method, here we extend the results, including the Integrated Gradient method in Figure S1. It corroborates with the findings in the paper and shows a different pattern for ViT on ImageNet-9. The signal-to-noise for ViT models is lower than the ResNet-18 when we use the Integrated Gradients, but when employing the Saliency, it is the inverse.

1.3. Correlation results

Figures S2 and S4 present the correlation between the accuracies of the challenges, as well as between the original and BG-Gap metrics. These results are important to help us understand the implications of the findings presented in Table 1. In summary, this analysis indicates that high original accuracy is not an indication of background robust models, and the classification models behave in the same way when processing original and mixed same images, as well as mixed rand and mixed next images.

1.4. On the relation between original accuracy and BG-Gap

Figure S3 show the relationship between performance in the Original scenario and BG_Gap for each dataset. Each bar in the chart represents a method, and its height indicates the BG_Gap rank for that method. The red dots represent the performance rank in the original scenario for each method. For each chart, we can observe the relationship between the original rank and BG_Gap for all methods in a specific dataset. This provides a direct visualization of the trade-off between model accuracy in the original scenario and its robustness to changes in the background, as indicated by the BG_Gap value.

In particular, if a method has a higher performance rank in the original scenario (indicated by a lower red dot) and a higher BG_Gap value (indicated by a taller bar), it suggests that the method may be overestimating the background of the images.

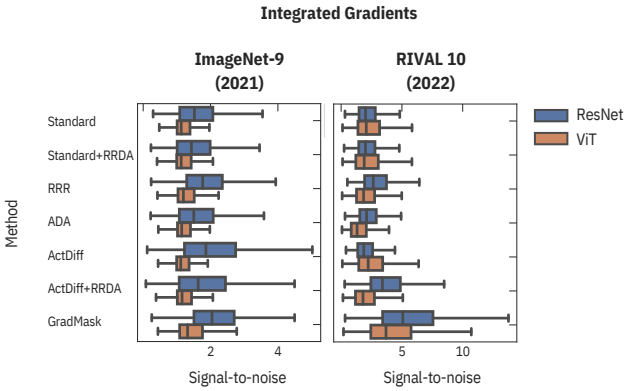

Figure S1. **Analysis of Signal-to-Noise Ratio.** This pipeline follows the same steps as in Figure 6. However, this scenario uses the Integrated gradients [22] interpretability method instead of Saliency [20].

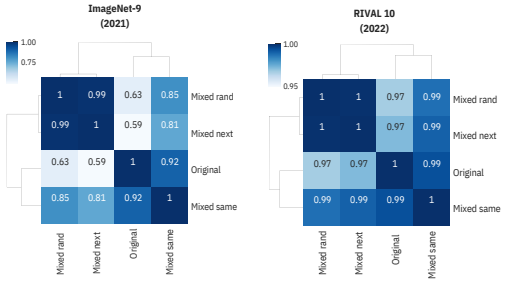

Figure S2. **Correlation between the challenges.** For each dataset, we compute the Person correlation between the challenge results. It shows a positive correlation between all pairs, but the correlation between Mixed same and Original, and Mixed rand and Mixed next are the higher values in both datasets.

[2]ResNet18 model from https://pytorch.org/vision/main/models/resnet.html

[3]vit_base_patch16_224_in21k model from https://github.com/huggingface/pytorch-image-models

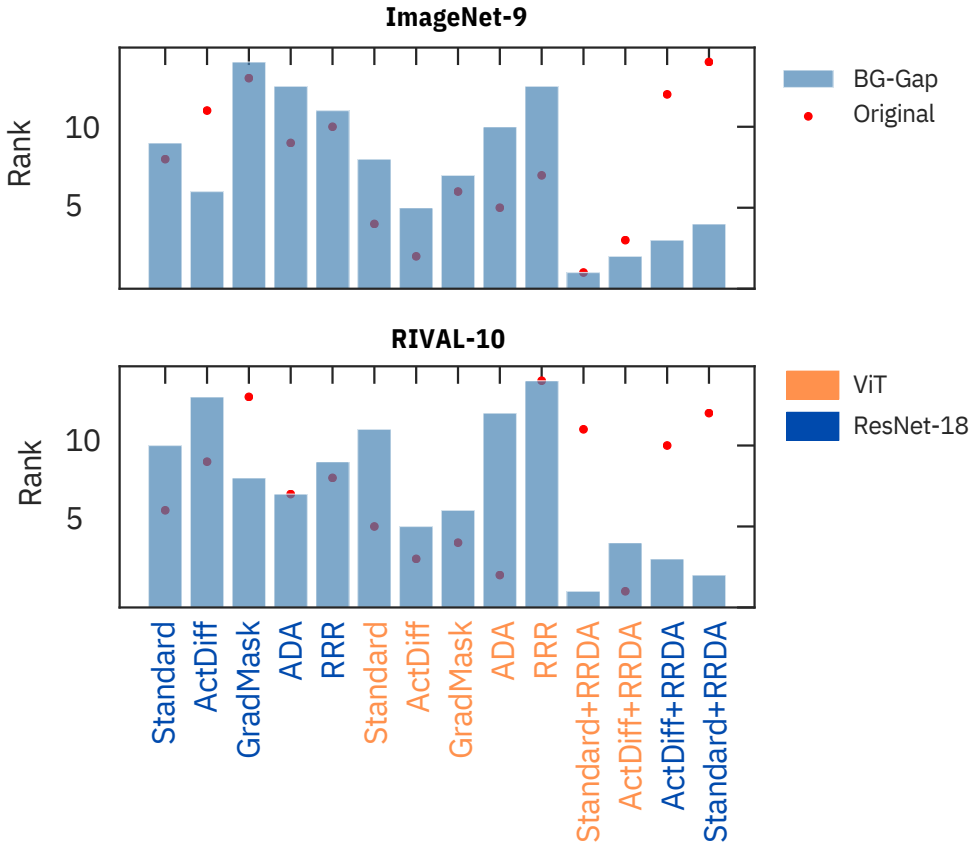

Figure S3. **Relation between original accuracy and BG-Gap.** We rank all method's accuracy on the original test set and its BG-gap. The visualization allows us to verify if some methods are overestimating the background of the images.

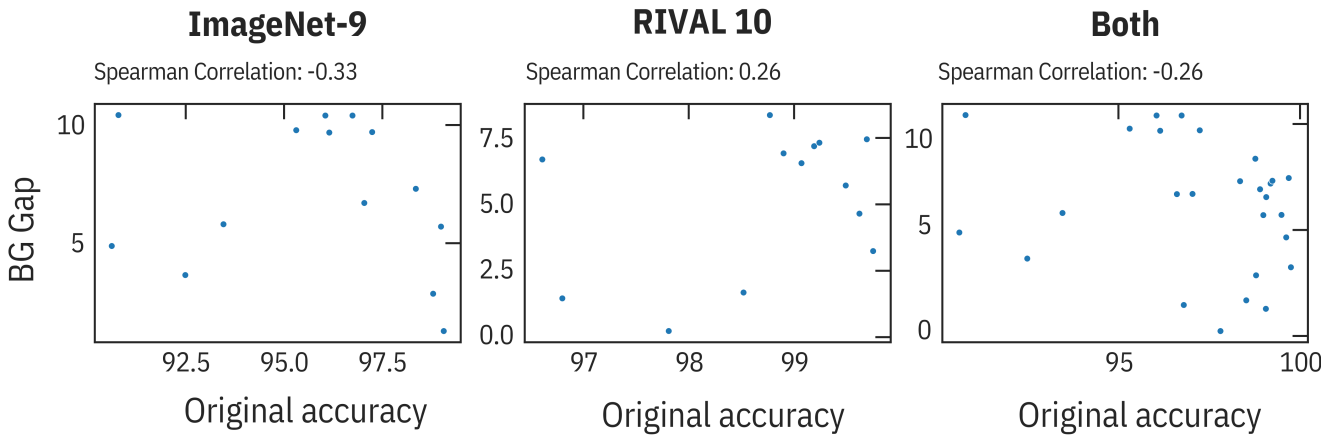

Figure S4. **Correlation between BG-Gap and Original accuracy.** We compute the Spearman correlation between all original accuracies higher or equal to 80% and BG-Gaps for each dataset scenario. In addition, we concatenate them both and compute the Spearman correlation (i.e., Both' scenario). The results do not present a strong positive or negative correlation between the values in all scenarios.