

EXAMINING WHY PERTURBATION-BASED FIDELITY METRICS ARE INCONSISTENT - SUPPLEMENTARY MATERIAL

S1 CAM DISAGREEMENTS

The illustration in Figure S1 presents saliency maps on randomly sampled images from the CIFAR-10, Imagenette, Oxford-IIIT Pets and PASCAL VOC 2007 datasets for pretrained ResNet50 model (imagenet weights) using AblationCAM (Ramaswamy et al., 2020), GradCAM++ (Chattopadhyay et al., 2018) and GradCAM (Selvaraju et al., 2017). It can be noted from Figure S1, the saliency maps generated using AblationCAM and GradCAM++ show a high degree of agreement, highlighting the importance of the body, neck, and head of the horse for an image from Cifar-10 dataset (1st row). However, the saliency map generated using GradCAM completely misses highlighting the head of the horse. In the 2nd row it can be observed that AblationCAM and GradCAM++ highlight not only the head of the fish but also other areas in the background as compared to GradCAM. Similarly, the saliency maps generated for the Oxford-IIIT Pets dataset image (3rd row) and PASCAL VOC 2007 image (4th row) show high inconsistency.

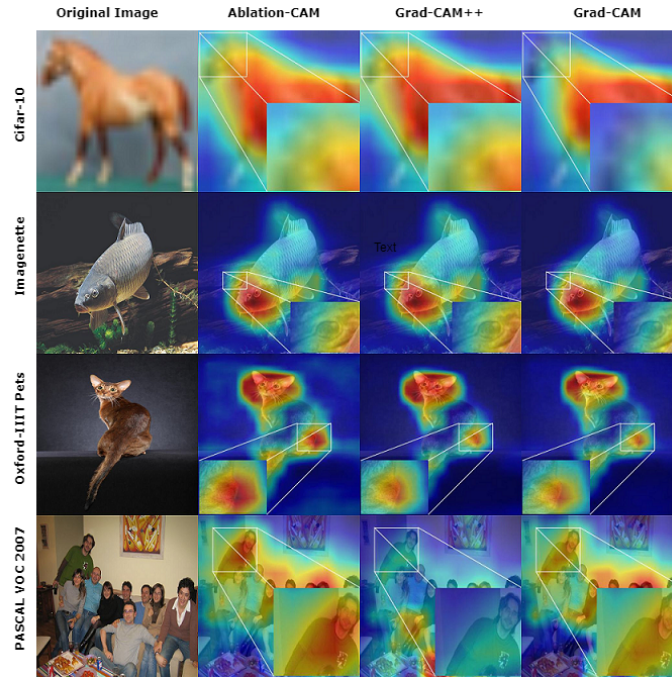


Figure S1: Disagreement between saliency maps generated using Ablation-CAM, Grad-CAM++ and Grad-CAM for ResNet50 model with imagenet weights. Each row represents a randomly chosen image from CIFAR-10, Imagenette, Oxford-IIIT Pets and PASCAL VOC 2007 datasets and their corresponding saliency maps

S2 PIXEL SELECTION AND RANKING

Selection of pixels for our analysis is another critical aspect for our analysis. As the size of the input images are typically 299×299 , 224×224 or 600×600 pixels for models, it is computationally expensive to conduct an analysis on all pixels. We therefore conduct our analysis on a subset of pixels which were randomly selected (based on (Tomsett et al., 2020)). Our approach to randomly select the pixels can be further justified from a theoretical perspective as explained below.

Let Q be a set of pixels such that $|Q| > 1$. We can define a hypothetical function $\psi(Q)$ that measures the importance of Q for the decision-making process of the model as:

$$\psi : Q \rightarrow \{1, 2, \dots, |Q|\} \subseteq \mathbb{R}$$

where \mathbb{R} is the set of all real numbers and a greater value of $\psi(Q)$ indicates greater importance.

We can define an image \mathbb{A} as an ordered set of pixels sorted according to their importance using function ψ .

$$\mathbb{A} = \{a_1^u, a_2^v, a_3^w, \dots, a_i^z\} \quad (10)$$

where, R_0 is the ordered set of pixels. $1 \rightarrow i$ are importance for the pixel index/ids $u \rightarrow z$ generates by ψ i.e. $\psi(a^u) = 1, \psi(a^v) = 2 \dots \psi(a^z) = i$ etc, where a greater value of $\psi(Q)$ indicates greater importance of the pixel set Q in the image.

Let us assume that \mathbb{B} is a randomly selected subset of pixels. Thus \mathbb{B} can be defined as below:

$$\begin{aligned} \mathbb{B} &= \{a_1^x, a_2^y, a_3^z, \dots, a_j^n\} \subseteq \mathbb{A} \quad \text{s.t.} \\ a^e &\neq a^f \quad \text{for } e \neq f \end{aligned} \quad (11)$$

where e and f are two random pixels. Let us assume that the order of pixels in \mathbb{A} and \mathbb{B} are different. This implies according to induction:

$$\begin{aligned} \exists \quad (a^p, a^q) &\in \mathbb{B} \quad \text{s.t.} \\ \psi(a^p) &> \psi(a^q) \in \mathbb{B} \quad \wedge \quad \psi(a^p) < \psi(a^q) \in \mathbb{A} \end{aligned} \quad (12)$$

However, $\psi(a^p) > \psi(a^q) \in \mathbb{B}$ and $\psi(a^p) < \psi(a^q) \in \mathbb{A}$ cannot be true at the same time, we can by mathematical induction deduce that $\nexists (a^p, a^q) \in \mathbb{B}$ that satisfy both conditions given in Equation (12). As such the order of pixels as per their importance are same in both \mathbb{A} and \mathbb{B} . We leverage this property that the order of importance of the pixels do not change even in randomly selected (without repetition) subsets for our analysis. If the selected pixels have the same importance ranks, their relative orders are not considered to affect the rank correlation.

S3 DROP SCORES FOR ALL DATASETS

S4 PAIRWISE DISTRIBUTION PLOTS FOR PERTURBATIONS

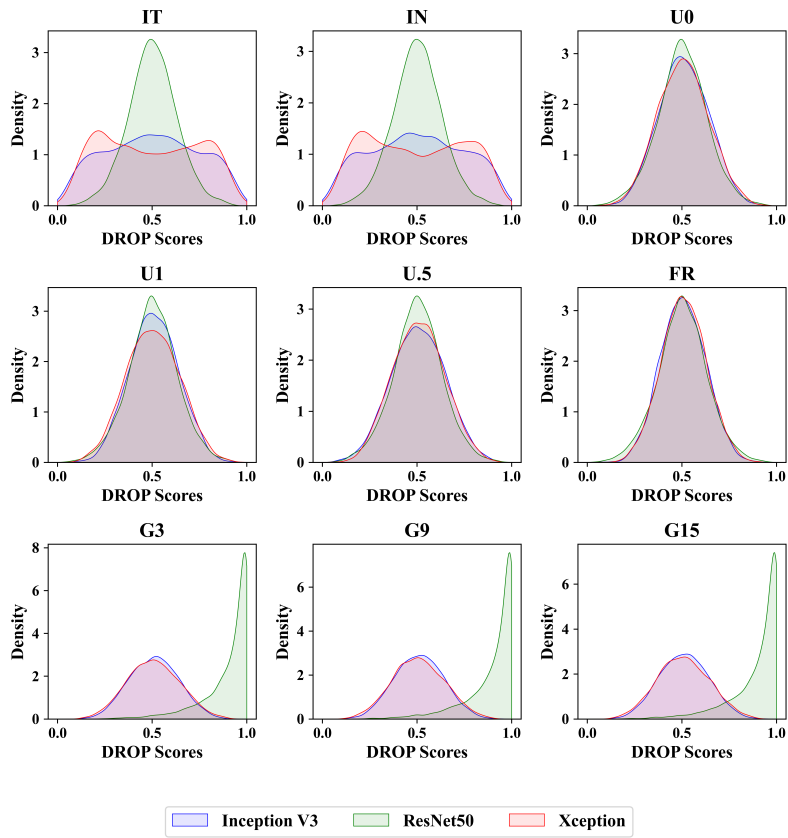


Figure S2: Distribution of *DROP* scores for all perturbations for Oxford-IIIT Pets Dataset

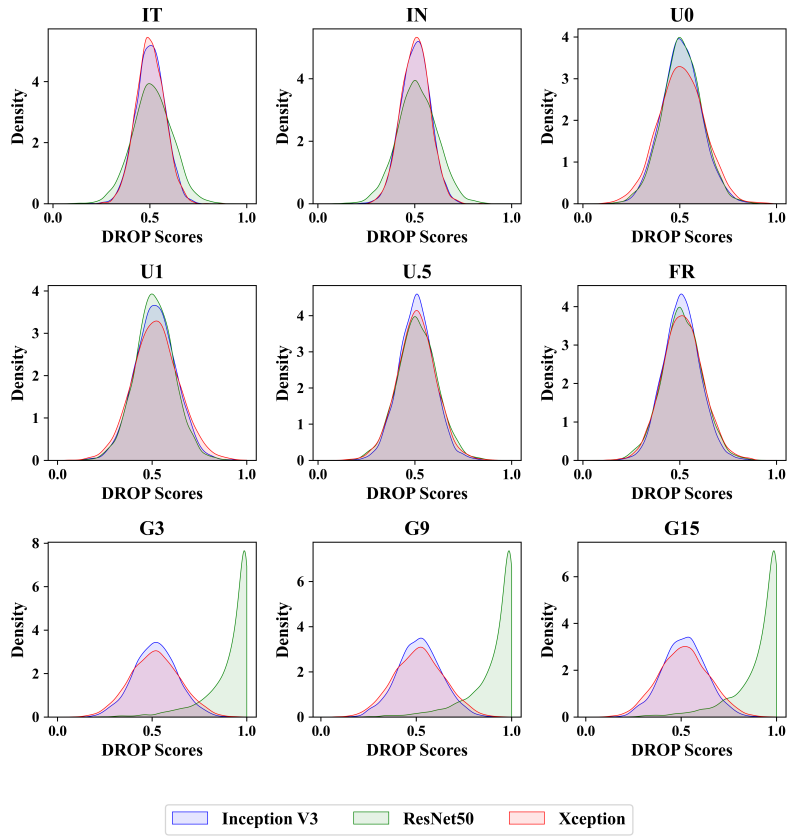


Figure S3: Distribution of *DROP* scores for all perturbations for PASCAL VOC Dataset

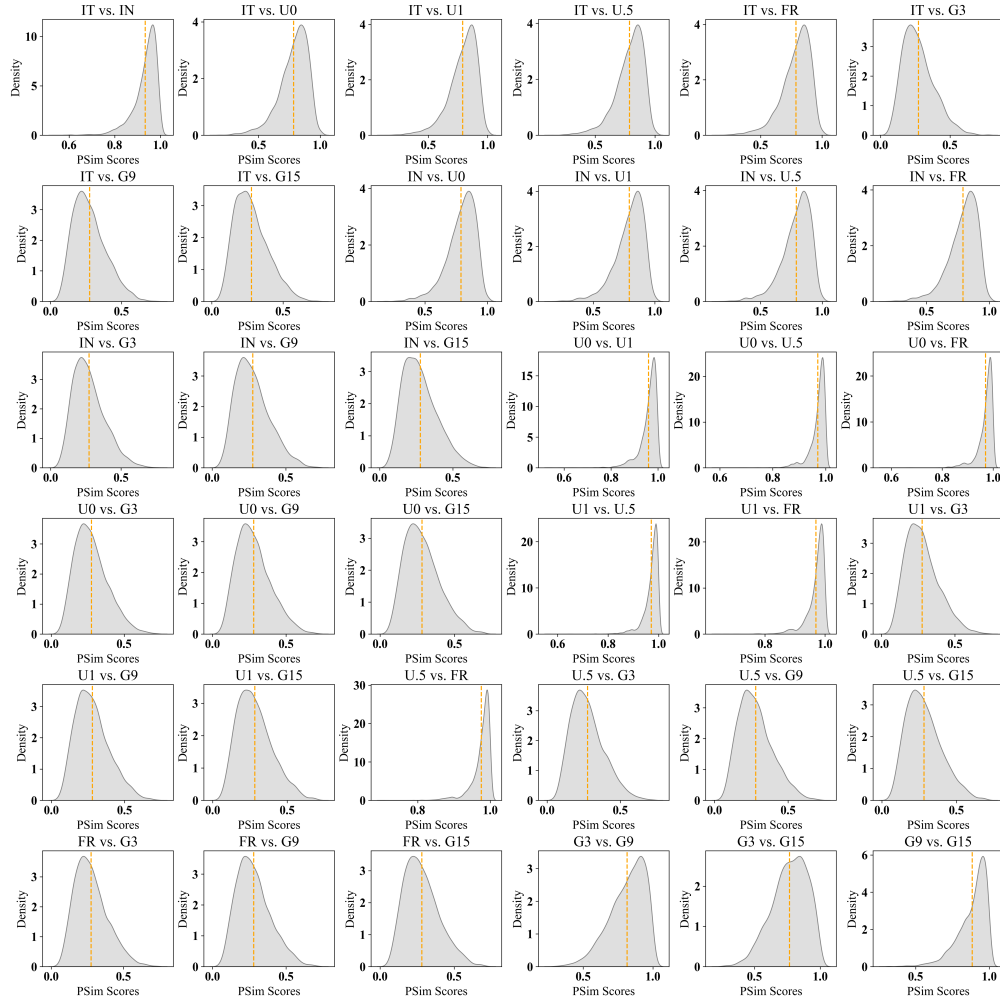


Figure S4: Distribution of pairwise $PSim$ scores for all perturbations for Resnet50 model on Imagenette Dataset

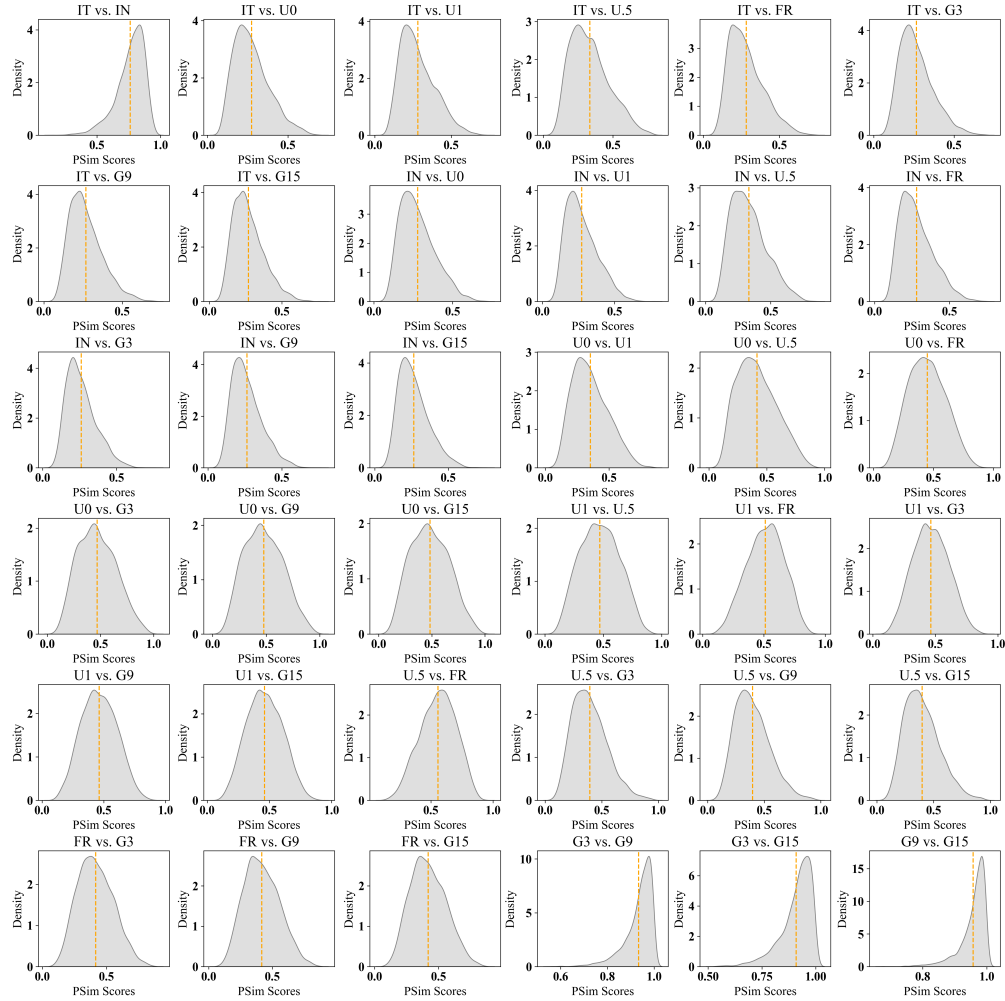


Figure S5: Distribution of pairwise $PSim$ scores for all perturbations for Xception model on Imagenette Dataset

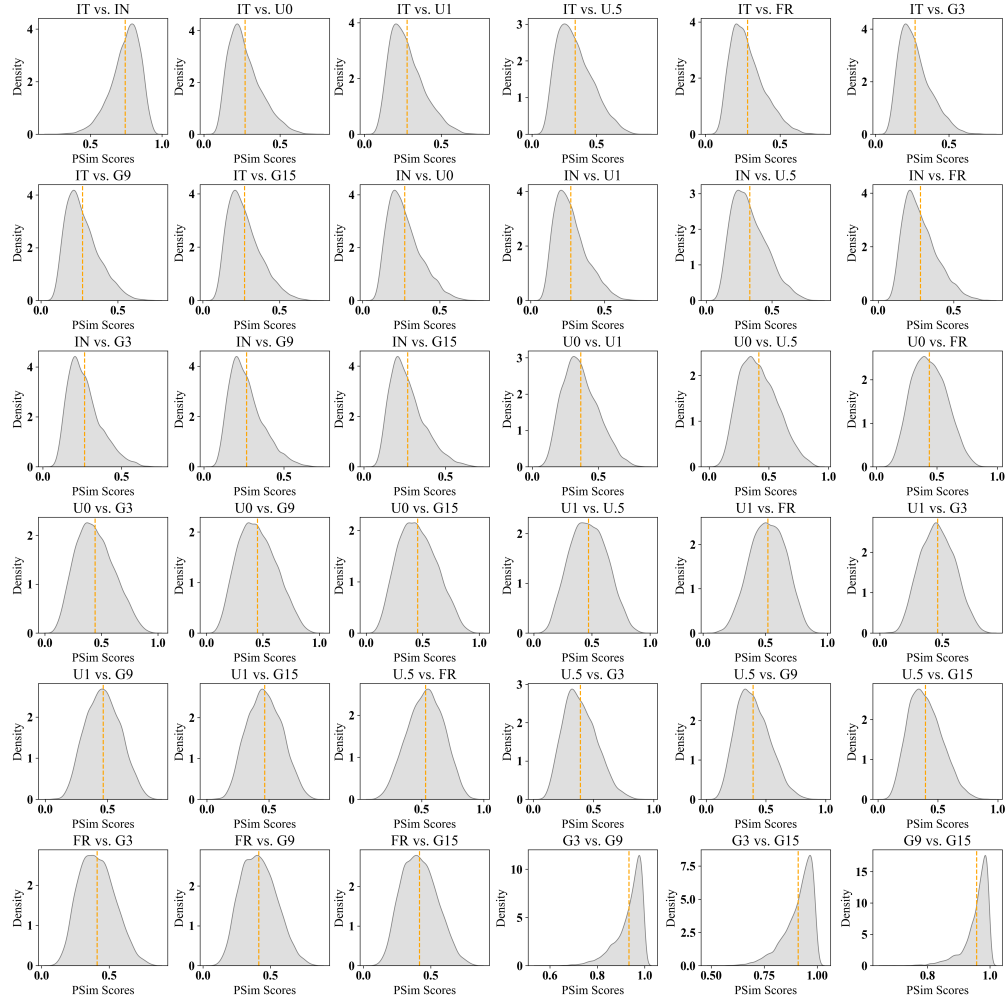


Figure S6: Distribution of pairwise *PSim* scores for all perturbations for Inception V3 model on Oxford-IIIT Pets Dataset

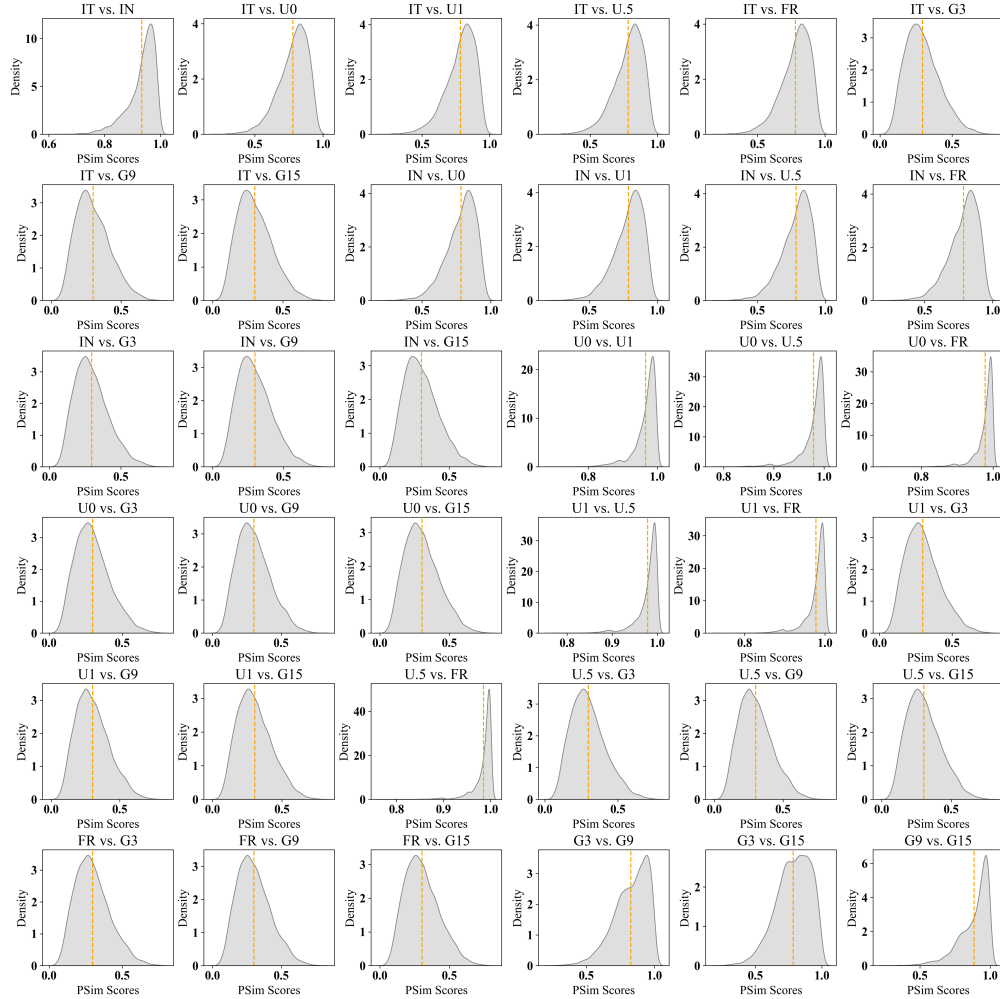


Figure S7: Distribution of pairwise $PSim$ scores for all perturbations for Resnet50 model on Oxford-IIIT Pets Dataset

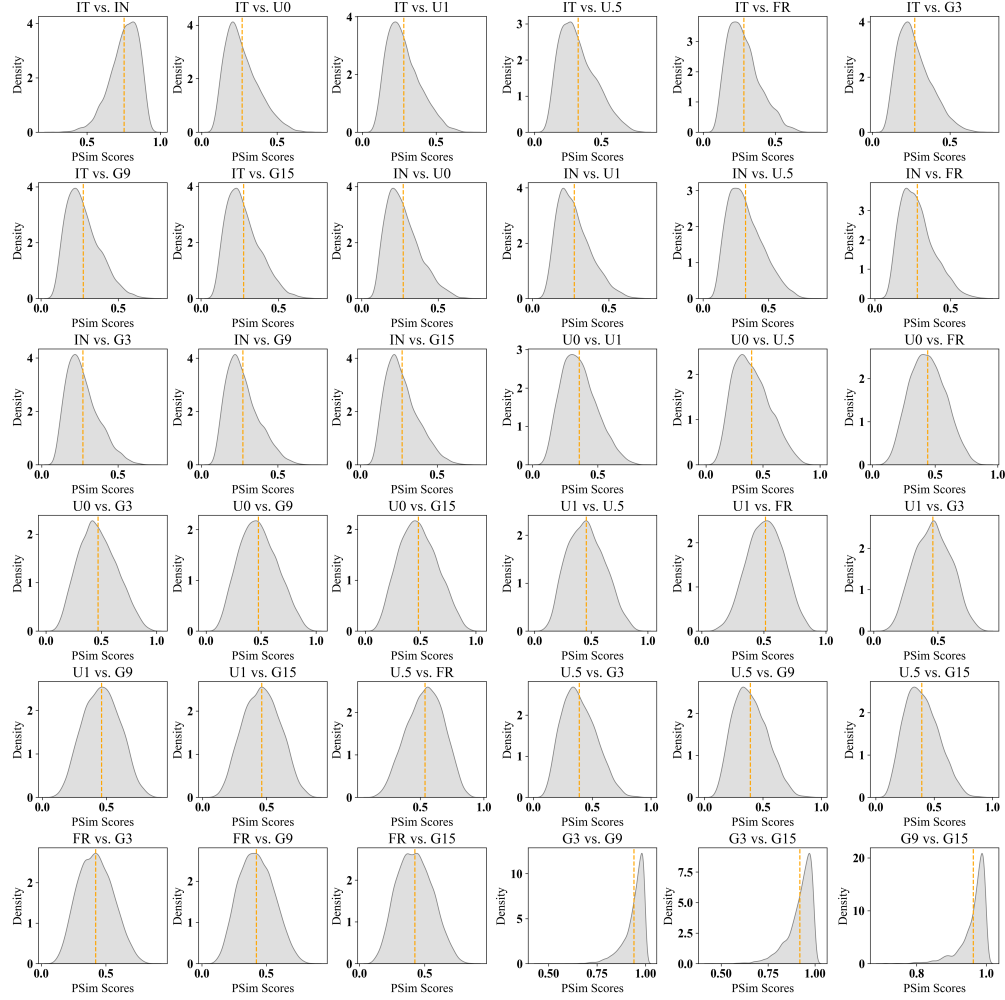


Figure S8: Distribution of pairwise *PSim* scores for all perturbations for Xception model on Oxford-IIIT Pets Dataset

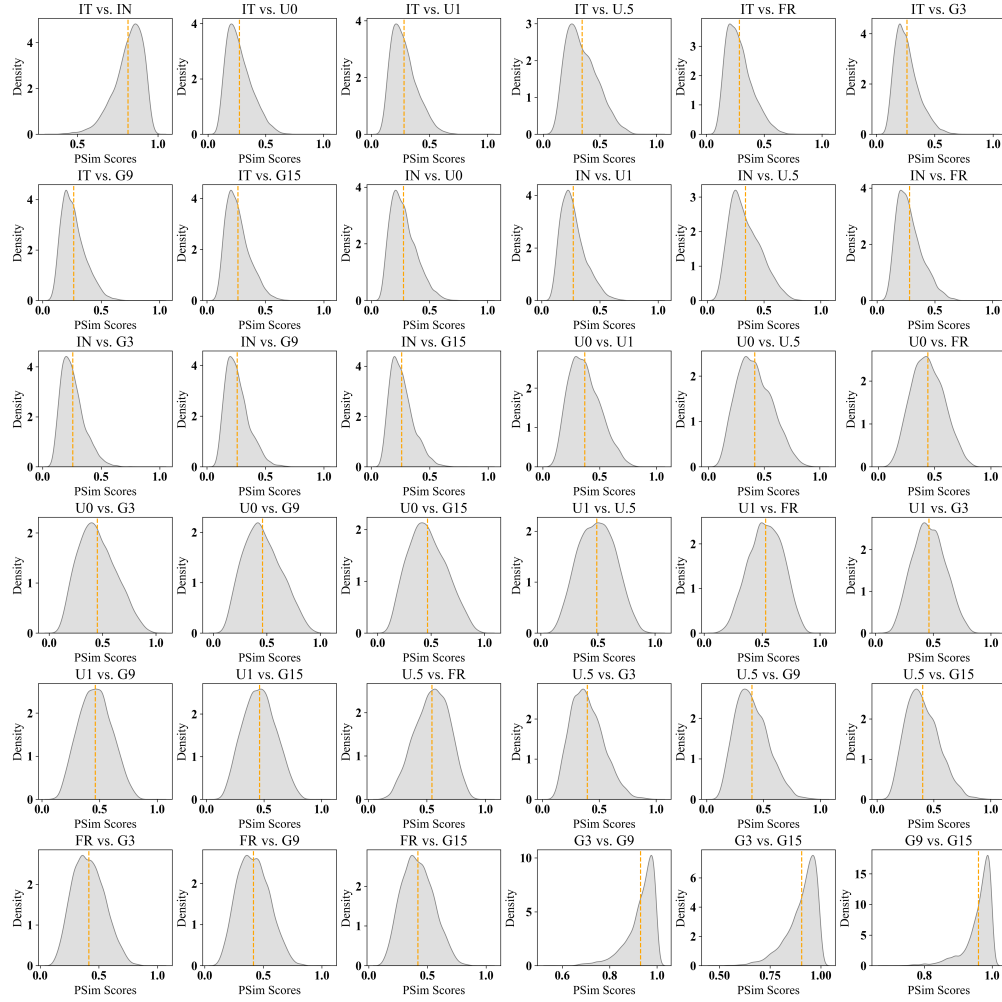


Figure S9: Distribution of pairwise $PSim$ scores for all perturbations for Inception V3 model on PASCAL VOC Dataset

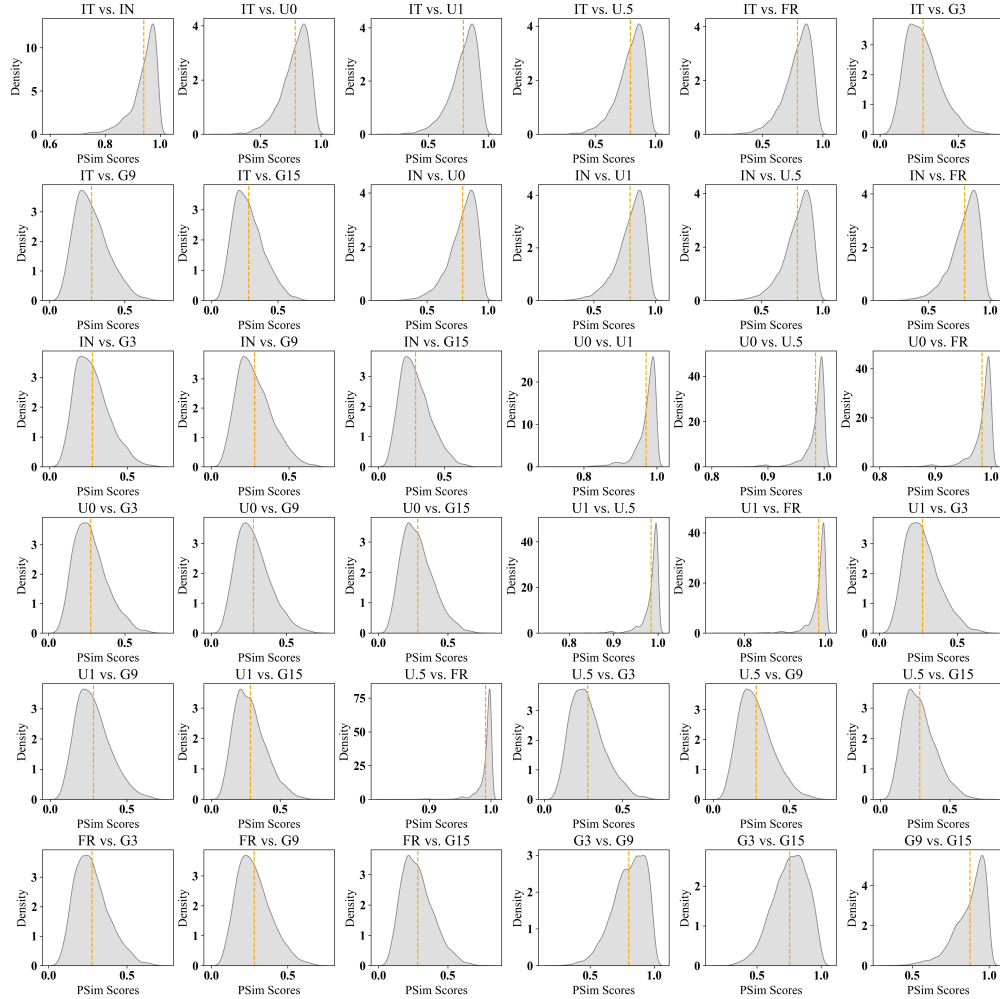


Figure S10: Distribution of pairwise $PSim$ scores for all perturbations for Resnet50 model on PASCAL VOC Dataset

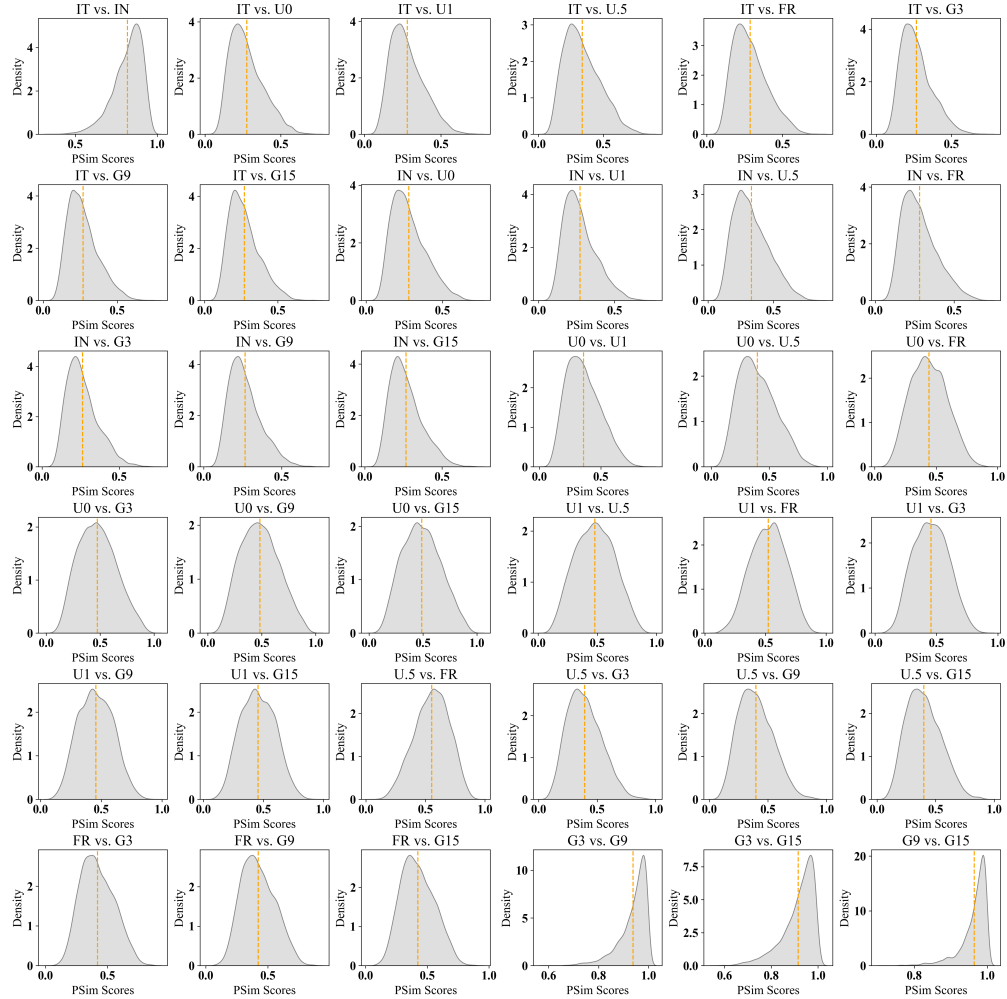


Figure S11: Distribution of pairwise $PSim$ scores for all perturbations for Xception model on PASCAL VOC Dataset

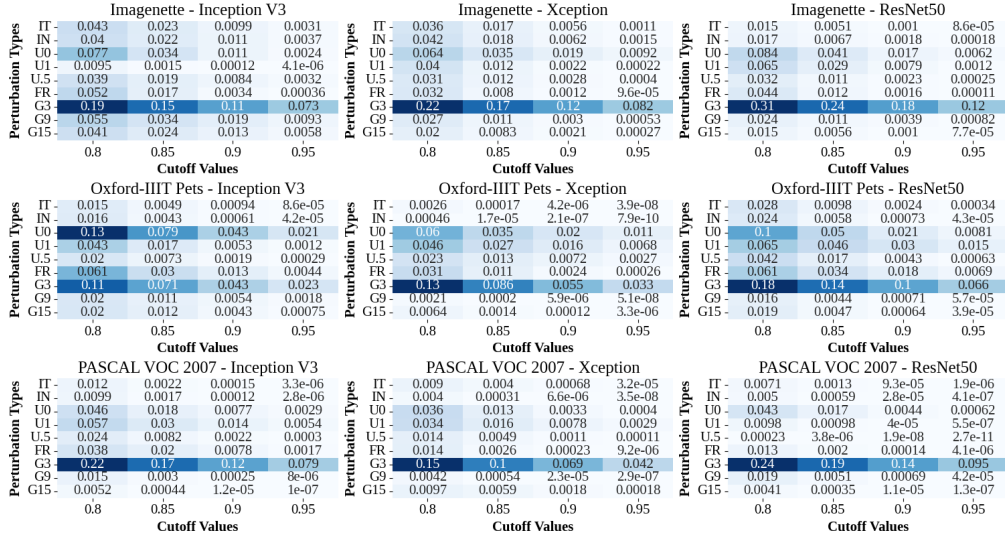


Figure S12: Distribution of *DROP* scores for dataset, model, perturbation types using segment-wise perturbation scheme

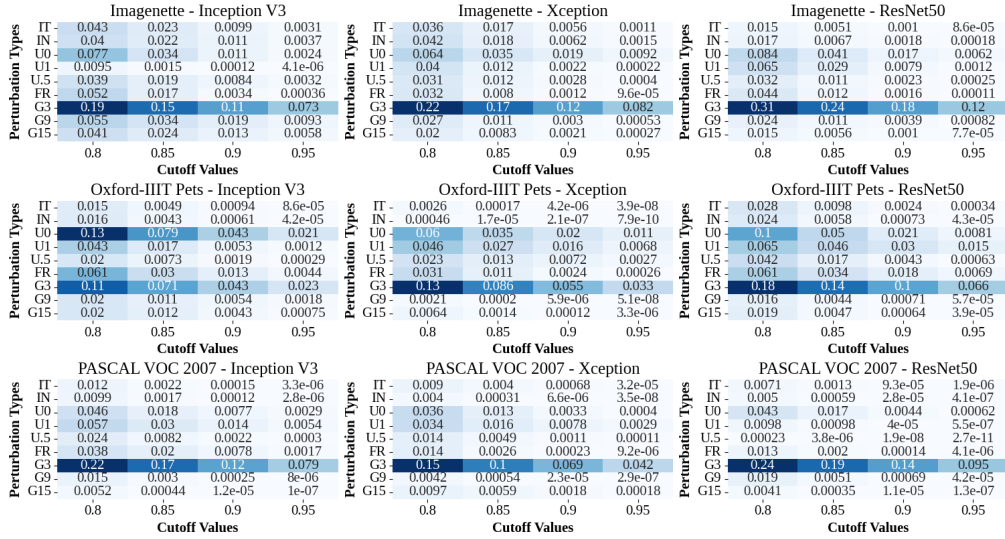


Figure S13: Distribution of *DROP* scores for dataset, model, perturbation types using pixel-wise perturbation scheme

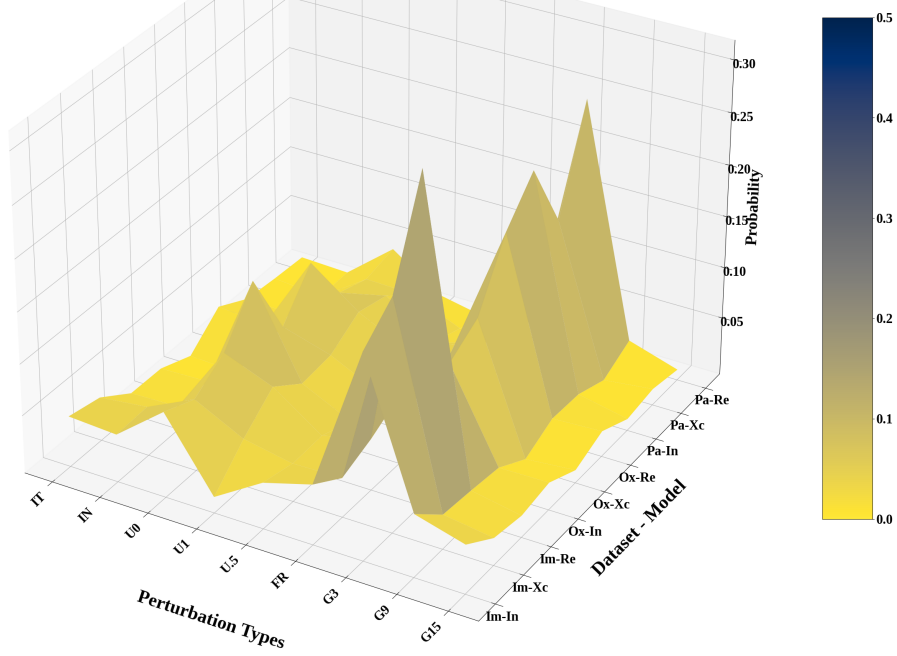


Figure S14: Probabilities of *DROPScore* to be above 0.80 for different perturbation pairs across all dataset-model combinations for segment-wise perturbation scheme.

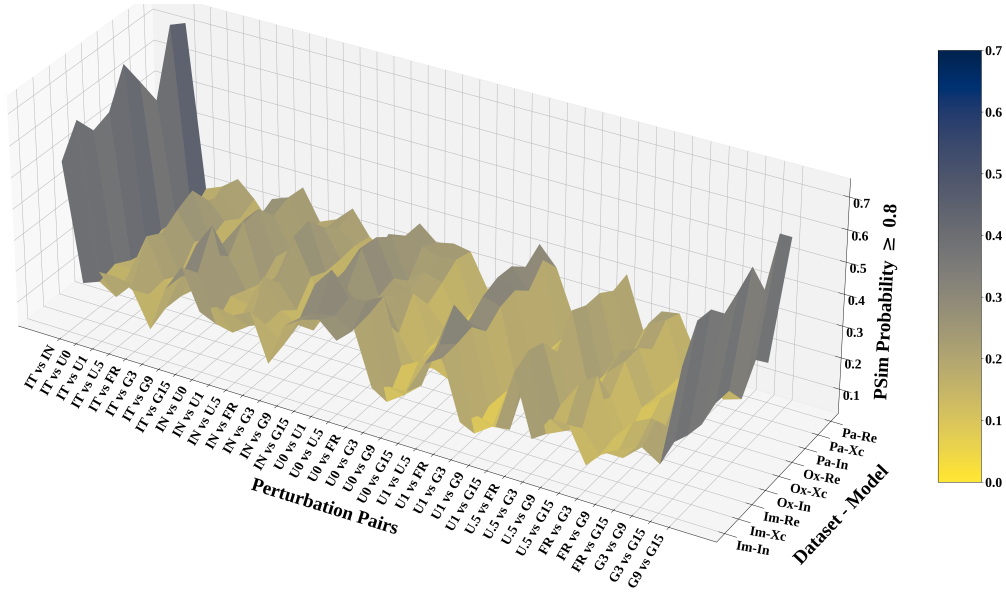


Figure S15: Probabilities of *PSim* score to be above 0.95 for different perturbation pairs across all dataset: model combinations