A1 Implementation details of FIDO-CA

For all the results of FIDO-CA, we followed the implementation details in the code released on Github `https://github.com/zzzace2000/FIDO-saliency` by the authors [29]. FIDO-CA was ran using the "preservation" objective in conjunction with the DeepFill-v1 [33] inpainter that we also harnessed in this paper. For the optimization, we used Adam optimizer with a learning rate of 0.05 and a regularization coefficient of 0.001. A coarse 56×56 mask was optimized using a ResNet-50 classifier for the ImageNet-S and Places365-S datasets respectively. The mask was finally upsampled to the full image size, i.e., 224×224, using bilinear interpolation.

A2 LIME-G is more robust than LIME on images of scenes, close-up and tiny objects

We have shown that LIME-G is more robust than LIME consistently on all 3 different similarity metrics (see Sec. 4.4 in the main text). Here, we aim to understand the image distributions where LIME-G was more robust than LIME and vice versa.

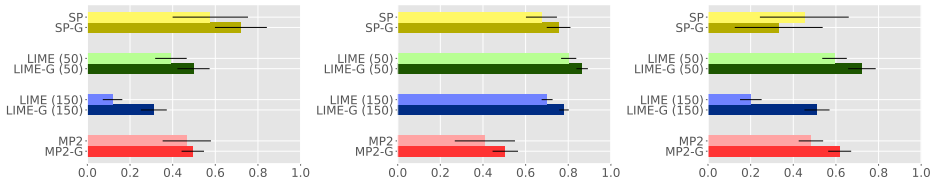
For each of the three metrics, we computed a set of top-100 score differences between LIME-G vs. LIME. Interestingly, we found the intersection of the three sets contains images of mostly scenes, close-up or tiny objects (see Fig. S1). In contrast, the common set of images where LIME is more robust than LIME-G contains mostly birds and medium-sized objects. These image distributions intuitively align with the domains where DeepFill-v1 is capable of inpainting and suggest that the performance of G-methods can be improved further with class-conditional inpainters.

Table S1: The results in this table are the number forms of the ImageNet sensitivity results in Fig. 6. G-methods are more robust to hyperparameters across different sensitivity metrics.

Method	Similarity Metrics		
	SSIM	Pearson correlation of HOGs	Spearman
SP	0.698±0.114	0.604±0.106	0.404±0.261
SP-G	**0.781±0.095**	**0.691±0.093**	0.317±0.206
LIME (50)	0.553±0.060	0.848±0.028	0.573±0.077
LIME-G (50)	**0.647±0.057**	**0.896±0.022**	**0.667±0.065**
LIME (150)	0.163±0.045	0.708±0.025	0.155±0.072
LIME-G (150)	**0.371±0.051**	**0.776±0.022**	**0.379±0.059**
MP2	0.476±0.155	0.453±0.096	0.522±0.088
MP-G	**0.479±0.064**	**0.569±0.051**	**0.698±0.054**

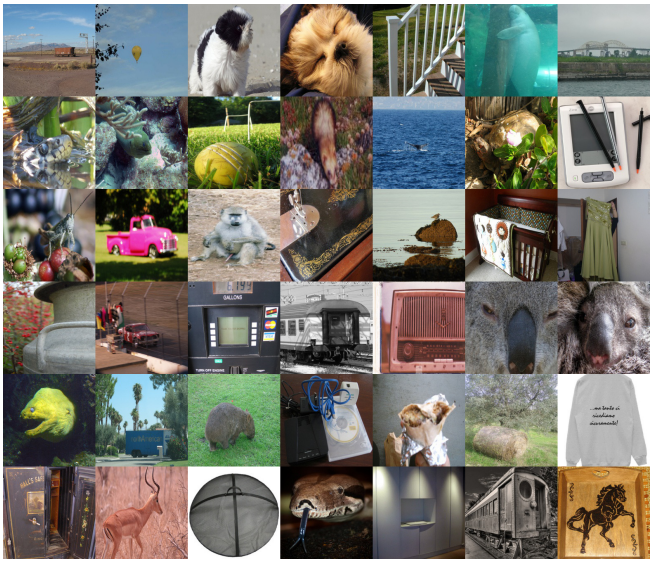
Table S2: The results in this table are the number forms of the Places365 sensitivity results in Fig. S2. The results follow the same trend as the ImageNet dataset.

Method	Similarity Metrics		
	SSIM	Pearson correlation of HOGs	Spearman
SP	0.577±0.177	0.674±0.073	0.452±0.288
SP-G	**0.720±0.122**	**0.755±0.056**	**0.332±0.208**
LIME (50)	0.392±0.074	0.802±0.036	0.594±0.078
LIME-G (50)	**0.498±0.076**	**0.865±0.027**	**0.722±0.058**
LIME (150)	0.118±0.046	0.701±0.026	0.201±0.071
LIME-G (150)	**0.312±0.061**	**0.780±0.022**	**0.511±0.051**
MP2	0.466±0.113	0.409±0.141	0.483±0.140
MP2-G	**0.494±0.053**	**0.505±0.060**	**0.618±0.057**



(a) SSIM (b) Pearson correlation of HOG features (c) Spearman rank correlation

Fig. S2: Bar plots comparing the robustness (higher is better) of G-methods and their counterparts when changing hyperparameters (described in Sec. 4.4) under three different similarity metrics: SSIM (a), Pearson correlation of HOG features (b), and Spearman rank correlation (c). Each bar shows the mean and standard deviation similarity score across 1000 pairs of heatmaps, each produced for one random **Places365** image. G-methods are consistently more robust than their counterparts across all metrics. The exact numbers are reported in Table S2.



Images where LIME-G outperformed LIME across all three sensitivity metrics

Images where LIME-G underperformed LIME across all three sensitivity metrics

Fig. S1: Common images across all three metrics where LIME-G is consistently more robust than LIME (top) and vice versa (bottom). Interestingly, we found the intersection of the three sets contains images of mostly scenes, close-up or tiny objects (top). In contrast, the common set of images where LIME is more robust than LIME-G contains mostly birds and medium-sized objects (bottom).

(a) Real (b) Mask (c) Preserve (d) Delete (e) Real (f) Mask (g) Preserve (h) Delete

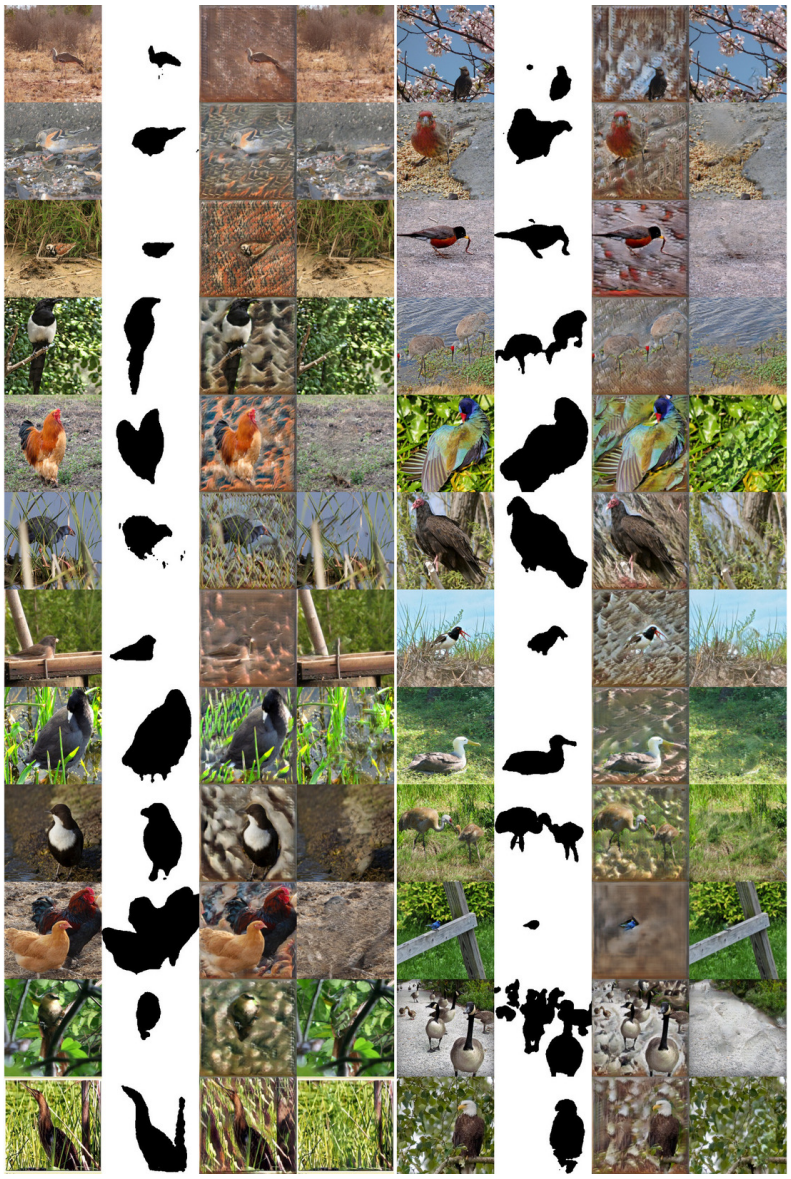
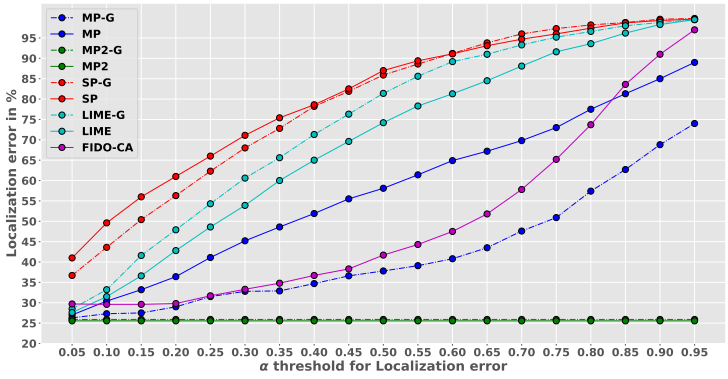


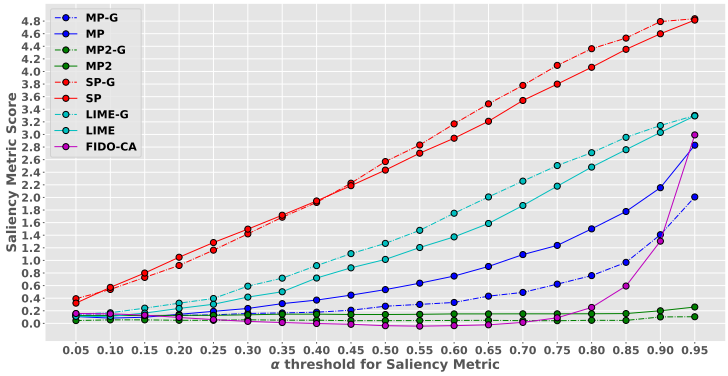
Fig. S3: Inpainting using the preservation objective generates unrealistic samples (Sec.4.1). We randomly chose 50 validation-set images (a) from 52 ImageNet bird classes and compute their segmentation masks via a pre-trained DeepLab model [39] (b). We found that using the DeepFill-v1 inpainter to inpaint the foreground region (i.e. our "deletion" task) yields realistic samples where the object is removed (d). In contrast, using the inpainter to fill in the background region (i.e. "preservation" task) yields unrealistic images whose backgrounds contain features (e.g. bird feathers or beaks) unnaturally pasted from the object (c).

Fig. S4: Top-10 cases where the LIME-G outperformed (left) and underperformed (right) LIME on the object localization task (IoU scores). From left to right, on each row: we show a real image with its ground-truth bounding box, LIME heatmap & its derived bounding box, LIME-G heatmap & its derived bounding box. See `https://drive.google.com/drive/u/2/folders/10JeP9dpuoa0M16xe2FloBEWajQ7PNKSX` for more examples of the LIME and LIME-G IoU results.

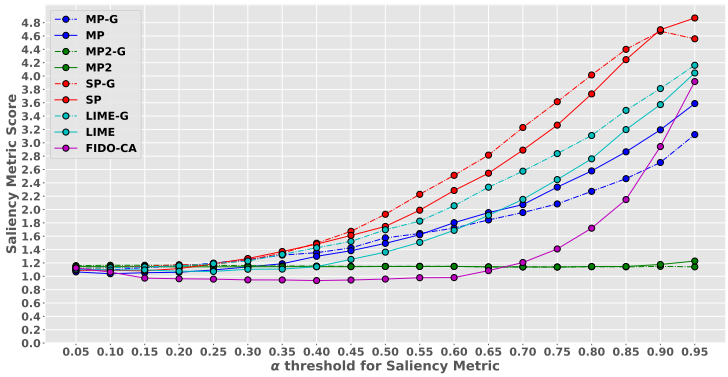
Fig. S5: Top-10 cases where the SP-G outperformed (left) and underperformed (right) SP on the object localization task (IoU scores). From left to right, on each row: we show a real image with its ground-truth bounding box, SP heatmap & its derived bounding box, SP-G heatmap & its derived bounding box. In the cases where SP-G has a lower IoU score than SP (right panel), we observed the heatmap localizes some unique features of the object as compared to the images in the top cases where the heatmap covers the entire image. See `https://drive.google.com/drive/u/2/folders/1XJ6M0AMHxZrXxLLw6m3Bx7sjvsyqN6JC` for more examples of the SP and SP-G IoU results.



(a) α vs Localization error for ImageNet



(b) α vs Saliency Metric for ImageNet



(c) α vs Saliency Metric for Places365

Fig. S6: Localization error (a) and saliency metric (b, c) performance of different attribution methods on a held-out set of 1000 images for different α threshold values. For each method, we search for the optimal α value on this held-out set and use the subsequent threshold for computing the scores on the 2000 images in the object localization and saliency metric experiments in Sec. 4.3.

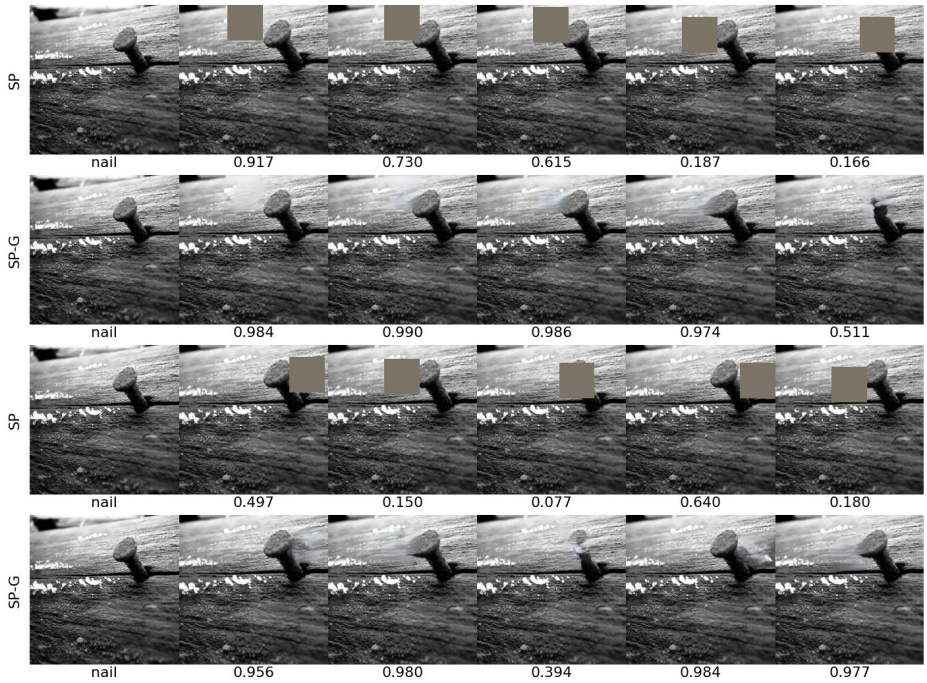


Fig. S7: Random intermediate perturbation samples by SP and SP-G on the same image from the nail class in ImageNet. SP-G drops the target-class probability only when the patch cover a major area of the nail (e.g. the center 0.394-probability sample in the bottom panel). This figure is a zoom-in version of the samples in Fig. 7.

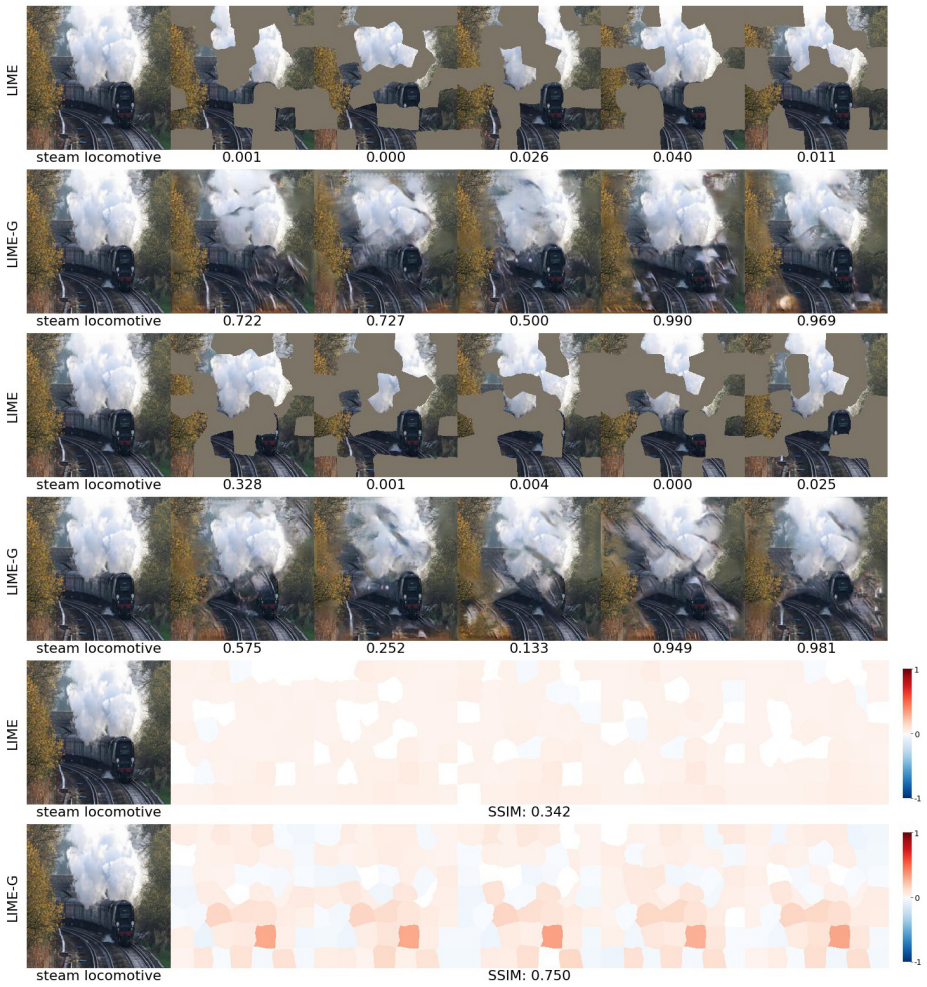


Fig. S8: Qualitative evidence supporting the LIME-G vs. LIME sensitivity experiment in Sec. 4.4. For both LIME and LIME-G, per image, we compute an average SSIM score across all 10 pairs of 5 heatmaps. We then take the difference between LIME-G and LIME and sort them in the descending order. This steam locomotive image is a random image from the top-100 ImageNet-S cases where LIME-G outperformed LIME. **Top four rows:** Here, we compare pairs of LIME vs. LIME-G perturbation samples that were created from the same random superpixel masks. LIME-G samples cause large probability drops only when some discriminative feature is removed from the image and thus results in more localized heatmaps. **Bottom two rows:** 5 heatmaps by LIME and LIME-G, each from a random seed. While LIME-G heatmaps are more consistent, LIME heatmaps is noisy and varies. See Fig. S9 and Figs. S10-S11 for similar observations in ImageNet-S and Places365-S dataset respectively. See `https://drive.google.com/drive/u/2/folders/1sKWig4Xk5Pm50kdONdAS9SkiTBhJRAkw` for more examples.

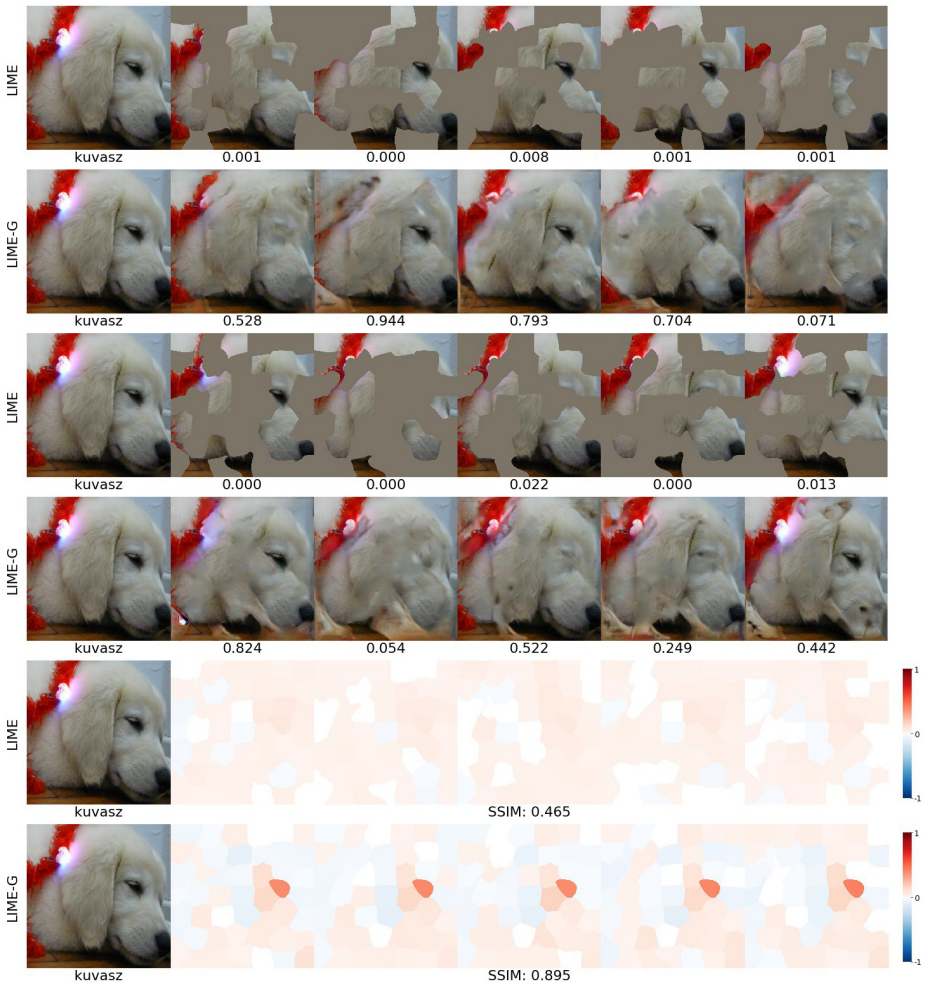


Fig. S9: Here, we show the same figure as Fig. S8 (see its caption) but for another random image among the top-100 ImageNet-S cases where LIME-G outperformed LIME on the SSIM similarity metric. See `https://drive.google.com/drive/u/2/folders/1sKWig4Xk5Pm50kdONdAS9SkiTBhJRAkw` for more examples.

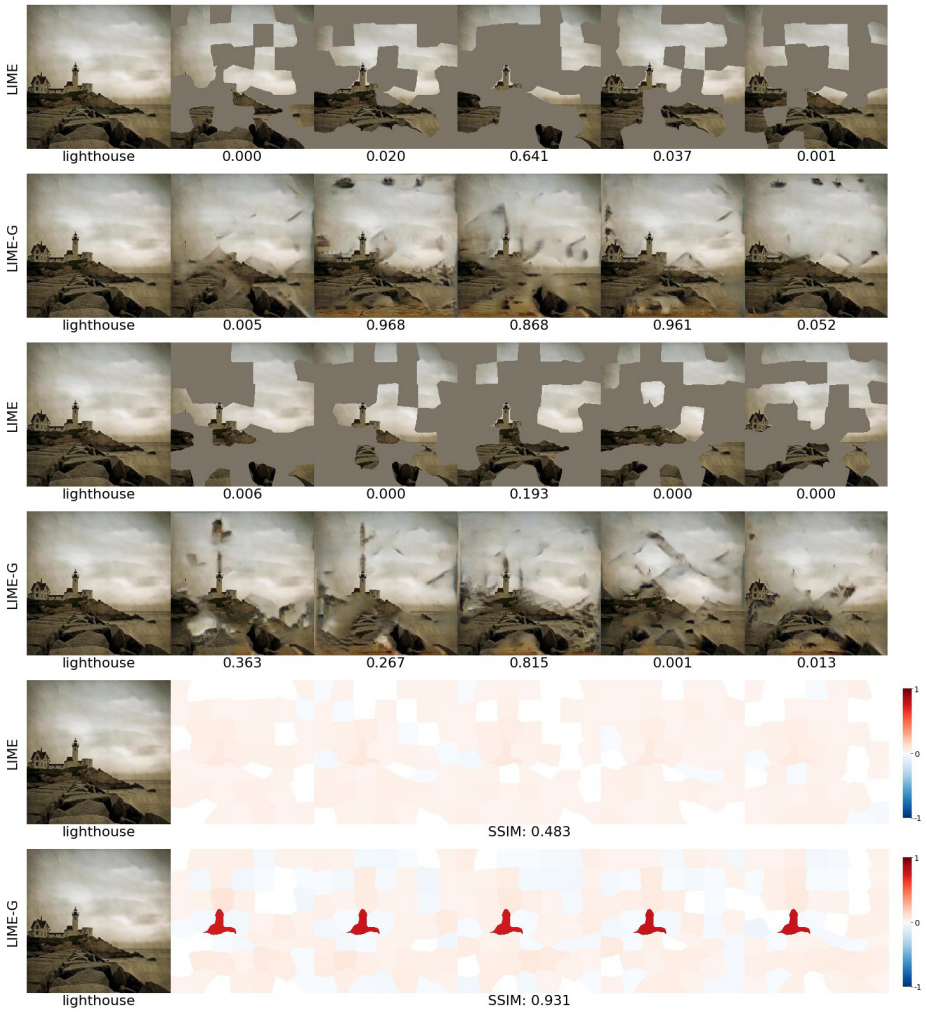


Fig. S10: Here, we show the same figure as Fig. S8 (see its caption) but for a random image among the top-100 Places365-S cases where LIME-G outperformed LIME on the SSIM similarity metric. See `https://drive.google.com/drive/u/2/folders/1aXyDFBq0HlcI0kQJpJyspNf2rtwLj35Z` for more examples.

Fig. S11: Here, we show the same figure as Fig. S8 (see its caption) but for a random image among the top-100 Places365-S cases where LIME-G outperformed LIME on the SSIM similarity metric. See `https://drive.google.com/drive/u/2/folders/1aXyDFBq0HlcI0kQJpJyspNf2rtwLj35Z` for more examples.

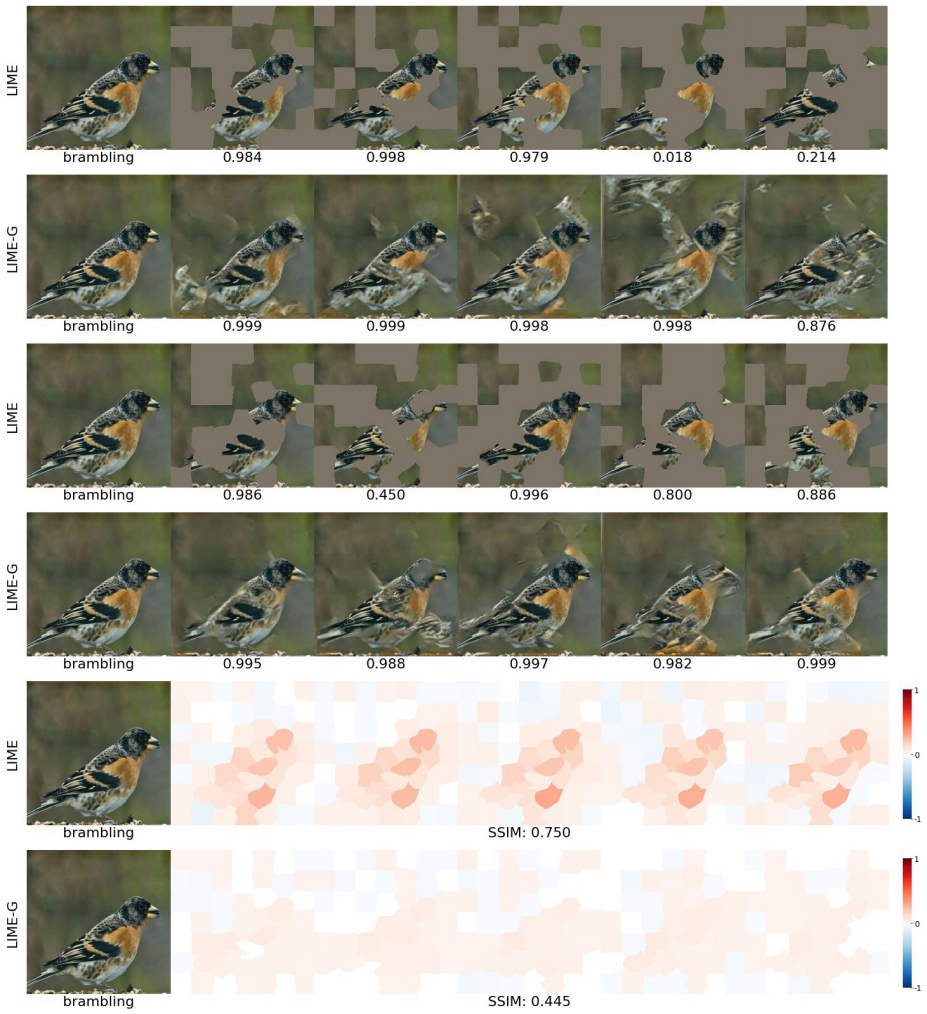


Fig. S12: Here, we show the same figure as Fig. S8 (see its caption) but for a random image among the top-100 ImageNet-S cases where LIME-G underperformed LIME on the SSIM similarity metric. LIME-G samples remain at high target-class probabilities and therefore produced heatmaps that are more sensitive than those of LIME. Similar observations can be found in Fig. S13 and Figs. S14-S15. See `https://drive.google.com/drive/u/2/folders/1sKWig4Xk5Pm50kdONdAS9SkiTBhJRAkw` for more examples.

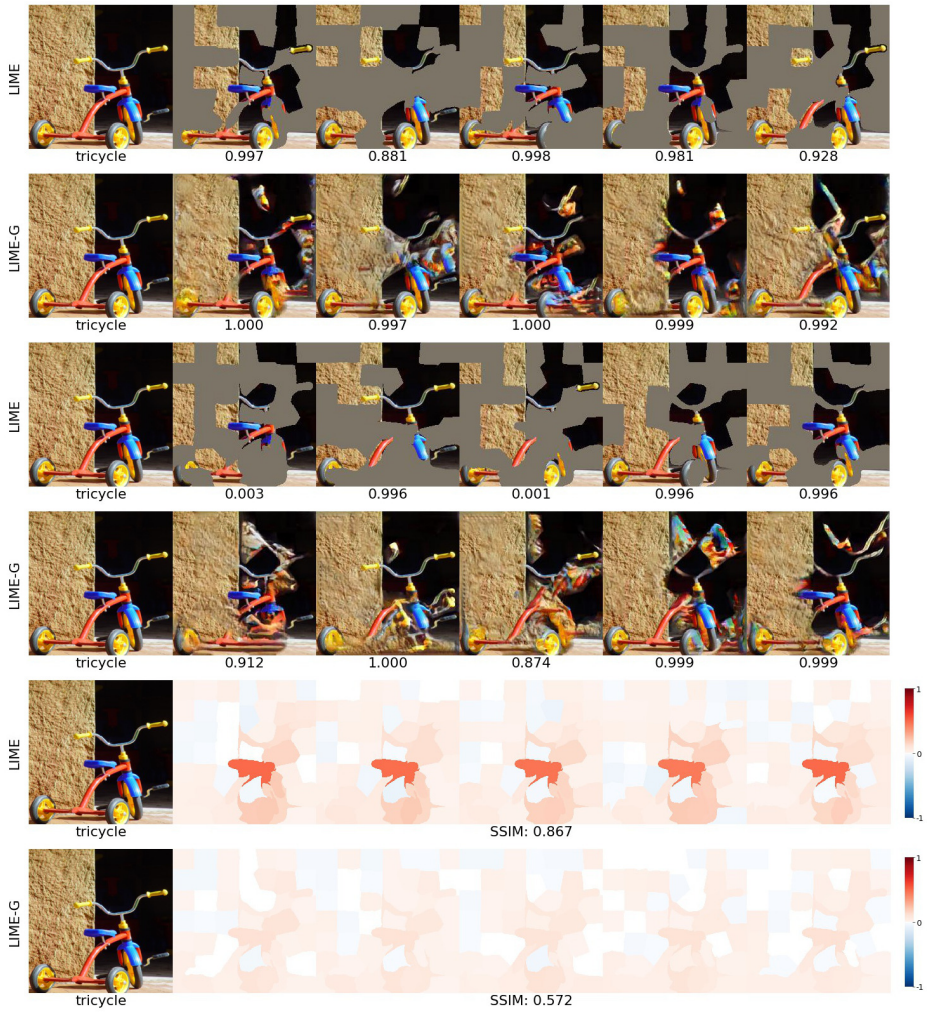


Fig. S13: Here, we show the same figure as Fig. S8 (see its caption) but for a random image among the top-100 ImageNet-S cases where LIME-G underperformed LIME on the SSIM similarity metric. See `https://drive.google.com/drive/u/2/folders/1sKWig4Xk5Pm50kdONdAS9SkiTBhJRAkw` for more examples.

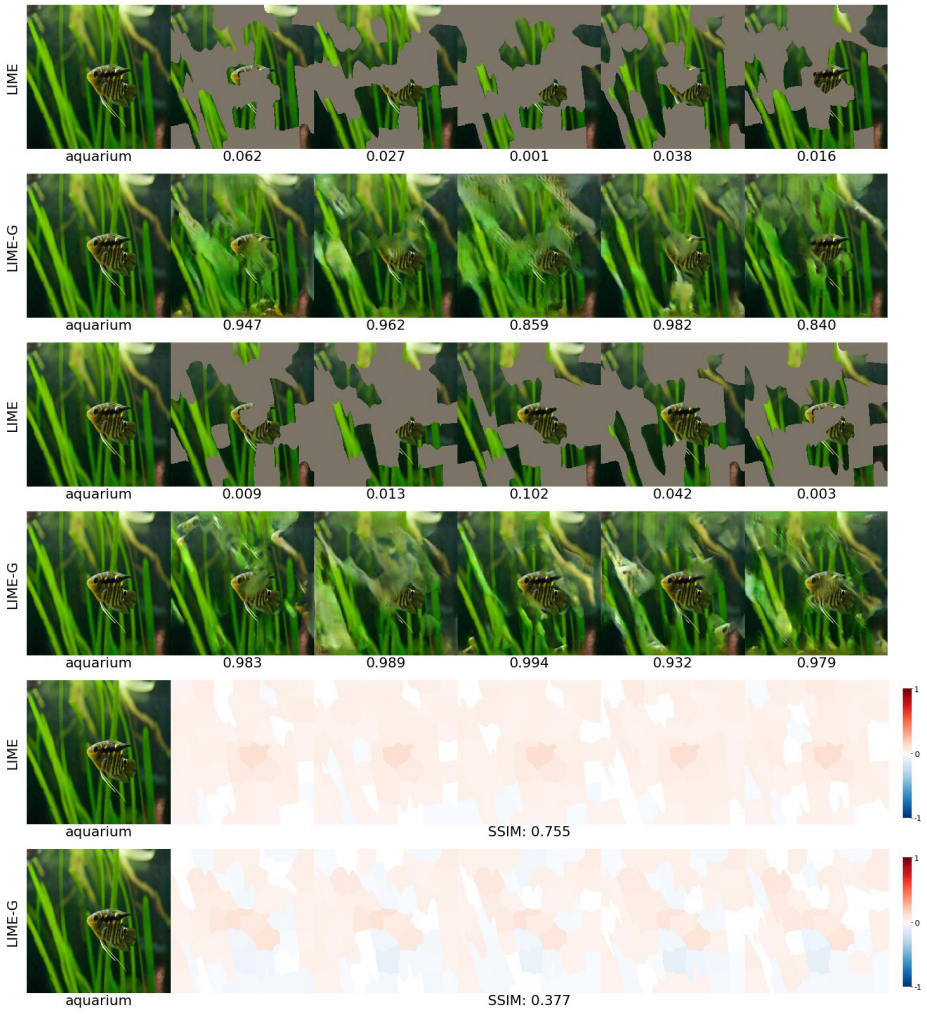


Fig. S14: Here, we show the same figure as Fig. S8 (see its caption) but for a random image among the top-100 Places365-S cases where LIME-G underperformed LIME on the SSIM similarity metric. See `https://drive.google.com/drive/u/2/folders/1aXyDFBq0HlcI0kQJpJyspNf2rtwLj35Z` for more examples.

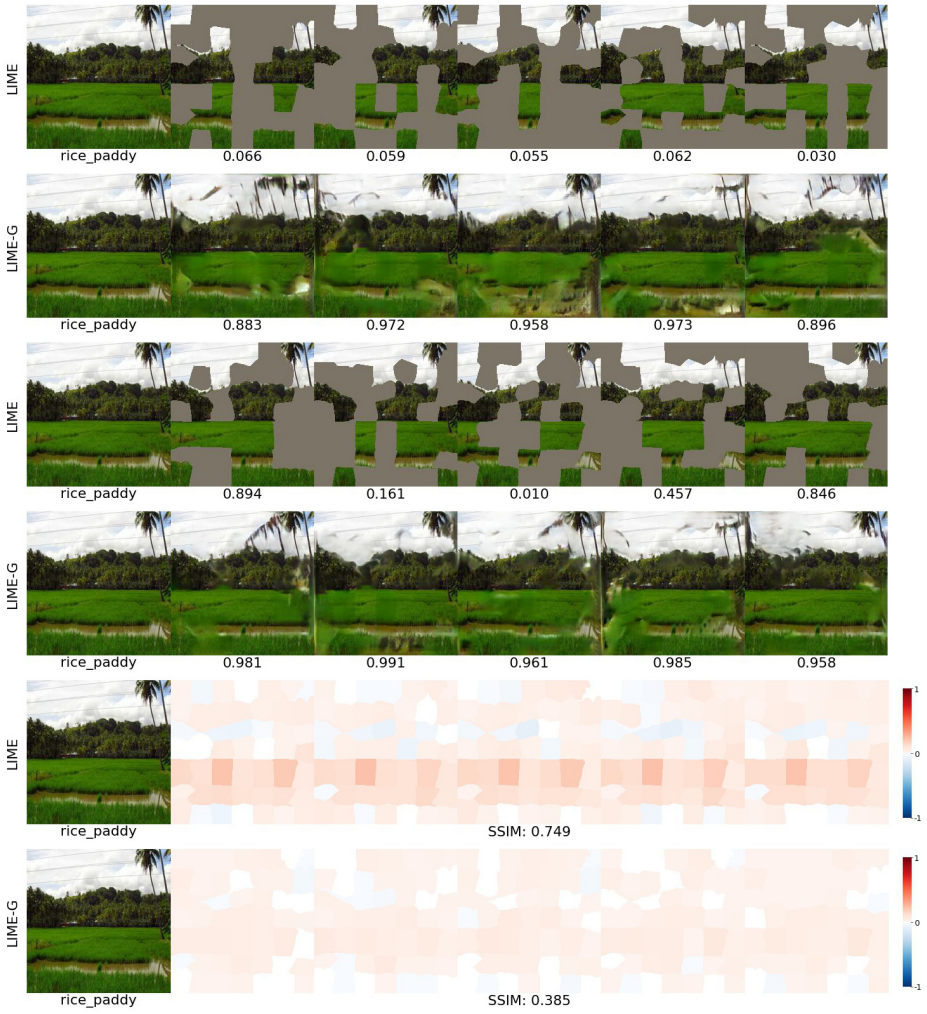
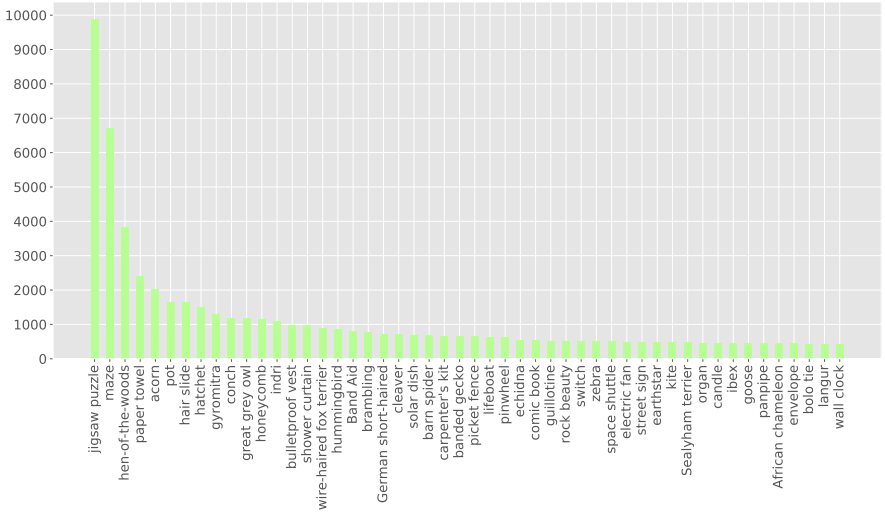
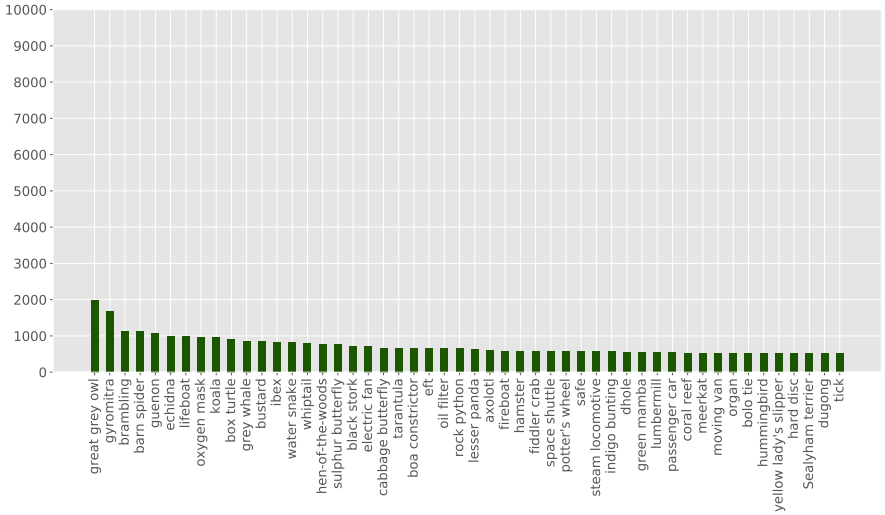


Fig. S15: Here, we show the same figure as Fig. S8 (see its caption) but for a random image among the top-100 Places365-S cases where LIME-G underperformed LIME on the SSIM similarity metric. See `https://drive.google.com/drive/u/2/folders/1aXyDFBq0HlcI0kQJpJyspNf2rtwLj35Z` for more examples.

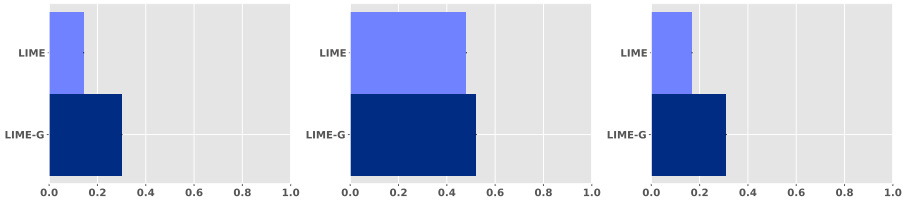


(a) LIME histogram distribution is skewed



(b) LIME-G histogram distribution is almost uniform

Fig. S16: We ran LIME and LIME-G on 200 images, each run has 500 intermediate perturbation samples. Here, for LIME (a) and LIME-G samples (b), we show a histogram of the top-1 predicted class labels for all 200 runs $\times 500$ samples = 100,000 images. The set of 200 images comprises of cases where LIME-G outperformed (100 images) and underperformed (100 images) LIME on the SSIM sensitivity metric (Sec. 4.4). LIME perturbed samples are highly biased towards few jigsaw puzzle, maze classes (top panel), which is somewhat intuitive given the gray-masked images (see Figs. S8–S13). In contrast, the histogram of LIME-G samples are almost uniform. **x-axis:** For visualization purposes, we sorted the top-1 labels and showed only first 50 labels.



(a) SSIM (b) Pearson correlation of HOG features (c) Spearman rank correlation

Fig. S17: Bar plots comparing the LIME vs. LIME-G robustness (higher is better) across two different numbers of superpixels $S \in \{50, 150\}$ under three different similarity metrics: SSIM (a), Pearson correlation of HOG features (b), and Spearman rank correlation (c). For each image in 1000 random ImageNet-S images, we produced a pair of heatmaps by running LIME (light-blue) or LIME-G (dark-blue) with two different numbers of superpixels $S \in \{50, 150\}$. Each bar shows the mean similarity across all 1000 heatmap pairs. LIME-G is consistently more robust than LIME, specifically by $\sim$200% under the SSIM (a) and Spearman rank correlation (c).