We thank all the reviewers for their constructive reviews, and apologize in advance for not being able to answer all questions or provide detailed experimental results due to the lack of space.

Digit-flipping experiment (suggested by R1): Broadly, the task here is to turn images of the MNIST digit "8" into those of the digit "3" by removing pixels which provide positive evidence of "8" and negative evidence for "3". We perform experiments with a setting similar to the DeepLIFT paper, except that we use a VGG-like architecture. The change in log-odds are as follows for different methods. Random - 1.41 ± 8.82 , Gradient - 11.93 ± 17.87 , Integrated gradient - 11.95 ± 17.19, FullGrad (with post-processing) - 8.48 ± 20.81, FullGrad (raw) - $\mathbf{12.93 \pm 18.21}$. Higher numbers are better. Here, "raw" FullGrad refers to naive aggregation without using any post-processing, except up-sampling for max-pooled maps.

We also compare raw FullGrad against the FullGrad with post-processing, on the **pixel perturbation task on MNIST**, and we find obtain the following fractional output-change values upon removing 70% of least salient pixels - FullGrad (with post-processing) - $\mathbf{0.08 \pm 0.11}$, FullGrad (raw) - 0.35 ± 0.23. Smaller is better here. Similar trend holds for other fractions of pixel removal and for Imagenet experiments as well. Note that we take absolute value of all heatmaps as this test requires unsigned heatmaps. Also note that absolute value of raw FullGrad $\neq$ FullGrad with post-processing.

From these experiments, we make the following conclusions: (1) FullGrad explanations are indeed class sensitive (as measured by digit-flipping), (2) Different interpretability tests may require different different forms of post-processing. This surprising fact is consistent in spirit with our theory (Proposition 1), which states that **a single saliency map cannot satisfy all intuitive properties we wish to impose**. The digit-flipping experiment emphasizes signed heatmaps, and hence raw FullGrad does better, while pixel perturbation / ROAR experiments place emphasis on finding important pixels regardless of their direction of influence, and hence layer-wise post-processing works better here. A principled method to find suitable post-processing in a task-dependent manner is a non-trivial problem for future work. We conjecture that this is perhaps analogous to the problem of finding appropriate reference inputs for Integrated gradients / DeepLIFT (see Q2). Finally, we thank R1 for suggesting this experiment that helped us obtain more insight about these aspects, and we will add related discussion in an update of the manuscript.

Response to R1:

We have responded to concerns about class-sensitivity and the importance of post-processing in the section above.

Q1) *"It is known that when bias terms are included in the attributions, the approach of Layerwise Relevance Propagation reduces to gradient*input for ReLU networks and satisfies completeness."* While it is true that LRP reduces to gradient * input in some cases, **it is incorrect that gradient * input satisfies completeness for ReLU nets with bias**. Proposition 2 shows that completeness for gradient * input holds only when the ReLU net has no biases. Completeness with a baseline also cannot be satisfied as $f(\mathbf{x}) - f(\mathbf{x}_0) = \nabla_{\mathbf{x}} f(\mathbf{x})^T \mathbf{x} + f^b(\mathbf{x}) - \nabla_{\mathbf{x}_0} f(\mathbf{x}_0)^T \mathbf{x}_0 - f^b(\mathbf{x}_0)$, does not reduce to a gradient * input term, even when $\mathbf{x}_0 = 0$, as the bias-gradient terms are not equal ($f^b(\mathbf{x}_0) \neq f^b(\mathbf{x})$) in general $\forall \mathbf{x}$ except in certain pathological cases (such as ReLU nets with no bias, or linear models).

Q2) *"Benchmark against Integrated Gradients scores averaged over multiple choices of the reference."* We test this with pixel perturbation experiments on Imagenet, where we choose references drawn from $N(0, 1), N(0, 0.1)$ and average the resulting maps. Specifically, for 1% pixel removal, the fractional change in output is as follows. Zero-reference $\rightarrow \mathbf{0.035 \pm 0.056}$, $N(0, 0.1) \rightarrow 0.042 \pm 0.061$, $N(0, 1) \rightarrow 0.065 \pm 0.078$. Smaller is better. Similar trend holds across other removal fractions. This suggests that we may need better heuristics for reference selection for high dimensional problems such as Imagenet classification.

Response to R2:

Q3) *"Definition 2 does not seem to be a sufficient characterization for completeness."* Thanks for bringing this to our notice! This can be easily fixed by forcing ϕ to depend on $S(\mathbf{x})$, i.e.; by requiring that $\phi(S(\mathbf{x}), \mathbf{x})$ is **not** a constant function of $S(\mathbf{x})$. This slight modification does not change the proofs or the implications.

Response to R3:

Q4) *"I would like to see 1) a better explanation of why the least k features cannot cause high-frequency edge artifacts, and 2) some experiments...to back this up empirically."* (1) Removing least-k features still causes artefacts, but the test measures which saliency methods identify pixels whose change doesn't affect the output, even if they have large artefacts. For the largest-k feature removal test, **random heatmaps would create maximal artefacts without identifying important pixels**, and can cause large output changes and falsely be considered a strong baseline. (2) For largest-k test, with 1% pixel removal, we obtain following output-change values. Random - 0.14 ± 0.11, Integrated gradients - 0.15 ± 0.11. Here larger is better, and it (falsely) seems that both methods are almost identical. For least-k, random remains the same, but Integrated gradient gets 0.036 ± 0.056. Here smaller is better, and the gap between the methods is more evident. We will add related discussion in an update of the manuscript.