We thank all reviewers for their constructive and helpful reviews. We will incorporate their suggestions in the final version, and in particular, expand our discussions of related work as suggested by R2. We will also add clarifications and additional experiment results discussed in this response to our final draft.

Originality and significance of the infidelity measure (R1, R3): We believe that the main novelty of infidelity (Definition 2.1) is the introduction of the random variable that represents the perturbation, which can be chosen by the user. Previous evaluation methods consider the perturbation to be fixed (set to some baseline value), and also consider all possible combinations of features. Definition 2.1 moreover allows us to focus on only problem specific perturbations of interest that incorporate prior knowledge into the problem, and may also be more computational feasible. Surprisingly, this new degree of freedom in setting the perturbation enables us to show that many existing explanations optimize the infidelity measure with respect to some perturbations (which also shows the importance of the additional flexibility). Additionally, we are able to introduce new explanations by simply defining a new perturbation. For instance, in the human evaluation experiment we only care about whether the model looks at the image or the caption, and we define our perturbation correspondingly, which is not possible for previous evaluation metrics for explanations.

Comprehensive Experiments (R1): We set up a sanity check experiment on MNIST when the perturbation follows that in SHAP (Defined in Prop.2.5), and we verify that SHAP has the lowest infidelity for this perturbation, which verifies Prop.2.5. (The infidelity for each explanation is GBP-SG: 7.0, SHAP: 2.0, Square: 3.7, Grad-SG: 6.6, GBP: 10.5, IG-SG: 5.0, IG: 6.6, Grad: 12.0). We will include a more complete version of such sanity check experiments in our final version. We also verify that the infidelity for IG and SHAP is 0 when the perturbation is a constant (Prop.2.1). The 0.2 radius for SG is the parameter that optimized infidelity score by Square, but the optimal radius for SG in MNIST for NB perturbation is around 1.0 (as in Fig.6). We redid experiments where we choose the SG radius for each setting by validation, and observe little changes in the relative results (OPT > SG > vanilla explanations with a great margin).

Smoothing Operation and General Kernel (R1): A smoothed explanation can be defined as $\Phi_k(\mathbf{f}, \mathbf{x}) := \int_{\mathbf{z}} \Phi(\mathbf{f}, \mathbf{z})\, k(\mathbf{x}, \mathbf{z}) d\mathbf{z}$ with some probability kernel $k(\cdot, \cdot)$ (which implies $k \geq 0$ by definition), which is equivalent to Smooth-Grad when $k(\mathbf{x}, \mathbf{z})$ is a Gaussian kernel. When $k(\mathbf{x}, \mathbf{z})$ is not a probability kernel, we may write the smoothed explanation as $\Phi_k(\mathbf{f}, \mathbf{x}) := [\int_{\mathbf{z}} k(\mathbf{x}, \mathbf{z})]^{-1} \int_{\mathbf{z}} \Phi(\mathbf{f}, \mathbf{z})\, k(\mathbf{x}, \mathbf{z}) d\mathbf{z}$ to make it invariant of linear scaling of the kernel. Mathematically, the optimal solution of Proposition 2.1 is a smoothed explanation where $k(\mathbf{x}, \mathbf{z})$ is replaced by $\mathbf{I}\mathbf{I}^T$, which is reminiscent of Smooth-Grad by replacing the Gaussian kernel by $\mathbf{I}\mathbf{I}^T$. In our experiments, we observe sensitivity and infidelity results are close when using Gaussian and uniform kernel, but the uniform kernel (or Truncated Normal) may be beneficial when we have a hard restriction for the input region (as Gaussian Kernel is unbounded).

Invertible Case (R1): When I is deterministic, the integral of $\mathbf{I}\mathbf{I}^T$ is rank-one and cannot be inverted, but being optimal with respect to the infidelity can be shown to be equivalent to satisfying the Completeness Axiom (see Proposition 2.2 where we address this case). The optimal solution then is no longer unique, since many feature attributions satisfy the Completeness Axiom. One way to achieve an unique solution is to add a non-deterministic noise to the baseline (as in noisy Baseline in l167-l172, which gives an unique explanation that satisfies a "robust Completeness" Axiom.) To enhance computational stability, we can replace inverse by pseudo-inverse, or add a small diagonal matrix to overcome the non-invertible case, which works well in experiments.

Advantage of Definition 3.1 (R1): While the benefit of our definition 3.1 compared to local Lipschitz continuous is not the main contribution of the paper, we point out that in certain cases, local Lipschitz continuity may be unbounded in a deep network (such as using ReLU activation function for gradient explanations, which is a common setting), but definition 3.1 is always finite given that explanation score is bounded, and thus is more robust to estimate.

Solve challenges of Localizing (feature-based) explanations (R2): Three major critiques of feature-based saliency maps are: (a) there is no fair way to evaluate the explanation, (b) feature-based explanations may not be faithful to the model (as shown by saliency sanity checks), and (c) explanations may not be robust to small perturbations. To address the first two challenges, our work proposes an evaluation metric (infidelity) that is more general than previous evaluation methods, allowing the user to define the perturbation of interest based on the context. In the Human Evaluation experiment, our infidelity measure is able to evaluate the quality of many explanations for this specific task and lead to an optimal explanation that improves human evaluation accuracy and passes sanity check. We also show that smoothing explanation and adversarial training (in Supplement A.3) allow us to obtain more robust explanations.

Differences between fidelity and sensitivity (R3): We *emphasize* that fidelity and sensitivity are **not isomorphic**. The easiest way to see this is that a constant explanation with minimum sensitivity has high infidelity. Low infidelity means that the dot product between the *fixed* explanation and the perturbation $\mathbf{I}$ is close to the function change after perturbation (so that the explanation is able to capture – with fidelity – the effect on the model function given $\mathbf{I}$), while low sensitivity means that the explanation does not change much after small perturbations to the input. Thus, while both involve perturbations, their mathematical characterizations are quite different (for fidelity perturbation is for function, and for sensitivity perturbation is for explanation). The key reason that we can reduce both sensitivity and infidelity is that sensitivity of many explanations (with dot product with $\mathbf{I}$) is higher than prediction sensitivity (which is observed [11]), which makes C1 (in l249) small and leads to a lower upperbound in Theorem 4.2. One relationship between sensitivity and infidelity is shown in Appendix C, where we show that "a robust version infidelity will be large if sensitivity of explanation is much larger than prediction sensitivity". Again, this relationship only holds when sensitivity of explanation is (much) larger than prediction sensitivity. Therefore, we emphasize that while we can improve infidelity and sensitivity, this is not an universal statement and the relationship between sensitivity and fidelity is non-trivial.