Thank you all for your thoughtful reviews. The following document seeks to address some reviewer concerns:

Motivation and Contribution (All): Reviewer 3 succinctly summarized our contribution in his review: "which explanation approach provides explanations humans agree with the most, on average." Our study was not intended to reveal which explanation is of superior quality; we apologize if that was the perceived motivation. Instead, the aim is to offer insight as to which of the popular explanation methods are more appealing to an average end-user who may not possess knowledge of machine learning, i.e. the "non-expert" layperson. This distinction will be made more explicit in a camera-ready draft.

Visual Explanations (Rev 1, 3): Reviewers 1 and 3 astutely note that explanations are not inherently constrained to the visual domain. We acknowledge and agree with these observations; while not explicitly stated, our study was primarily restricted to the visual domain. Given that visual explanations are the most widely-used of available explanation methods, we restricted our focus to these methods, thereby limiting the overall scope of the study. When referring to "visual explanations" in our analysis and unification, we refer to an explanation that is perceived through sight (including text), not that the explanation is necessarily an image. Future works that introduce explanation methods spanning alternative domains would certainly mandate further analysis. We would be sure to clarify this constraint in a camera-ready submission.

Casting Other Domains to Image Classification Explanations (Rev 3): As Reviewer 3 mentions, casting image explanations to time-series data is undoubtedly a sub-optimal solution. Nevertheless, a survey of the current literature revealed that these explanation methods are frequently applied to these domains [1][2][3]. An emphasis of the shortcomings of existing solutions coupled with the need for explanation methods specifically targeting these input domains will be included in a camera-ready submission.

Unification of Explanation Presentation (Rev 1, 4): We agree with Reviewer 1 that our unification does not encapsulate explanations from interpretable architectures, adaptive NNs, and methods exploring the intra-layer and inter-layer statistical properties. Our unification encompasses visual explanation methods for open-loop, forward-path inferences given a pre-trained model and a test input. We will be sure to offer an appropriate positioning of our unification with respect to the literature and focus more on improving the textual explanation, as per the advice of Reviewer 4.

IRB Approval and Study Size (Rev 2, 4): Our study was submitted for IRB approval and received exemption. We chose to compensate participants according to the advocated minimum wage of $15 per hour [4]. Given our limited budget, this restricted the scale of our study, but nevertheless provided statistically significant results when considering either bootstrap confidence or population sampling confidence intervals. Related works surveyed at a similar magnitude [5]. These details will be included in our camera-ready submission.

Ensuring a Fair Comparison (Rev 3): For all of the explanation methods that required a specific layer input to produce an explanation, we universally selected the last convolutional layer to ensure a fair comparison. This selection for saliency methods is grounded in recommendations from prior work (please refer to submission citation 8). As such, the same layer is selected for the NN-query in explanation-by-example. Our survey validation questions specifically accounted for how often the user agreed with the model prediction; we only removed fast submissions if they were physically impossible to achieve (e.g. the survey participant used a bot to autocomplete the form). The statement regarding "well-formed" was not a quantitative assessment and will be retracted.

Methods Selected (Rev 3): We agree with Reviewer 3 that including localized and compact explanations for image classification are desirable for a more informative study. It is for this reason that we specifically selected Grad-CAM++ over Grad-CAM, as it also advocates for a more localized and compact explanation. As opposed to evaluating across saliency methods specifically, our study was intended to provide a comparative analysis across the various approaches to explanation as they are perceived by the average human. Given the wide body of available methods in the literature, we aimed to select the necessary but sufficient subset of methods that were most commonly explored in related works and explainability studies. The NN-query method specifically used in the explanation-by-example implementation is based on a number of prior works (e.g. [1], submission citation 17).

Related Works (All): Thank you all for pointing out these important related works. We will expand on our related works section to include all references to explainable NNs, methods that observe statistical properties of a DNN, other saliency methods, and the related explainability studies. We will be sure to include them in the camera-ready submission.

[1] Gee et al. "Explaining deep classification of time-series data with learned prototypes" 2019
[2] Arnout et al. "Towards A Rigorous Evaluation Of XAI Methods On Time Series" ICCVW 2019
[3] Assaf et al. "Explainable Deep Neural Networks for Multivariate Time Series Predictions" IJCAI 2019
[4] Williamson et al. “On the Ethics of Crowdsourced Research” PS: Political Science & Politics, vol. 49, no. 1, 2016
[5] Hase et al. "Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior" 2020