Dear reviewers, thank you for your feedback. We'll take them all into account. Below are our responses to major points.

Link between Interpretability and Ground Truth Annotations (from R5) The link between interpretability and annotation labels comes from the concept of *coherence* for explanation evaluation [1]. Coherence means the consistency between an explanation and human prior belief. Here, the annotation label is the relevant prior belief since it is a human-defined and domain-specific label deemed most appropriate for a feature group. In fact, [2] demonstrated that attribution methods that correlate - or cohere - well with annotation labels also increased human trust in model predictions. We used this same annotation-based metric in our paper at lines 226-235. Practically-speaking, coherence is important so that users can agree with model interpretations. Without coherence, a user may inadvertently think that a model fails when it actually works, e.g. the motivational Fig. 1. We will include this discussion in our paper revision.

Runtime Experiments (from R5) We have runtime experiments in Appendix J (referenced on line 225). We will add a summary of the runtime result in the main paper. Our proposed method outperforms the state-of-the-art.

Assumption 5 Clarifications

- **Text Example (from R1):** The interactions {a, very, good} and {not, very} are captured by the contexts of $\mathbf{x}^\star$ and $\mathbf{x}'$. Assumption 5 argues that we can remove/omit any special higher-order (>3-way) interaction associated with both target and baseline tokens because such interactions are not very interpretable. It wouldn't make sense for a special 5-way interaction to be created by {_, a, very, good, _} or {not, _, very, _, _} when there could just be low-order interactions {a, very, good} and {not, very} created (which make more direct sense).
- **Synthetic Example (from R7):** This example creates an interaction between exactly 5 target values and 5 baseline values, so indeed Assumption 5 would be violated. The question worth considering here is: does it make sense for conditions to be placed on both target and baseline feature values? Perhaps an interaction between just 5 target values is already meaningful enough, which satisfies Assumption 5.
- **Testing Assumption 5 (from R7):** Our "Interaction Redundancy" experiment in Section 5.2 was designed to show how well Assumption 5 is satisfied via the redundancy of new interactions as the assumption is violated.

On Faithfulness (from R4) We did not claim that the ground truth metric is for faithfulness. In the interpretability literature, *faithfulness* and *coherence* are two different concepts. *Faithfulness* - as R4 suggested - refers to an accurate description of model decision-making [3]. For our discussion on *coherence*, please see our response on "Link between Interpretability and Ground Truth Annotations". Arguably, many interpretation methods are *faithful* according to their special properties, such as our method, Shapley Taylor Interaction Index, and Integrated Hessians via Axioms. Here, we are interested in annotation labels to evaluate *coherence*. (the word correlation metric was used out of respect of [2])

Choice of Interaction Strength Threshold (from R7) The threshold lies on a continuum between the interpretability and completeness tradeoff of an explanation - fundamental to Explainable AI [3]. A principled way to address this question is through an interactive visualization means, where a user is able to adjust the interaction strength threshold (with a slider UI) to understand its effect on explanations. If this is not an option, showing a small number of interactions (compared to the input size) is desirable for explanation simplicity [1]. An automatic way to determine the threshold is to use prediction performance gains of a surrogate interaction model as a guide, where this model is trained on the data samples from ArchDetect and interactions are added to the model until performance stops improving, similar to [4].

Why these Axioms? (from R1) Was the question about how axioms lead to better interpretability? The completeness axiom is designed to tell how much feature(s) impact predictions, and is widely desired for individual feature attribution and some feature interaction methods (Section 6). The Set Attribution axiom essentially allows us to obtain independent attribution scores $\phi(\mathcal{I}_i)$ for different disjoint feature sets $\mathcal{I}_i$ due to the additive structure of a function, so we can do the analysis of Fig. 2c on each of those feature sets, thereby leading to more interpretable results.

Why subtract out the baseline in ArchAttribute? (from R2) To satisfy axioms. The intuition for ArchAttribute's interpretability was shown in Fig. 2c. Indeed, our experiments in Table 2 showed that ArchAttribute is interpretable.

How ArchDetect works for Image Classification? (from R2) Please see discussions in Appendix B (and line 196).

ArchAttribute with Other Interaction Detectors (from R7) In interest of space, we will include these comparisons in paper revisions. One can still judge the different interaction detections in Appendix K by manually merging them.

Hessians on many-times differentiable models (from R7) Table 2 shows Integrated Hessians (IH) on BERT which uses many-times differentiable GELU activations. IH is computationally prohibitive to run for image classification.

[1]. Miller, T. *"Explanation in Artificial Intelligence: Insights from the Social Sciences"* in *Artificial Intelligence* **267** (Elsevier, 2019), 1–38. [2]. Jin, X. et al. *"Towards Hierarchical Importance Attribution: Explaining Compositional Semantics for Neural Sequence Models"* in *ICLR* (2020). [3]. Gilpin, L. H. et al. *"Explaining Explanations: An Overview of Interpretability of Machine Learning"* in *DSAA* (2018), 80–89. [4]. Tsang, M. et al. *"Feature Interaction Interpretability: A Case for Explaining Ad-Recommendation Systems via Neural Interaction Detection"* in *ICLR* (2020).