1 Transductive Results on Higher Dimensional Inputs

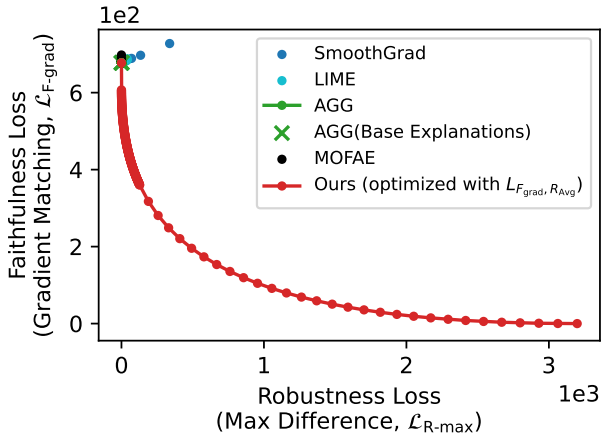

Figure 1: **Faithfulness vs. Robustness.** Comparison of our method against baselines for a ResNet (1) model on 10 images from ImageNet (2). We observe that our method provides explanations that are Pareto-optimal and capable of managing trade-offs between properties. In contrast, the baselines produce explanations that are not optimal and concentrated in a limited region of the Pareto front(upper left corner).

2 Cross Evaluation: Directly Optimized over one set of Properties and Evaluated on another set

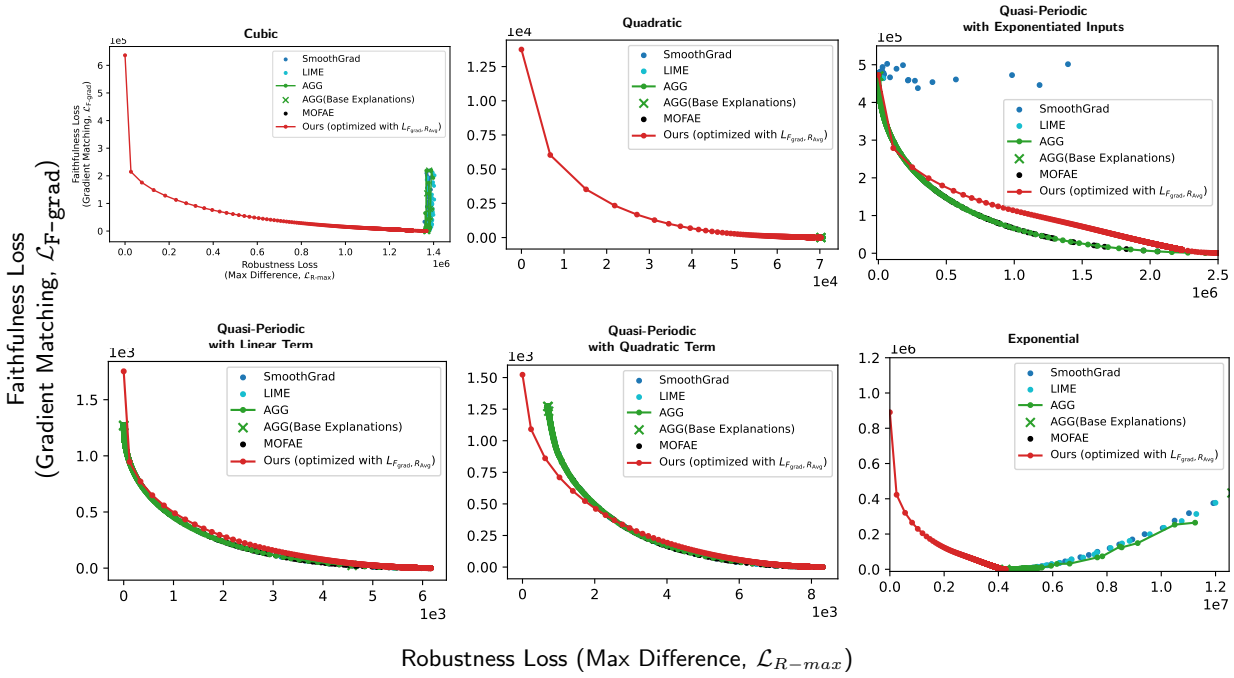

Figure 2: **Faithfulness vs. Robustness.** For AGG, MOFAE, and our method, we optimized over Gradient Matching faithfulness and average robustness formalizations, and evaluated using Gradient Matching faithfulness and the maximum distance formalization of robustness. We observe that for power functions such as quadratic and periodic, our method provides optimal explanations—even though it was not directly optimized for those robustness definitions. While our method is not as optimal for quasi-periodic functions, it still produces explanations that manage trade-offs between properties in a more controllable way compared to the baselines.

3 MLP Results

We trained neural network models with one hidden layer using the (3) dataset. The results show that our method provides more optimal solutions than the baselines. Furthermore, they highlight the limitations of AGG and MOFAE, as these methods are highly dependent on the quality of the base explanations used to generate the explanations.

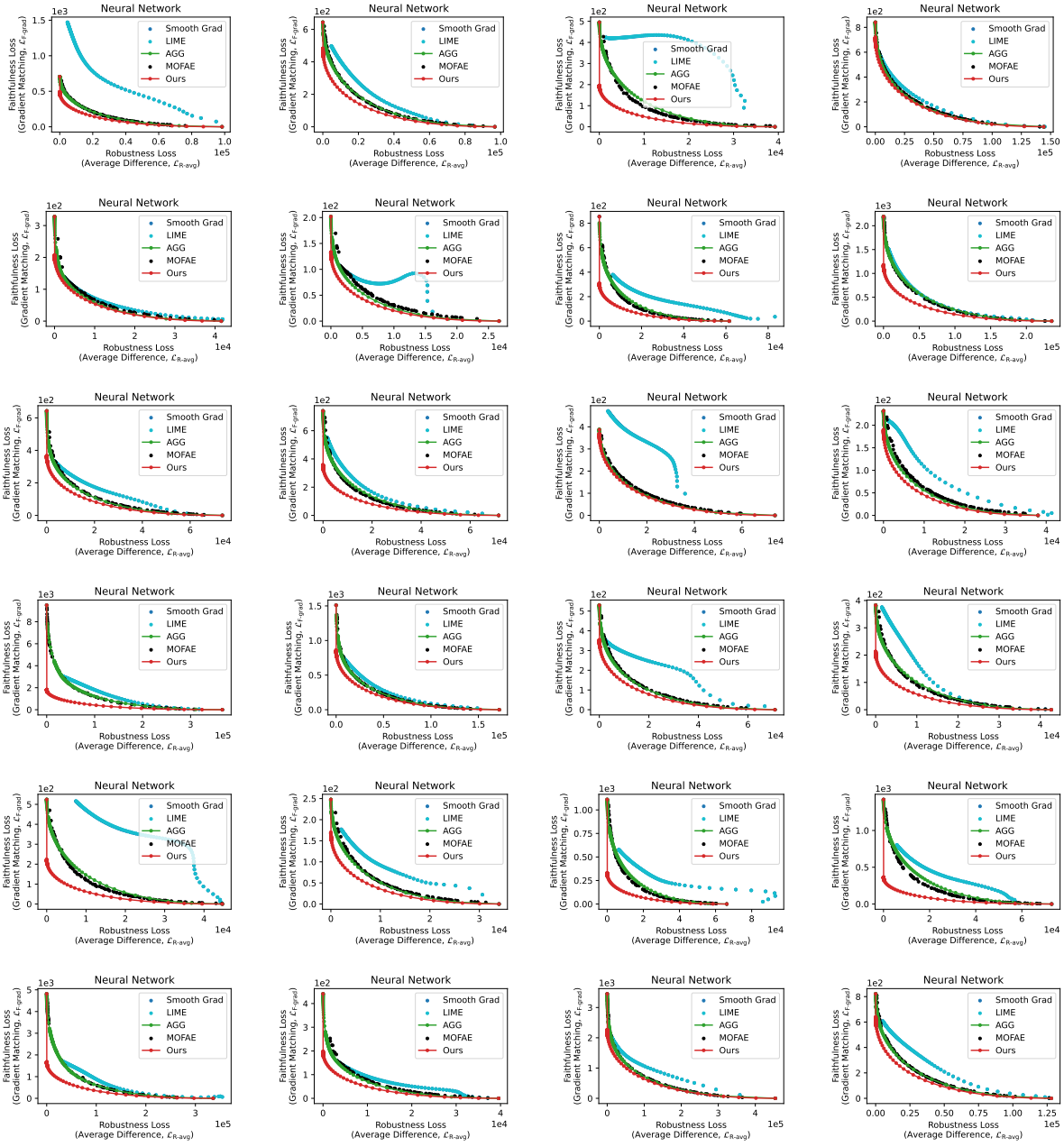

Figure 3: Neural network explanations (1–24).

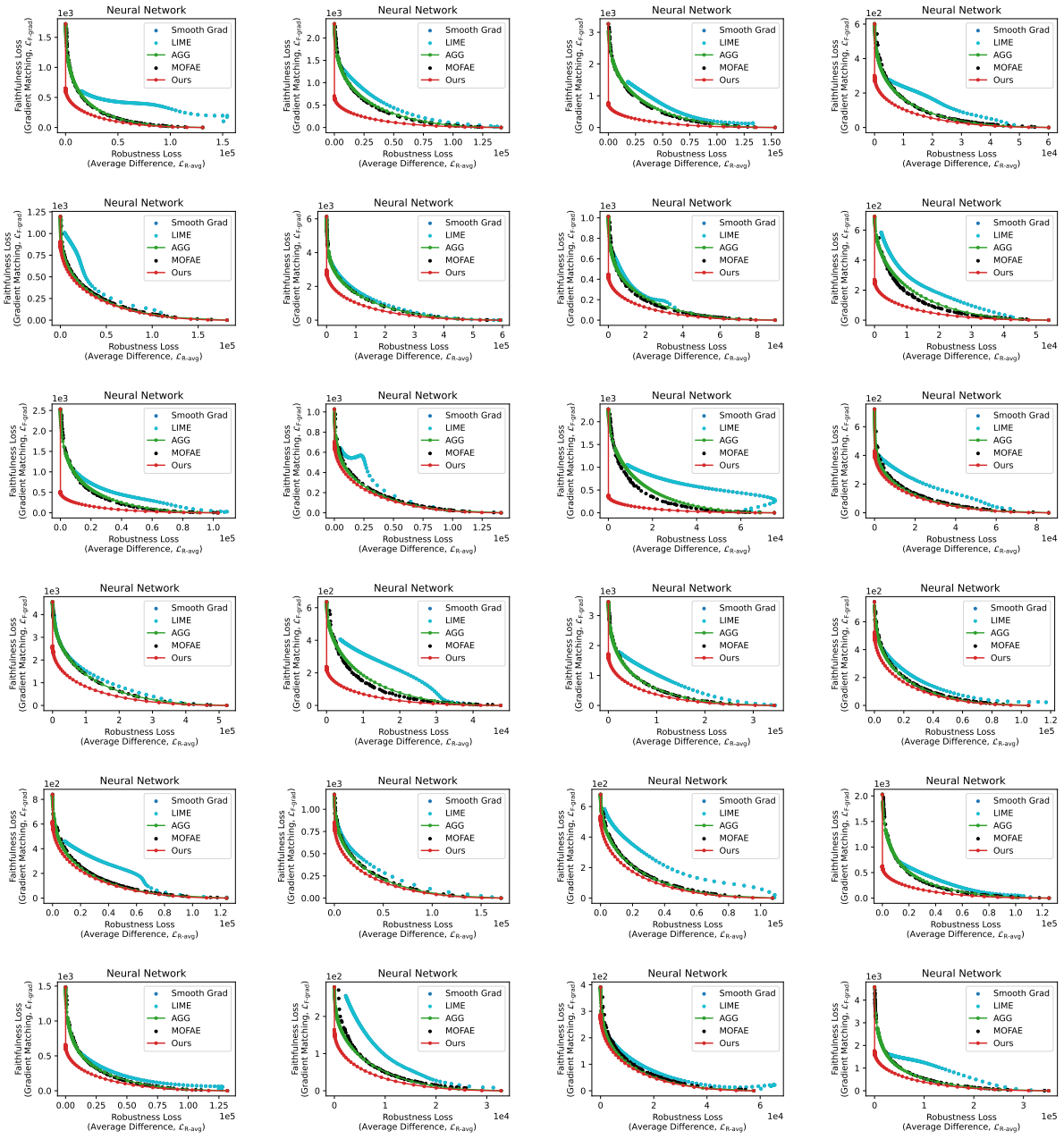

Figure 4: Neural network explanations (25–48).

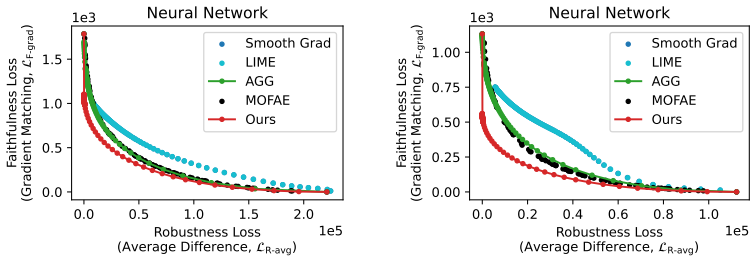

Figure 5: Neural network explanations (49–50).

4 Qualitative Comparison

We compared our method with baselines based on agreement on the top-1 most important feature, where a score closer to 1 indicates stronger agreement. Our method demonstrates consistently high agreement across varying trade-off hyperparameter values (λ), whereas the baselines show comparably lower agreement scores regardless of changes to their respective hyperparameters.

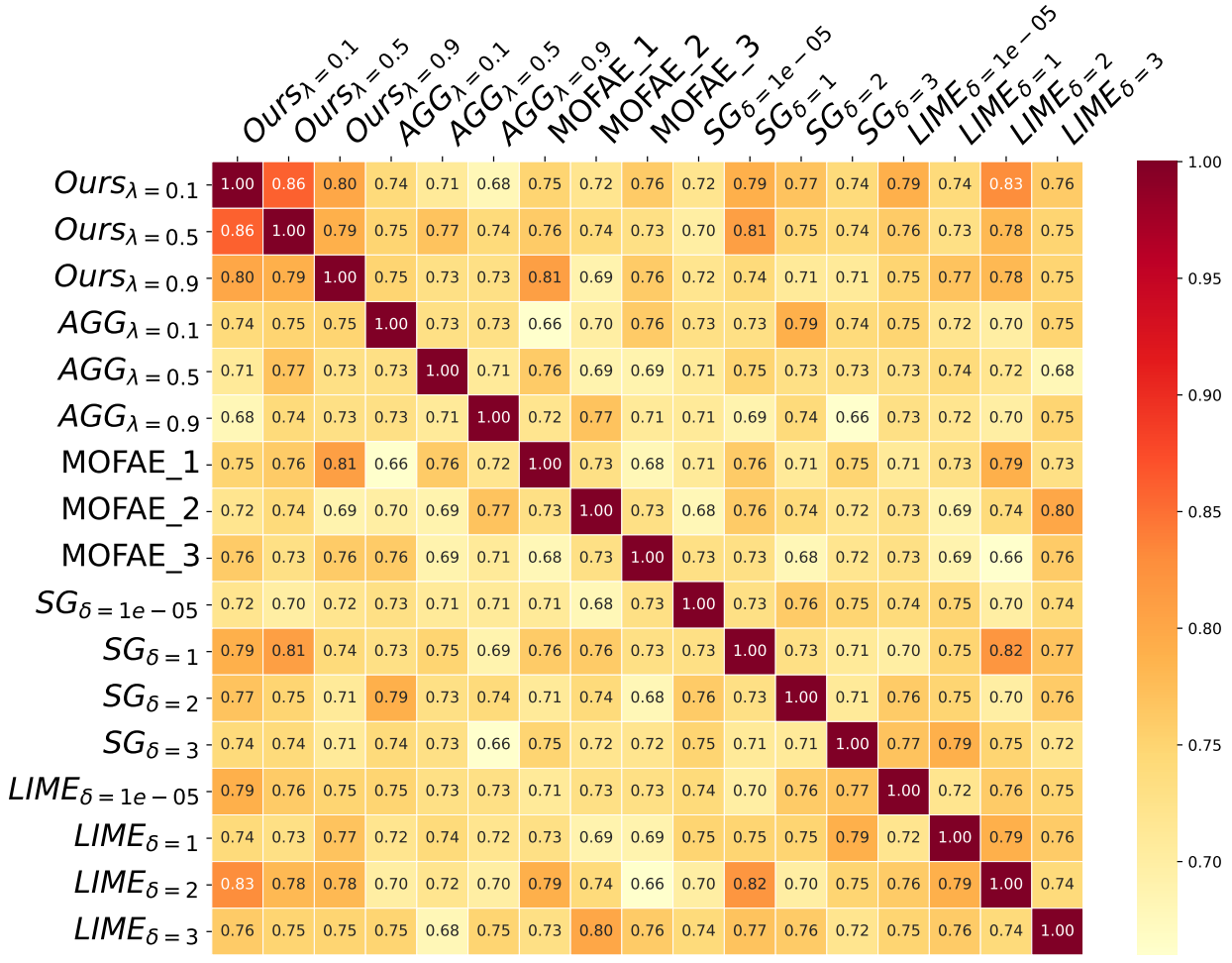

Figure 6: Comparison using Agreement on the top feature for a cubic function.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[3] "Solar Flare." UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C5530G.