We thank all the reviewers for helpful comments and we address the major concerns in the following.

(R1, R2) The use of CNN backbones in our visual reasoning tasks. The CNN backbone is capable of processing the visual stimuli of our BONGARD-LOGO tasks. We first demonstrated this in the ablation study in Appendix, Section C, where Table 2 shows that the best train/test accuracies on free-form shapes are 97.6% and 89.0%, respectively. This implies that isolating from the concept reasoning properties of *context-dependent* and *analogy-making* perception, the CNN backbones can achieve high recognition performance on our visual inputs. Furthermore, we performed standard supervised experiments on human designed shapes by training two separate binary attribute classifiers for `convex` and `have_two_parts`, respectively, using the same CNN backbone as in the paper. With each dataset of randomly sampled 14K positive and 14K negative samples, the model achieved the near-perfect train/test accuracies 98.7%/97.0% on `convex` and 98.0%/97.1% on `have_two_parts`.

(R2) Differences with RPMs. Our benchmark is complementary to RPMs: RPMs focus on relational concepts (such as progression, XOR, etc.) while our BONGARD-LOGO problems focus on abstract object concepts (such as stroke types, abstract attributes, etc.). In RPMs, the relational concepts come from a small set of five relations [17]. In our benchmark, the object concepts can vary arbitrarily with procedural generation. The three core properties (i.e., context-based, analogy-making, few-shot with infinite vocabulary) of human cognition captured by our benchmark also define a new challenge for current object perception methods. Besides, the model performance of 50-60% in our benchmark is not as high as it might seem. Each of our tasks asks for a binary decision, where chance performance is 50%. The majority of the model performances in Table 1 are within 10% better than chance. In contrast, the model performance of 60-70% accuracy achieved on RPMs [17] is much higher than its chance performance of 12.5% accuracy.

(R2) Visual reasoning on V-PROM and real images. Thank you for bringing up V-PROM, an interesting work on extending RPMs to real images. As V-PROM is a real image version of RPMs, it differs from our benchmark in the similar ways described above. We chose synthetic data because a) Bongard designed the original problem sets with simple line drawings and geometries, and we follow the same design principle. b) Furthermore, procedural generation of synthetic data gives us the precise control of concepts and nuisances, making a systematic study of model performance on Bongard problems feasible. Some early attempts of building our benchmark were indeed based on real-world images, such as CelebA and MS-COCO datasets. However, the quality is hindered by uncontrollable confounders in real images, resulting in ambiguous concepts.

(R2) The level of generalization in the test sets. Similar to prior work, the level of generalization in our benchmark can be varied flexibly, and models can be evaluated at multiple points. We chose "+1 extra action stroke" and "+1 more straight line" in the held-out sets as two basic cases for extrapolation. As these two tasks have been shown difficult for current methods (i.e. performances close to chance in Test Acc (FF) and (NV) as shown in Table 1), we did not include more challenging setups. We will expand our discussions on this in the future revision.

(R2, R3, R4) Strong philosophical claims. The statements of objectivism in traditional AI [14] and metaphysical realism in Bongard problems [15] are referenced from prior work. We included them to highlight the intellectual roots of our work and its connections to other disciplines. We concur with the reviewers and we will tone down our language, focusing on the technical merits of this work in the future revision.

(R3) Problem set in BONGARD-LOGO versus the original Bongard problem set. Taking advantage of procedural generation, we are able to produce an infinite amount of free-form shapes, which the original set does not have. The procedural generation method can be easily extended to incorporate new shape features, including filling-in of enclosed areas. We will also open-source our procedural generation code for the research community to create their own shapes.

(R4) Ablation study on model sizes versus performances. We vary the model size by dividing each layer size in the ResNet-15 backbone by a reduction factor α (α is a divisor of the number of parameters). Figure 1 illustrates the results of two top-performing models: ProtoNet and Meta-Baseline-SC. We see that Train Acc decreases consistently with smaller model sizes. On the test sets, results slightly vary across different models. ProtoNet is more sensitive to model size than Meta-Baseline-SC: All the Test Acc of ProtoNet tend to decrease with smaller model sizes, while some Test Acc (BA and CM) of Meta-Baseline-SC remain robust. We will add these ablation results in the future revision.

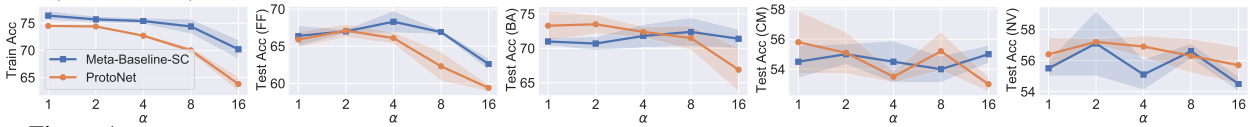

Figure 1: Performance of Meta-Baseline-SC and ProtoNet under different model sizes (model size is smaller with larger α).

(R4) Traditional ILP methods for BONGARD-LOGO tasks. ILP does not offer complete solutions to Bongard problems. The best ILP method so far [13] has relied on manually specified rules to tackle a subset of 39 carefully selected Bongard problems, making it infeasible for our benchmark which consists of tens of thousands of problems. We intentionally avoid highly customized methods that may have low generality and limited applicability, which deviate from the motivation of our benchmark as a step towards inspiring more robust and reliable AI systems. That being said, we very much agree with R4 that a *hybrid* system that combines data-driven and symbolic methods is a promising future direction for both generality and reliability, as we discussed in Section 5.