We thank the reviewers for their constructive feedback and suggestions. We deeply appreciate their efforts in these difficult times. It is very encouraging that the reviewers found our dataset valuable (R3) and well-tailored (R2); our data collection strategy rational (R3) and founded (R1); construction of minimally contrastive pairs clever (R3); and our paper clear and well-written (R1, R2, R3).

R2: This paper does not provide a proper solution to fully leverage this dataset to advance the relation reasoning from 2D. Proposing a proper solution to relational reasoning in 2D is important but out of scope for this paper. We hope that our work can help catalyze such advances. Although our primary contribution is a dataset for grounding spatial relations in 3D, we have provided insights that can be useful in advancing relation reasoning from 2D. Our results suggest that estimating 3D configurations could be a promising first step for understanding spatial relations in 2D. Also, our dataset exposes that recent 2D methods overly rely on bias, and hence can help build better models that overcome these shortcomings.

R3: Discuss how the work relates to earlier works by Chang et al. Thanks for bringing these works to our notice. Indeed the earlier works by Chang et al. are related to ours. In [Chang et al. EMNLP 2014], the authors model spatial knowledge by leveraging statistics in 3D scenes. For spatial relations, they create a dataset with 609 annotations between 131 object pairs in 17 scenes. We on the other hand, collect a relatively larger dataset for spatial relation grounding with 10K annotations, across 5K object pairs in 5K scenes. In [Chang et al. ACL 2015], the authors focus on creating a model for generating 3D scenes from text, and create a dataset of 1129 scenes from 60 seed sentences. We also create 3D scenes, but unlike [Chang et al. ACL 2015] our dataset focuses on only spatial relations not other properties like object color. Moreover, unlike [Chang et al. EMNLP 2014, Chang et al. ACL 2015], scenes in Rel3D occur in minimally contrastive pairs which control for potential biases like language bias. Finally, the objects in [Chang et al. EMNLP 2014, Chang et al. ACL 2015] are limited to those found in indoor scenes like chairs and tables, while we also consider outdoor objects like trees, planes, cars and birds. Hence Rel3D covers a wider array of spatial relations. We will add this discussion to the paper.

R3: How were specific object instances selected and were they pre-specified for a given crowdworker? The specific object instances given to crowdworker were based on the expert annotator' response. In the first stage of data collection, we showed randomly-selected object instances to the expert annotators. Based on the object instances, the expert annotator marked if a particular relation between them is natural. When asking the crowdworker to create the scene, we gave them the pre-selected object instances. Also, the crowd-workers had an option to select "Not Possible", if they deemed that the relation could not hold between the object instances. However, this option was used very infrequently ($< 5\%$ of times).

R3: Rationale for subsampling only 1/4 of all relations and what relations are excluded due to that? This is due to the limited budget for expert annotators. There are 134,670 possible relations and it would be too costly to annotate each of them as natural or not.

R3: To my understanding, a crowdworker's judgment that a relation is not valid in a presented image is deemed to imply that the relation is not using an intrinsic frame of reference – is this a reasonable assumption (i.e. are there other cases when a crowdworker might say the relation does not hold)? The assumption is reasonable because the same worker is asked to judge the relation from a different view that would negate the relation unless the worker is using an intrinsic frame. While collecting a scene, we first ask a crowdworker to place the subject and object so that they satisfy the relation. Once done, we show the same worker an image taken from a different camera view. For "to the left of", "to the right of", "in front of" and "behind", the new camera view is directly opposite from the initial one, i.e. from the back wall. For the relation "to the side of"; the new camera view is from the left wall. Since the task and the image are shown in quick succession, we assume that the worker would use the same frame of reference. Hence, if the worker did the task correctly but says that the relation is not valid in a different view, it implies that they did not choose an intrinsic frame of reference (intrinsic to the object) while doing the task, but rather did the task from an extrinsic frame of reference that is dependent on the camera view.

A possible corner case could arise if one object is completely blocked by another in the second view. But this is largely avoided by using a slanted topish view (as shown in Figure 2). Lastly, in the final filtering step, we ask independent workers to verify the relation while taking into account the inferred frame of reference. Hence any erroneous sample is filtered out in the final dataset.

R1: The scenes involve only two objects and thus not so realistic. We agree that more objects can improve realism. With 2 objects per scene, our data is still useful because it provides a stepping stone for more different cases. In addition, Rel3D can be easily augmented by inserting distractor objects or by merging multiple scenes, without collecting additional human annotations.