We thank all reviewers for their constructive comments. Due to space limit, we address the major concerns as follows.

Synthetic dataset (R1) - We note that many of the (synthetic or real) datasets prepared in robotic grasp learning research are not made fully public. For example, configuration of the simulation environment used in the closely related work of 6-DOF GraspNet [23] is not publicly available, which makes it difficult to directly and fairly compare different methods. This is the main reason why we have to prepare our own synthetic data. In terms of additional features of our dataset other than the doubled size, more categories and instances, a valuable one is that our dataset has antipodal label for each grasp sampling, which supports physically sensible models such as GPNet. To contribute to the community, we will make all our dataset, including configuration of simulation environments, publicly available.

Results of GPNet-Naive (R1) - In GPNet-Naive, we still use anchor coordinates in the grasp proposal phase, but not for grasp regression. The grasp proposals are distinguishable because we use an anchor-dependent method for feature extraction (Sec. 4.2). Without anchor coordinates as input features, the Antipodal Classifier does not work well and the regression loss is much higher than that of GPNet. We will improve the description of GPNet-Naive in the paper.

Novelties of GPNet (R1, R2, R3) - To our best knowledge and as R4 said, our work is the first one that uses 3D anchors to generate 6-DOF grasps. Our grasp parametrization using surface contact, grasp center, and 'pitch' angle (L149) enables definition of anchor based grasp proposals, which is intuitive and physically sensible. With this parametrization and grasp proposal module, we only need to regress 4 parameters of center offset and 'pitch' angle, instead of 6 ones as in [23]; reduction of variables makes learning much easier. Our novelty also presents in design of antipodal validity -> grasp regression -> grasp classification (cf. Fig.1), where we prune grasp proposals by antipodal constraint and take the regressed centers and angles as input when scoring the grasp candidates, while [23] does not consider antipodal constraint, and [4,22,28] output grasp candidates and their confidences in a parallel way, which is suboptimal for ranking grasps (L211-L213). In addition, our anchor-dependent feature extraction differs from prior works, and our GPNet is trained end-to-end, while [23] trains its Grasp Sampler and Grasp Evaluator independently.

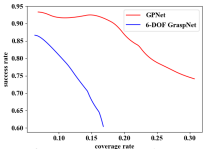
Pruning in the beginning (R2) - Thanks. We have in fact tried an affordance detection module in GPNet, where contact points suggested by affordance detection are picked for grasp proposal and the rest ones are removed. On average, 32% prediction time is saved with no sacrifice of success rate. We will include these results in the paper.

Rule-based evaluation (R2) - We measure the rotation difference by $\Delta_\theta = 2\arccos \mathbf{q}_1 \cdot \mathbf{q}_2$. For ease of understanding, we formulate it as L109. We have tested both and there is no much difference. We will revise the text accordingly.

More empirical evaluation on model variants (R3) - Thanks for the suggestion. We have in fact ever tried other modeling choices such as 2D Depth CNN: we use 2D Depth CNN and 2D grids (2D version of GPNet) to generate 6-DOF grasps, but it only produces 13.4% success rate@10% and 4.3% coverage rate@100%, much worse than results of GPNet in Tables 1 and 3. For other choices such as voxelizing point cloud as input of 3D CNN, we note that it would sacrifice precision of 3D location, as verified by recent methods on KITTI benchmark of 3D object detection; it is harmful for precise regression of grasp pose as well. We will include these additional evaluations in the paper.

More experiments on GPNet-Naive (R3) - We just pick one GPNet-Naive variant to show the importance of anchor coordinates as input for antipodal classification and grasp regression. Success rates@10% of GPNet-Naive are respectively $0.050, 0.119, 0.100, 0.061, 0.053$ when $(r, b) = (3, 22cm), (7, 22cm), (10, 22cm), (10, 10cm), (10, 30cm)$.

Standard evaluation metrics (R3) - Without losing fairness, we have selected 4 representative points on the curve of success rate vs coverage rate to report in the paper. A full curve is shown as the figure right; AUCs (Area Under Curves) of GPNet and 6-DOF GraspNet are 0.206 and 0.081, respectively. Note that we use a stricter criterion for coverage rate, where both center and rotation distances are considered (only center distance is considered in [23]).



Real-world evaluation (R3) - Thanks for the suggestion. Although our setting of real-world experiments is almost identical to that in [23], we will pursue more thorough real experiments in future research.

Computational analysis (R4) - Computations of existing methods are evaluated based on generating $10k$ grasps per object. Our GPNet takes $\sim 2.1s$ from input to scoring the generated grasps, while GPD [33] takes $> 10s$ for grasp sampling and grasp classification, and 6-DOF GraspNet takes $\sim 10s$ from grasp generation to grasp refinement.

Single-mode issue of 6-DOF GraspNet [23] (R4) - Grasp proposal of [23] is based on a trained variational auto-encoder (VAE). VAE suffers from an issue of "latent variable collapse" [Dieng et al.,AISTATS'19]. The posterior collapses to a simple prior and inference network fails to learn good representations, especially if the likelihood model is powerful, e.g. as in [23]. Thus, directly sampling the latent vector over normal distribution may make predicted grasps concentrate on a single mode. The lower coverage rate of 6-DOF GraspNet in Table 1 also supports this analysis.

Importance of regression module (R4) - Though we prune the grasp proposals based on antipodal criterion, there is no guarantee that the remaining ones are precise enough to be successful, especially for low-resolution scenario. The regression module is to regress a more precise grasp center and free rotation. If increasing the grid resolution for an improved precision, both the number of grasp proposals and the time of antipodal pruning will increase cubically.

Empirical evaluation of focusing or spreading anchors (R4) - In Table 1, we have shown that best coverage rate (more diverse) is obtained when spreading the anchors ($r = 10, b = 30cm$), and best success rate (more precise) is obtained when focusing the anchors ($r = 10, b = 22cm$). We will improve the description in the paper.

Intra-category generalization to unseen objects (R4) - Thanks and we will specify this in the introduction.