We calculate mIOU using masks with the best visible area IOU or best amodal GT IOU. Results show that by sampling (Fig. 4 above) and we also show (Fig. 3 of supp. material) that more occlusion correctly results in wider posteriors. (94.68% vs. 86.03%), mIOU on invisible area only drops by just 2.14% (62.85% vs. 60.71%). We crop images by GT bounding box and train a ResNet-PSP segmentation model. We predict the instance masks on the test set, yielding 80.83% amodal mIOU. We use the predicted instance masks as input to perform amodal completion using both our model and the baseline (Deocclusion [37]), which we outperform. We also find that our model is robust to instance mask corruptions. Even though full mIOU drops by 8.65% (94.68% vs. 86.03%), mIOU on invisible area only drops by just 2.14% (62.85% vs. 60.71%).

R2: Motivation of amodal completion & Amodal instance segmentation is more interesting. We agree that amodal instance segmentation is also an interesting problem. However, amodal instance segmentation can be decomposed into instance segmentation and amodal completion. We provide additional experiments on amodal segmentation (results in Tab. 2) and will update the paper. We crop images by GT bounding box and train a ResNet-PSP segmentation model. We predict the instance masks on the test set, yielding 80.83% amodal mIOU. We use the predicted instance masks as input to perform amodal completion using both our model and the baseline (Deocclusion [37]), which we outperform. We also find that our model is robust to instance mask corruptions. Even though full mIOU drops by 8.65% (94.68% vs. 86.03%), mIOU on invisible area only drops by just 2.14% (62.85% vs. 60.71%).

R2: Lack of quantitative results for multiple posterior predictions. We agree and quantitatively evaluate this aspect (paper will be updated). For each instance, we sample 20 latent codes from the approx. posterior distribution. We calculate mIOU using masks with the best visible area IOU or best amodal GT IOU. Results show that by sampling we find masks that significantly better match amodal GT than using the approx. posterior mode (Tab. 3 above). Hence, the approx. posteriors incorporate diverse plausible masks, correctly capturing the ambiguity. This suggests the use of multiple posterior samples in downstream applications. Sampling is more beneficial for cases with significant occlusion (Fig. 4 above) and we also show (Fig. 3 of supp. material) that more occlusion correctly results in wider posteriors.

R1 & R3 & R4: Results are only shown on KINS dataset. We primarily focus on driving scenes and exploit two datasets with multiple classes, KINS and Cityscapes. KINS contains 7 categories (pedestrian, cyclist, person sitting, car, van, tram, truck) and Cityscapes contains 8 categories (bicycle, bus, person, train, truck, motorcycle, car, rider). We show quantitative results only on KINS, because it is the only available large-scale amodal dataset with accurate human annotations. We plan to show quantitative results on more classes and non-rigid objects for the camera-ready version.

R1 & R4: Comparisons for downstream tasks. We appreciate the suggestions. However, our main contribution is on the amodal completion task. Comparisons to other downstream task methods are beyond the scope of our paper. Nevertheless, we compared FID scores with Deocclusion [37] on the image editing task.

R1: Need to compare with more baselines & Only one category. We apologize for the confusion and will clarify this aspect in the camera-ready version. However, all experiments are conducted on datasets with multiple classes. We use Deocclusion [37] as baseline, since it is the current state-of-the-art model on amodal completion with no amodal ground truth supervision. Also, our user study demonstrates that we outperform the annotation skills of humans, which further strengthens our conclusions. We will add additional details about the user study in the final version of the paper.

R1: Transformation network. The spatial transformation network is end-to-end differentiable and trained as part of the pipeline in stages (2) and (3), using the mask reconstruction loss as supervision (we will clarify in the final version).

R3: Accuracy versus occlusion percentage. We agree. We show results in Fig. 4.

Details, etc: We appreciate the suggestions and will incorporate them in the camera-ready version.