We thank all reviewers for their feedback and insightful comments. We are pleased to see our contributions warmly received: "It's inspiring to see the fusion of interactions and vision in this paper" (**R3**); "The addressed problem of reconstructing 3D shapes, with hand-object interaction, is novel and interesting" (**R4**); and "Overall, I really liked the idea of the paper" (**R1**). We will modify the paper according to reviewers' suggestions and we will introduce their comments with respect to the paper presentation.

R1. Antropomorphization of the motivation: We used human-inspired examples in the abstract and intro to motivate the complementarity of vision and touch when performing 3D object understanding. It was not our intention to antropomorphize the algorithm but rather to provide the task motivation. We will revisit the abstract and intro to clarify this. Model presentation: Our algorithm was built upon the use of graph networks for mesh deformation (e.g. [59] and [52]), due to their *strong performance* in 3D reconstruction tasks, and the *advantageous resolution properties* of meshes. Moreover, using charts enables the *disentanglement of visual and tactile information* and, as result, the model can enforce *touch consistency* in the final prediction. Hence, we introduced our algorithm as integrating charts within mesh-based approaches. Although we could have *alternatively* presented it as adding meshes to the AtlasNet charts, we disagree with the premise that one way of presenting is superior to the other. Nevertheless, we will revise the paper to motivate our approach from both perspectives. Charts: Atlases and charts are well known concepts in topology for describing manifolds and we did not intend to claim our use of them as a novel. We will change the presentation of the intro to reflect this, in particular, we will change "which we call charts" to "called charts [ref]". However, we do want to highlight that, to the best of our knowledge, our approach is the first to use these concepts for joint reconstruction from vision and touch signals. Universal approximation: As pointed out by **R1**, this is not a big limitation. However, the proof for universal approximation of AtlasNet relies on sampling and then passing sufficiently many points to an ideal MLP to epsilon-approximate a manifold. The FoldingNet decoder uses a single 2D grid of fixed points and transforms each point independently. By contrast, our model deforms the vertices of many charts by aggregating neighboring information at each layer of the graph network (see Eq. 1). We will add the suggested references. Pose information: The hand pose is available to the method for inference, and all charts are predicted within hand's reference frame. UNet role: The UNet model does currently model a non-invertible mapping due to smoothing in the rendering process. # iterations: 3 refinement iterations were used based on empirical validation, and backed by prior works [59, 52].

R2. Simulation simplicity: Our simulator is based on objects coming from a standard 3D reconstruction dataset - ShapeNet. While it is true that the objects are not highly complex (e.g. in terms of surfaces and scales), our dataset is the first of its type. In addition, despite its simplicity, the dataset was sufficient to clearly demonstrate the benefit of touch for 3D reconstruction. Touch information: In our setting, touch is useful for completely reducing the uncertainty of a surface's local position and structure. Adding surface details for 3D reconstruction would probably not lead to large performance variations as they amount to very little variation in overall shape.Touch-only experiment in Tab.2: Yes, in this experiment we report results for a single grasp. We observed that models leveraging touch only reach almost perfect reconstruction on touch site but poor global reconstruction quality, confirming that the benefits are local. Comparison to prior work: The comparison for a variety of visuotactile reconstruction baselines is reported in Tab. 1. We do not compare directly to prior work, as we are unaware of work directly applicable to our setting (see l. 93–104), though we would have been happy to compare to or discuss the relevance of other work, had **R2** highlighted them.

R3. Direct use of intermediate representations: We consistently found our approach to outperform models naively conditioned on intermediate touch representations, e.g. Tab. 1 (rows 9 and 11) where the reconstruction algorithm receives intermediate features from the touch signals directly and yet performs notably worse. We hypothesize that this happens due to information imbalance in vision vs. touch for the reconstruction task. Touch only setting: Touch signals are exclusively passed to the algorithm (no vision). Thanks for the pointer, we will resolve this ambiguity.

R4. Limitations: (1) Predicted charts sometimes poorly overlap creating a noisy boundary, as opposed to smoothly connecting (see Figure 8). (2) Charts are not forced to form a continuous, connected surface, and so from our qualitative evaluation $\sim 5\%$ of predictions possess a chart detached from the remaining surface. (3) Our algorithm's dependence on full 3D scene information, which is perhaps unrealistic for un-simulated scenarios. We will include this discussion in the paper. Vertex neighborhood/charts: Vertex neighborhood is defined as the set of other vertices which share an edge with it *within the same chart*. Given this neighborhood definition, charts *do not share edges*, and so we enable communication among charts as described in l. 140–152. Interpreting of touch signals: Touch signals often look similar as the finger might only touch the object lightly, or the object may be smooth, making the touch reading not human-interpretable. Note that additional touch readings are depicted in Fig. 2 in the appendix. Simulating touch: The Phong model is not used to compute the depth signal, but to render a touch sensor reading from the depth representation. We provide a complete description of this process in Section 1.5 of the appendix. Prediction resolution: Decreasing the resolution of charts (# faces) decreases the representation power of the method and so reduces performance, while increasing it boosts performance, though only up to a limit. This is a property common to all representations in 3D reconstruction. Single modality comparison: For the vision only single modality setting see Section 3.6 and Tab. 4 in the appendix. Simulation simplicity: See R2 answer. Fig. 3: It depicts a whole shape made of multiple charts. Thanks for pointing this out, we will resolve this ambiguity.