We thank the reviewers for their thoughtful comments! We are encouraged that all reviewers voted to accept, finding the paper to be creative and well-motivated [**R1**], extremely clear / well-written [**R1 R2 R3**], and that the technical contributions of the paper – a novel model for language-conditioned instruction following – was judged to have medium-high [**R1**] to high [**R2**] significance. We address specific concerns below (and will incorporate all feedback).

[R1] Comparison to prior work. Thanks for pointing out the relevant ablated comparisons in [4] and [7], which achieve val-unseen success rates (SR) of 42% and 43.6% respectively. With additional hyperparameter tuning, we found the results of our model increased to 33% SR / 30% SPL on val-unseen (up from 31% / 27%). Compared to these works, we use less powerful ResNet-34 image features (instead of ResNet-152) and our model is much less dependent on the structure of the navigation graph (see L36-41 below). We will update Table 2 with these comparisons instead of [1].

[R1] Impact of agent heading. Excellent point for which we will add discussion. Heading indicates important cues about a trajectory which the model can leverage in the belief state. For instance, is is unlikely for an agent following the true path to turn 180 degrees midway through (unless this is commanded by the instruction). Similarly, without knowing heading, the model can represent instructions such as 'go past the table' but not 'go past with the table on your left'.

[R1] Are the mapper-filter experiments in Sec 5.2 produced by training without the policy? Are the filter and semantic map parameters updated while training with the policy in Sec 5.3? Yes, to both. In Sec 5.2 we train without the policy, using only the filter loss (KL-divergence between the predicted belief $\boldsymbol{b}_{1:T}$ and the true state $\boldsymbol{s}^*_{1:T}$). In Sec 5.3 we train concurrently with both the filter loss and the policy loss (cross-entropy loss to maximize the likelihood of the ground-truth target action). In both cases we train all parameters end-to-end (except for the pretrained CNN). We have verified that the stand-alone performance of the filter is not unduly impacted by the addition of the policy.

[R1] Is the true state sequence $s^*_{1:T}$ (L245) always the human trajectory, or does it include the exploration that is done by the agent during training? Yes it is always the ground-truth (human) trajectory. Recall that the filter subcomponent is simply trying to estimate the path of an ideal agent following the instructions, given a partially-observed semantic map $\mathcal{M}$. However, the agent's exploration does determine the content of the map (i.e., the map contains the observations from the agent's explorations in the current episode). We will clarify.

[R1] No 'direct' access to the instruction and map We will add this qualification at L238.

[R2] "Middling" performance compared to SoTA. As in the paper, we freely admit that the proposed method does not approach SoTA performance, but we agree with the reviewer guidelines in that *"Solid, technical papers that explore new territory or point out new directions for research are preferable to papers that advance the state of the art, but only incrementally."* Moreover, our novel formulation for instruction-conditioned goal-identification in VLN as state tracking provides increased inspectability of agent beliefs and reasons over 2D space rather than just navigable nodes. As such, it may be of broader interest outside the VLN task.

[R2] Lack indications that this architecture will lead to SoTA. The performance of the model improved with additional hyperparameter tuning (see L5-9 above). Given the rebuttal period, it was impractical to incorporate the orthogonal additions like RL training, data augmentation from a trained speaker model, or trajectory re-ranking which have lead to much of the advancement at VLN.

[R2] The action space of the agent... requires significant domain and task knowledge in the form of the structure of the navigation graph. We would like to clarify that the navigation graph is defined by the R2R dataset [1]. Previous works [2-8] use LSTMs that operate directly over the nodes in this graph. Our approach is far less reliant on this structure than prior work (at test time, our mapper and filter operate on arbitrarily discretized space, with no knowledge of the navigation graph). We anticipate this could be an advantage for sim-to-real transfer (i.e., in real robot scenarios where a navigation graph is not provided, and could be non-trivial to generate).

[R3] How does the state rejection mechanism compare to the one of Weib et el. arxiv.org/abs/1511.06458. Can one think of the language component as a refined prior? We are not sure what the reviewer is referring by a "state rejection mechanism" in our model, but are happy to hear more details in an updated review. With regard to Weib et al. - we do not need to employ such rejection or importance sampling techniques because we use a histogram representation for the posterior, not a set of weighted samples as in particle-based Bayesian state tracking. At a high-level, we agree that language does define the set of expected observation and actions an agent should take throughout the trajectory.

[R3] At L134 what are XY? X and Y are the initial spatial dimensions of the semantic map in grid cells. In experiments this dimension is 96 (refer supplementary). In practice the map could be dynamically resized if necessary.

[R3] What does 'strong baseline' in abstract refer to? We are referring to the LingUNet model and the Hand-coded baseline that exploits dataset biases (Sec 5.2). Note that existing models [2-8] do not explicitly model the goal location, and are thus not capable of predicting the goal location from a provided sequence of observations.