Cross-modal Map Learning for Vision and Language Navigation Supplementary Material

Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan,
Eleni Miltsakaki, Dan Roth, Kostas Daniilidis
University of Pennsylvania
{ggeorgak,karls,kwanchoo,sohamdan,elenimi,danroth,kostas}@seas.upenn.edu

Abstract

In this supplementary document we provide the following items:

1. *Discussion on societal impact and limitations.*
2. *Implementation details.*
3. *Analysis of path prediction learning with regards to auxiliary loss and start position input.*
4. *Analytical results over semantic map prediction to assess the contribution of cross-modal map prediction.*
5. *Additional results on the effect of stop decision threshold.*
6. *Additional qualitative navigation results and visualizations of the learned attention representations.*

1. Societal Impact and Limitations

Potential negative societal impact. Our current method is trained on scenes from Matterport3D which contains scans of homes from North America and Europe. Since we do not model out-of-distribution scenarios, deploying our method in safety critical situations such as rescue operations or hospitals could have negative outcomes. Furthermore, house layouts strongly correlate with regions of the world and with socio-economic factors, making it likely that agents using our algorithm will underperform when deployed in other parts of the world or in poor or minority houses which are frequently underrepresented in datasets.

Limitations. While our approach achieves results comparable with the state of the art, we acknowledge that there is much room for improvement.We would like to point out three limitations of our method. First, since we predict the path from the semantic map, we are not utilizing information from the instructions that describe object attributes such as color, (i.e., "brown table", "red table"). This can be important in situations where we need to distinguish between two instances of the same category. Second, we depend on the pretrained BERT representation, after fine-tuning its final layer, to provide all relevant information about the instruction. We do not use any explicit language representation, which could allow for better decomposition of instructions. Third, our method is limited by size of the local egocentric map. We cannot spatially ground information to locations outside of the local map, and while increasing the size of the local map can significantly improve performance, it is also computationally expensive.

	TL	NE	OS	SR	SPL
CM^2-GT, w/o $\mathcal{P}_t^0$, $\lambda_\xi = 0$	**9.37**	6.80	32.9	29.3	22.2
CM^2-GT, w/o $\mathcal{P}_t^0$	10.62	6.18	38.4	34.3	26.5
CM^2-GT, $\lambda_\xi = 0$	12.61	5.04	54.3	49.1	39.0
CM^2-GT	12.60	**4.81**	**58.3**	**52.8**	**41.8**

Table 1. Analysis of our path prediction strategy demonstrating the contributions of $\mathcal{P}_t^0$ and the auxiliary loss using navigation metrics on *val-seen* set.

2. Implementation details

Our method is implemented in PyTorch [1]. The UNet [2] models used in our method have four encoder and four decoder convolutional blocks with skip connections. The entire model is trained with the Adam optimizer and a learning rate of 0.0002. During training all λs are equal to 1. The training data for both the map and waypoint prediction were sampled from the ground-truth paths provided in VLN-CE train split. We used around 700K examples to train *CM^2* and around 500K to train *CM^2-GT*. The semantic segmentation that produces $\hat{\chi}$ is another UNet which we pre-trained separately from the rest of the model on RGB observations from the Matterport3D scenes. The egocentric map and waypoint heatmap dimensions are $h' = w' = 192$ and $u = v = 24$ respectively. Each pixel in the egocentric

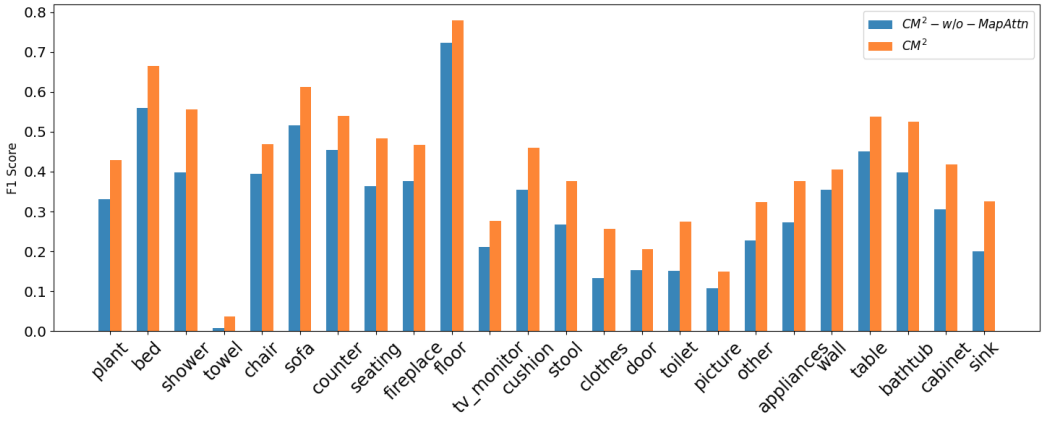

Figure 1. Per-class semantic map predictions with and without cross-modal map attention. Performance gains are more noticeable for object categories over floor and wall.

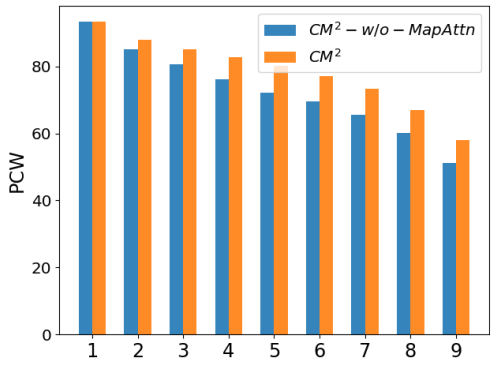

Figure 2. Per-waypoint path prediction results with and without cross-modal map attention. Waypoint 9 corresponds to the goal, while waypoint 0 is used as input to our method.

map corresponds to physical dimensions of $5cm \times 5cm$. We use $k = 10$ waypoints and $c = 27$ semantic classes from the original 40 categories of Matterport3D. For the controller we define stop distance threshold $\tau = 0.5$ and goal confidence threshold $\gamma = 0.6$. Our method does not use any recurrence or an implicit state representation so the map and path predictions are temporally independent. However, during a navigation episode we maintain a global occupancy map using the ground-projected depth o_t that is registered using Bayesian updates. The input to the model is an egocentric crop from this global map, so the agent is aware of previously observed occupancy.

3. Analysis of path prediction learning

We investigate the contribution of certain choices we made to mitigate the ambiguity over waypoint placements during path prediction learning as discussed in section 3.3 of the main paper. In particular, we train the following variants of our CM^2-GT model: 1) without using the starting position heatmap $\mathcal{P}_t^0$ as input, 2) without the auxiliary loss for predicting whether a waypoint has been traversed ($\lambda_\xi = 0$), and 3) without $\mathcal{P}_t^0$ and $\lambda_\xi = 0$. The variants are evaluated against our proposed approach on *val-seen* using the navigation metrics from section 4.1 of the main paper (Table 1). We observe that without the auxiliary loss success rate drops by 3.7%, while not using the starting position further decreases success rate by 18.5%. The worst performance by far is recorded when both are not utilized. The results justify our choices and suggest the importance of anchoring the prediction of the entire path to a starting location in the egocentric map, complemented by an auxiliary objective that forces the model to predict its current position on the path.

4. Analytical results for cross-modal map attention

In section 4.2 of the main paper we investigated the importance of the cross-modal map attention component by comparing our approach to the baseline CM^2-w/o-MapAttn that is unaware of the language instruction during map prediction. Here, we show additional per-class and per-waypoint results over F1 score (Figure 1) and PCW (Figure 2) respectively. First, in Figure 1 we observe that the model trained with the cross-modal map attention (CM^2) performs better on all semantic categories against the baseline. Furthermore, the performance gain is more pronounced over object categories (e.g., toilet 12.4%, sink 12.6%) as opposed to semantic classes referring to the structure of the scene (e.g., floor 5.6%, wall 5.1%). This reinforces our initial hypothesis that the attention component is able to pick semantic cues from the instruction and improve the map prediction. Additionally, in Figure 2 we demonstrate path prediction results over individual waypoints (1-

	Val-Seen					Val-Unseen				
	TL	NE	OS	SR	SPL	TL	NE	OS	SR	SPL
CM^2, $\tau = 1.5$	**9.54**	6.06	42.4	38.8	34.6	**9.07**	**7.01**	35.2	31.3	27.7
CM^2, $\tau = 1.0$	10.72	**5.88**	49.2	42.6	**35.9**	10.04	7.09	39.0	33.3	**27.9**
CM^2, $\tau = 0.5$	12.05	6.10	**50.7**	**42.9**	34.8	11.53	7.02	**41.5**	**34.3**	27.6

Table 2. Additional results on the effect of stop distance threshold on VLN.

9). Waypoint 0 is omitted since it is used as input to our method, while waypoint 9 corresponds to the goal location. As expected, waypoints earlier in the path have larger PCW. However, an interesting observation is that the gain in performance increases for waypoints closer to the goal rather than in the beginning of the path, thus demonstrating that improved map prediction is crucial for predicting waypoints far from the starting position.

For additional qualitative comparisons of semantic map predictions between the baseline and our approach see Figure 6.

5. Additional results on effect of stop distance threshold.

We repeat the experiment presented in section 4.2 of the main paper regarding the effect of the stop distance threshold on the VLN task using our CM^2 (no GT map) agent on both val-seen and val-unseen splits. In Table 2 we observe a similar trend as that shown in Table 4 of the main paper. Success rate is higher when τ is low, because the agent takes the stop action more cautiously, while trajectory length is best when τ is high.

6. Additional visualizations

Finally, we share additional visualizations of navigation episodes (Figure 5) and more examples of spatial and semantic grounding of the learned representations. Figure 3 shows the attention decoder output H_t^s and Figure 4 presents more examples of the cross-modal attention. See section 4.3 of the main paper for more details.

References

[1] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 1

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

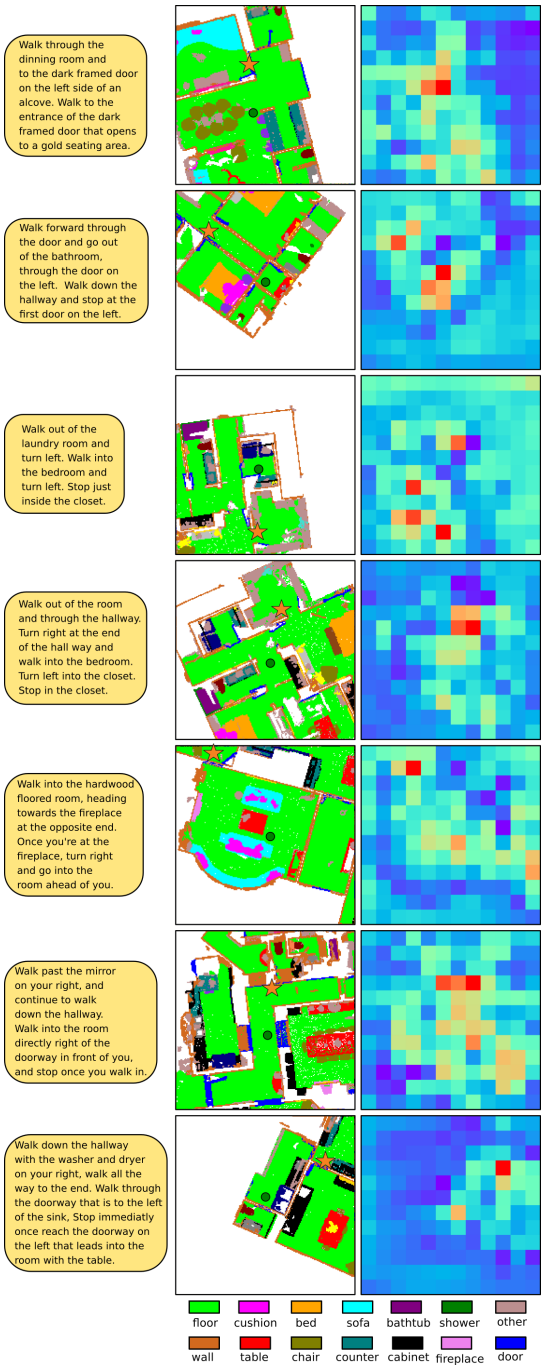

Figure 3. Visualization of attention decoder output H_t^s that focuses on areas around goal locations and along paths. The agent's location is denoted with a green circle and the goal with an orange star.

Figure 4. Visualization of the cross-modal attention representation between map and specific word tokens. The representation tends to focus on semantic areas of the map that correspond to the object referred to by the token. Note that in the example on the 4th row the representation focuses on the area where stairs are located, even though we do not use a specific semantic label for stairs in the map.

Figure 5. Navigation examples using our method *CM*^{2} on *val-seen* (first from top) and *val-unseen* (last three). The top row of each example shows the RGB observations of the agent, while bottom shows the path prediction on the egocentric maps (the agent is in the middle looking upwards shown as the green circle). The red waypoints represent our path prediction at the particular time-step. Observe that the goal, shown as an orange star, is neither visible nor within the egocentric map at the beginning of the episodes. The ground-truth map and path are depicted in the bottom left corner.

Figure 6. Semantic map predictions with and without cross-modal map attention.