A. Background: Transformers

The proposed spatial planning method is based on the Transformer model (Vaswani et al., 2017). A Transformer layer, denoted by f_{TL}, takes a tensor $X \in \mathcal{R}^{d\times S}$ as input, where d is the embedding size and S is the size of the input. It consists of two sublayers, a multi-head self-attention layer (f_{SA}) and a position-wise fully connected layer (f_{FC}). There is a residual connection around each sublayer, followed by layer normalization (Ba et al., 2016) (LN):

$$R = \text{LN}(f_{\text{SA}}(X) + X), Y = f_{\text{TL}}(X) = \text{LN}(f_{\text{FC}}(R) + R)$$

where $R, Y \in \mathcal{R}^{d\times S}$ are the intermediate and final representations, respectively.

The multi-head self-attention (f_{SA}) layer has h attention heads, each computes a scaled dot-product attention over queries Q, keys K and values V, which are all different projections of the input X:

$$Q_i = W_{Q,i}^T X, \quad K_i = W_{K,i}^T X, \quad V = W_{V,i}^T X$$

$$Z_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i$$

where $Q, K \in \mathcal{R}^{d_k\times S}$, $V \in \mathcal{R}^{d_v\times S}$, $i \in 1, 2, \ldots, h$ d_k and d_v are hyper-parameters and all Ws are parameters. The output of all attention heads, Z_is, are concatenated and projected to the same dimension as the input. Finally, the position-wise fully connected (f_{FC}) layer applies two linear transformations to each position with a ReLU activation to the output of the multi-head attention.

B. Dataset Details

We generate synthetic datasets for training the spatial planning models for both navigation and manipulation settings. For the navigation setting, we perform experiments with $M \times M$ maps with two different map sizes, $M \in \{15, 30\}$. We randomly generate $o_{min} = 0$ to $o_{max} = 5$ obstacles in each map, where each obstacle is an rectangle at a random location with each side being a random length from 1 to $M/2$. All the rectangular obstacles are rotated in two random orientations.

For the manipulation setting, we consider a reacher task using a planar arm with 2 degrees of freedom. We use an operational space of size $P \times P$. Each link of the arm is of size $P/4$. The arm is centered at the center of the operational space. Let the orientation of two links be denoted by θ_1 and θ_2. We assume both the links can freely rotate in a plane, $\theta_1, \theta_2 \in [0, 2\pi)$. For each environment, we generate $o_{min} = 0$ to $o_{max} = 5$circular obstacles centered at a random location $0.25P$ to $0.75P$ distance away from the center, with a random radius between $0.05P$ and $D - 0.15P$ where D is the distance of the center of the obstacle from the center of the operational space. We convert each environment to a configuration space map of size $M \times M$, where each cell (i, j) denotes whether the arm will collide with an obstacle when $\theta_1 = 2\pi i/M$ and $\theta_2 = 2\pi j/M$. We experiment with two map sizes, $M \in \{18, 36\}$, corresponding to 20° and 10° bins for each link. The choice of P does not affect the map as the collision check for each cell in the configuration space is performed in the continuous operational space where all distances are relative to P.

C. Navigation Mapper Architecture Details

The Navigation Mapper module predicts a single value between 0 and 1 for each image in o indicating whether the cell in the front of the image is an obstacle or not. The architecture of the Navigation mapper consists of ResNet18 convolutional layers followed by 3 fully-connected layers of size 256, 128, and 1 as shown in Figure 8. Each cell can have up to 4 predictions (from images corresponding to the four neighboring cells facing the current cell), which are aggregated using max-pooling to get a single prediction.

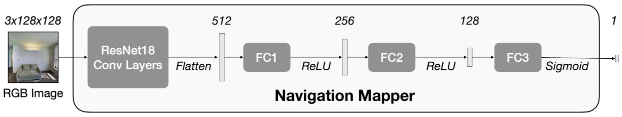

Figure 8. **Navigation Mapper Architecture.** Figure showing the architecture of the Navigation Mapper.

D. Attention Visualization

We show the visualization of attention maps corresponding to two different locations in Figure 9. Interestingly, we noticed three consistent patterns: a) at least one of the attention head out of eight captures obstacles (left), b) one of the attention heads focuses on goal location (middle), and c) some attention maps focus on nearby obstacles to get accurate planning distance (right).

E. Examples

We show additional examples for navigation task for in-distribution test set (in Figure 10), out-of-distribution More Obstacles test set (in Figure 11) and Real-World test set (in Figure 12) each with map size $M = 30$. Additional examples for manipulation task are shown for in-distribution test set (in Figure 13) and for out-of-distribution More Obstacles test set (in Figure 14).

We also visualize examples for the end-to-end mapping and planning experiments for the manipulation task. We show examples of map and action distance predictions using the SPT model trained with dense and perfect supervision in Figure 15 and with noisy and sparse supervision in Figure 16.

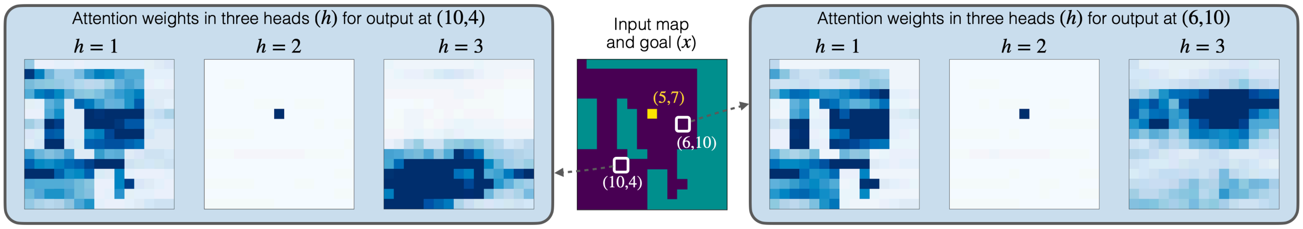
Attention weights in three heads (h) for output at $(10,4)$
$h = 1$
$h = 2$
$h = 3$
Input map
and goal (x)
(5,7)
(6,10)
(10,4)
Attention weights in three heads (h) for output at $(6,10)$
$h = 1$
$h = 2$
$h = 3$

Figure 9. **Attention Visualization.** Visualization of the attention heads learned by Spatial Planning Transformers. SPTs learn an attention for each location in the map with respect to every other location.

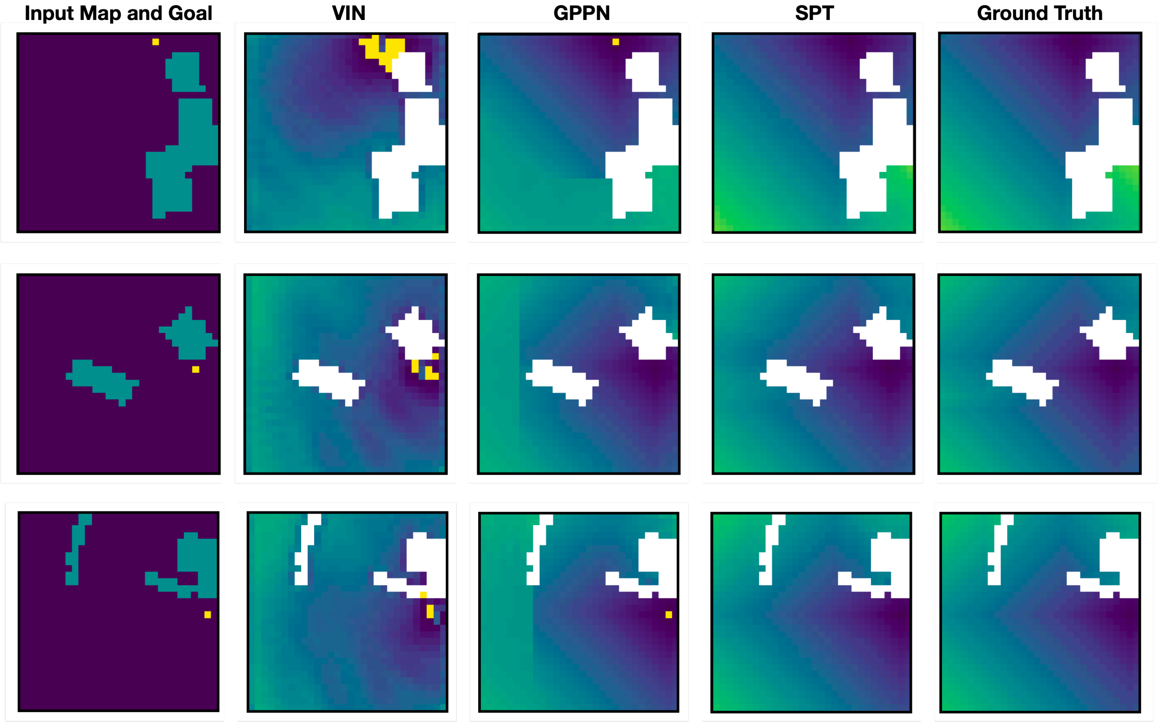

Figure 10. **Navigation in-distribution test set examples.** Figure showing 3 examples of the input, the predictions using the proposed SPT model and the baselines, and the ground truth for the Navigation in-distribution test set for map size $M = 30$.

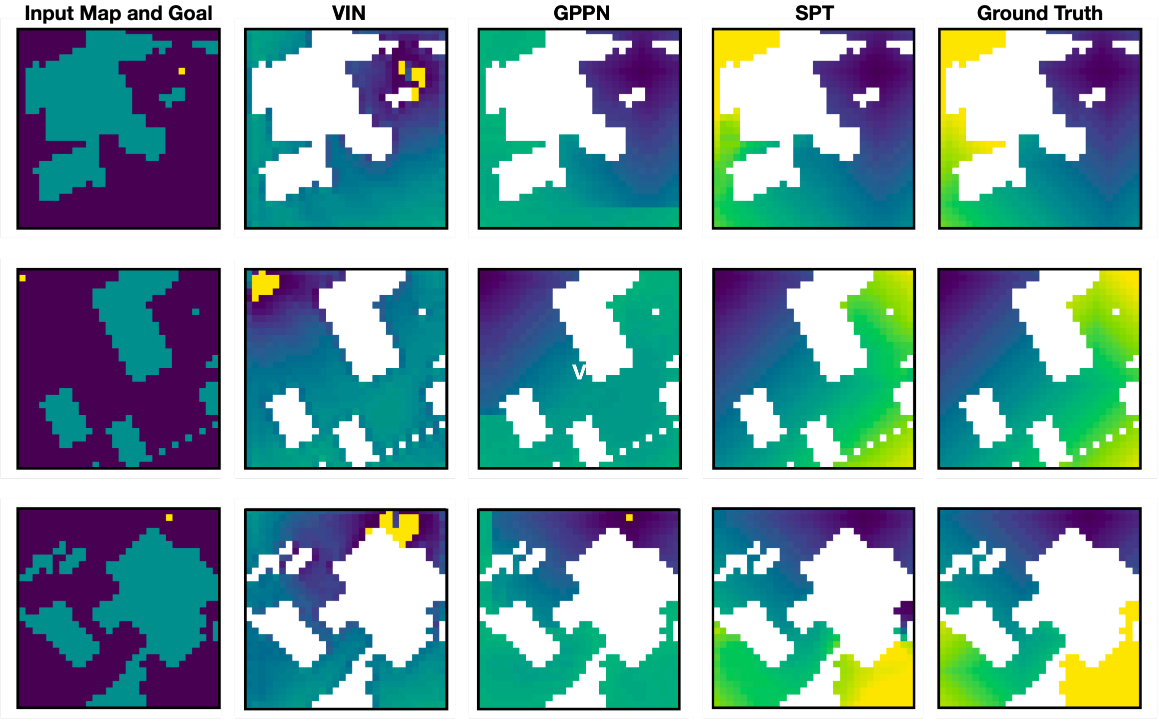

Figure 11. **Navigation out-of-distribution More Obstacles test set examples.** Figure showing 3 examples of the input, the predictions using the proposed SPT model and the baselines, and the ground truth for the Navigation out-of-distribution More Obstacles test set for map size $M = 30$.

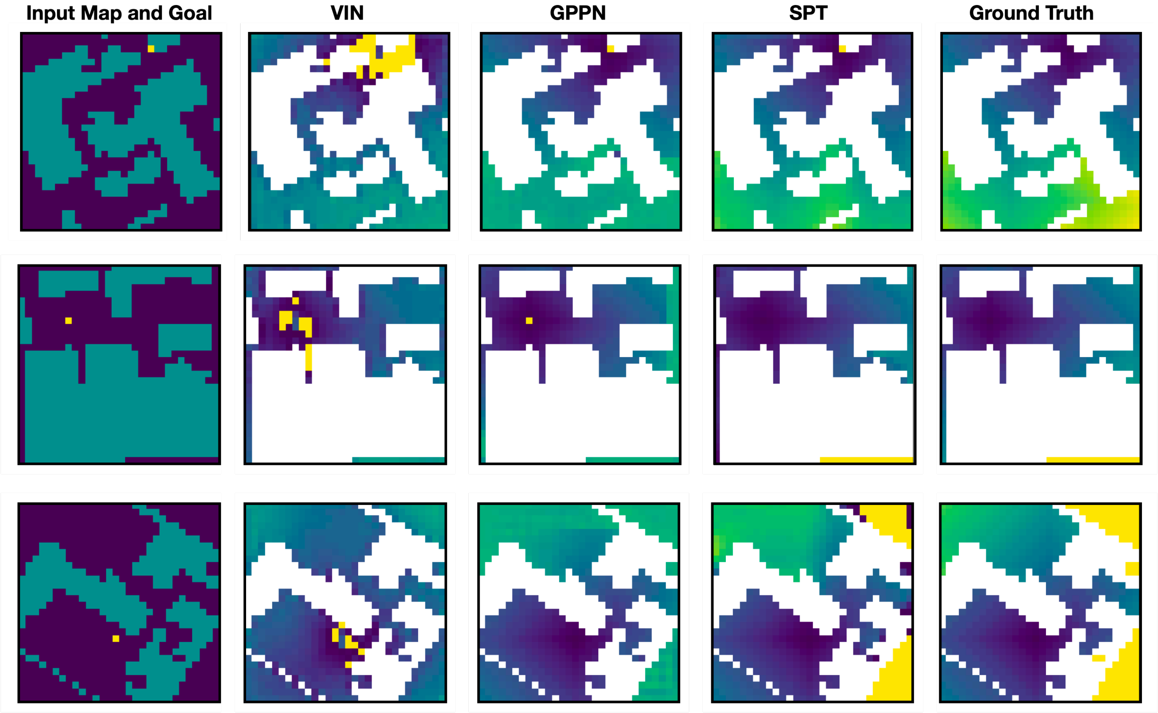

Figure 12. **Navigation out-of-distribution Real-World test set examples.** Figure showing 3 examples of the input, the predictions using the proposed SPT model and the baselines, and the ground truth for the Navigation out-of-distribution Real-World test set for map size $M = 30$.

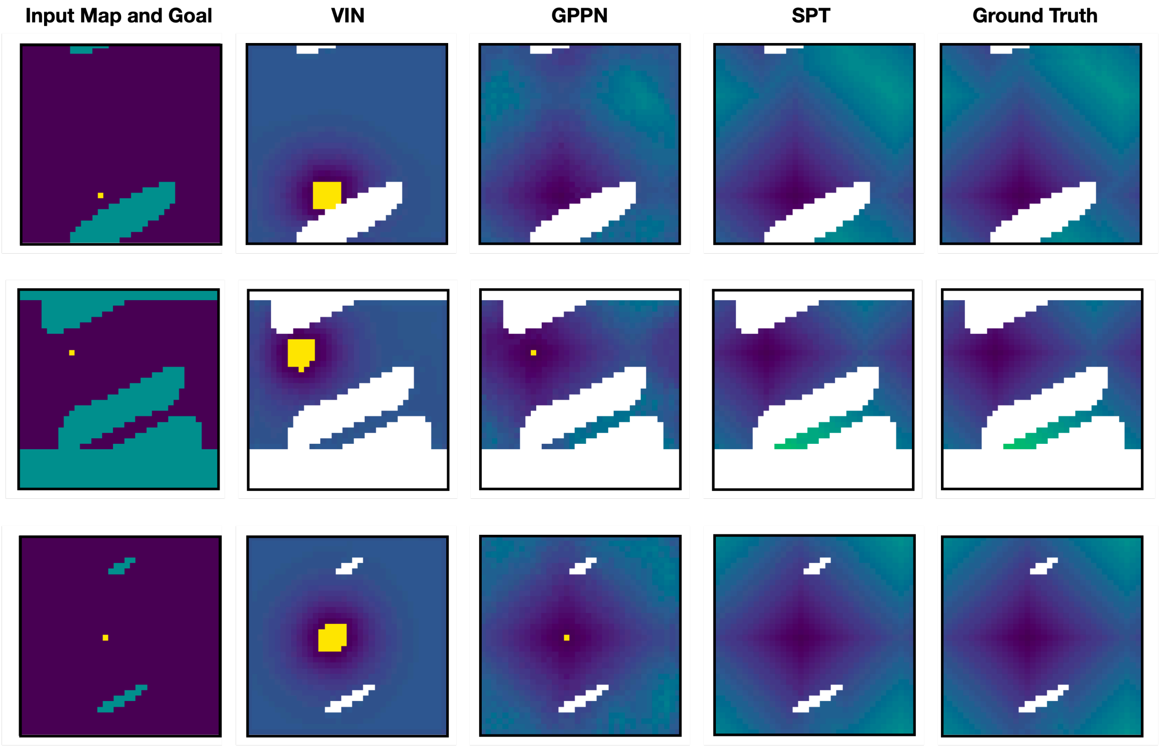

Figure 13. **Manipulation in-distribution test set examples.** Figure showing 3 examples of the input, the predictions using the proposed SPT model and the baselines, and the ground truth for the Manipulation in-distribution test set for map size $M = 36$.

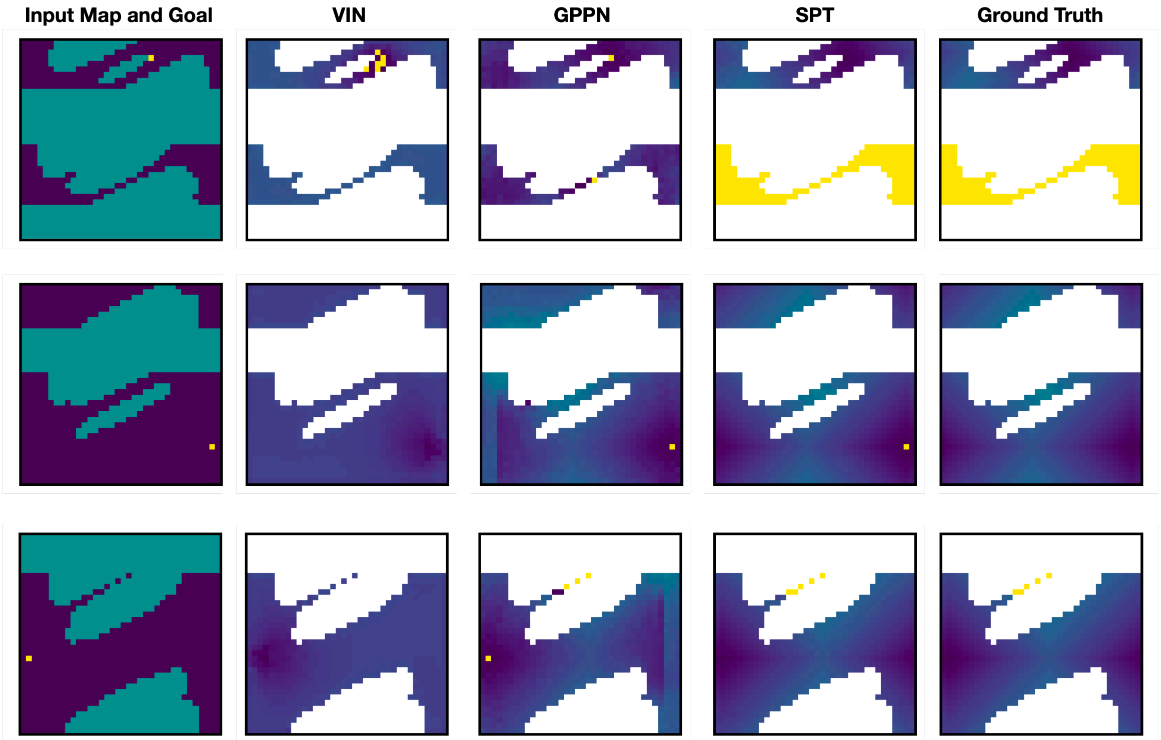

Figure 14. **Manipulation out-of-distribution More Obstacles test set examples.** Figure showing 3 examples of the input, the predictions using the proposed SPT model and the baselines, and the ground truth for the Manipulation out-of-distribution More Obstacles test set for map size $M = 36$.

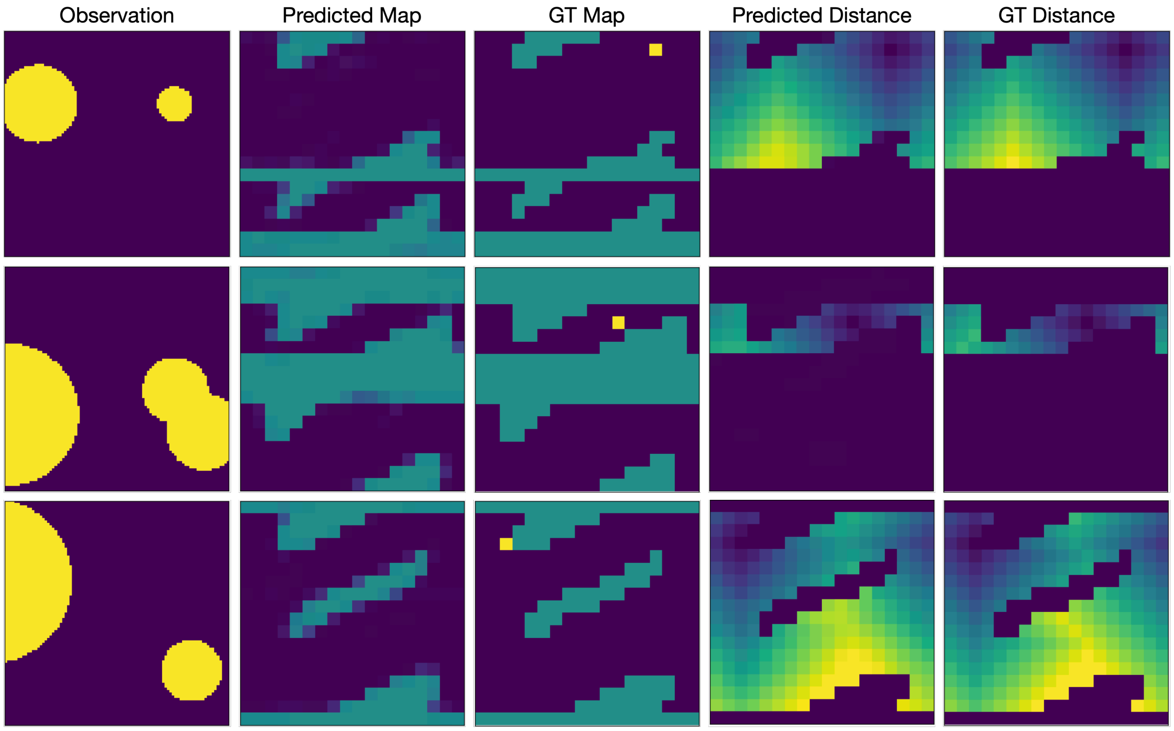

Figure 15. **Dense and Perfect Supervision.** Figure showing examples of map and distance predictions using the SPT model trained with dense and perfect action-level supervision.

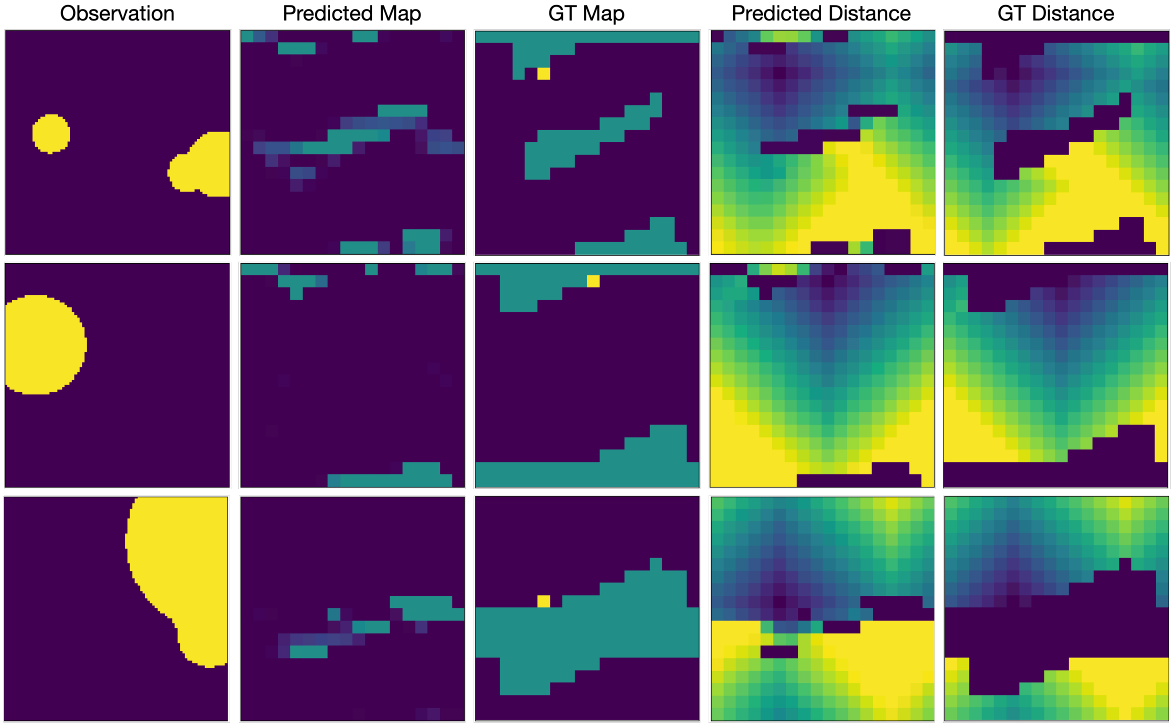

Figure 16. **Sparse and Noisy Supervision.** Figure showing examples of map and distance predictions using the SPT model trained with sparse and noisy action-level supervision.