A Input Past Trajectory Experiment: Additional Experimental Details

Data. We use the same data and input representation as Phan-Minh et al. [38], but we filter out any data with less than $1\,\mathrm{s}$ of past trajectory information to enable decoding of the agent's past trajectory. The input past trajectory OOD split with a threshold of $10\,\mathrm{m}$ for the heuristic distance allows for sufficient ID training (25,669), validation (7,344), and test (7,270) examples, while still having a reasonable number of OOD examples (validation: 2,521, test: 3,267).

Architecture and Training Details. For the CoverNet, ensemble, and Post-CoverNet baseline models we use a ResNet-50 backbone to extract features, following the procedure used by Phan-Minh et al. [38]. For the ISAP model, to compensate for the added compute associated with the interpretable architecture, we use a ResNet-18 backbone to extract features. The backbone features are fed into two linear layers for the baseline models, whereas in ISAP there are three blocks of linear layers, one block for each semantic concept. We found the coefficients: $\lambda_{\text{agent}} = 1$, $\lambda_{\text{map}} = 1$, and $\lambda_{\text{sc}} = 10$ for the loss to work well in training. The coefficient for the social context decoding is higher than the rest as this representation is spatially sparse, and otherwise the decoding collapses to a stable local minimum of predicting no other agents in the scene. We train the model for 25 epochs using the Adam [59] optimizer with a 0.001 learning rate, a batch size of 16, and a weight decay of 5×10^{-4}. We note that we train the ISAP model for 25 epochs with no early stopping as the different loss components have varying convergence speeds. All baselines are trained according to the same training set-up as ISAP, but we save the best model according to the validation loss.

In Post-CoverNet, we learn a radial normalizing flow [51] of eight layers for each of the 64 anchors. We place a batch normalizing layer before the normalizing flows, per the advice by Charpentier et al. [34]. The normalizing flows learn a density over a four-dimensional latent space. For ISAP, we learn a set of 64 normalizing flows for each of the semantic concepts, for a total of $64 \times 3 = 192$ normalizing flows. Setting the total certainty budget to $\sum_c N_c = e^6$ worked well empirically.

Following the procedure outlined by Phan-Minh et al. [38], the CoverNet and ensemble models use a modified cross-entropy loss, called the constant lattice loss, for the classification task. The ground truth label is the anchor with the trajectory in the anchor set closest to the true future trajectory according to the minimum average point-wise Euclidean distance. For Post-CoverNet, we use the ELBO loss defined in Eq. (3). This loss corresponds to a Bayesian loss with an uninformative Dirichlet prior [34]. For ISAP, the reconstruction losses from the decoders are added to the ELBO loss. We found scaling the KL divergence term by 10^{-5} to work well empirically.

All reconstruction losses are the sum of squared errors. Since the agent's past behavior information is low-dimensional compared to the size of the input x, we make a design decision to decode a single vector for this latent variable. The agent decoder output includes the trajectory of the agent of interest for the past $2\,\mathrm{s}$ and the agent's speed, acceleration, and heading change rate. The decoder consists of two linear layers. For the map and social context latent variables, we decode them into the respective subcomponents of the spatial representation in the input x (see Fig. 1). Each pixel in the spatial representation is predicted to be in $[0, 1]$ along three RGB channels. Instead of decoding from the latent encoding z, which is four-dimensional, we decode the map and social context from an upstream feature layer of dimension $4,096$ to increase the representational capacity of the latent space. These decoders consist of convolutional components inspired by the VQ-VAE model [60].

Runtime. The considered models run on average at: $4.6\,\mathrm{Hz}$, $0.920\,\mathrm{Hz}$, $0.460\,\mathrm{Hz}$, $1.789\,\mathrm{Hz}$, and $0.797\,\mathrm{Hz}$ for CoverNet, the small ensemble ($N = 5$), the big ensemble ($N = 10$), Post-CoverNet, and ISAP, respectively. Our ISAP model is thus more efficient than the larger ensemble while achieving better uncertainty estimation performance. The Post-CoverNet model provides a one-shot epistemic uncertainty estimation approach that is more efficient than both ensembles.

B Map-Based Experiment: Additional Experimental Details

Data. To further test our approach, we conduct a map-based experiment. We sub-sample the NuScenes [40] dataset based on HD map information. Starting from the data used by Phan-Minh et al. [38], we again filter out any data with less than $1\,\mathrm{s}$ of past agent trajectory information to enable decoding of the agent's past trajectory. We then split the data into ID and OOD examples according to the metadata associated with the HD map provided by NuScenes [40]. ID examples are chosen

Table 3: Trajectory prediction results for the CoverNet [38] baseline on ID (OOD) test set data for both the input past trajectory and map-based OOD data splits. Lower is better. We see a substantial drop in performance from ID to OOD data for both experiments, hence OOD detection in this setting is important.

Experiment	minADE_1	FDE
Input Past Trajectory	4.327 (7.130)	9.474 (13.632)
Map-Based	4.732 (6.111)	10.590 (13.464)

to be from Singapore's Holland Village and Queenstown neighborhoods (left-side driving) and to not contain 'roundabout' or 'big street' in the description. OOD data is taken from Boston (right-hand driving) and contains 'roundabout' in the description. We note that although 'roundabout' may be in the metadata, this refers to the scene, and not necessarily the current local map surrounding the agent of interest. Thus, although the majority of examples contain roundabouts, we have some straight roads without roundabouts in the OOD data as well. Similarly, despite filtering out 'big street' scenes from ID data, there may still be some larger roads in the ID dataset. This split allows for sufficient training (8,110), validation (318), and test (2,186) examples for ID, while still having a reasonable number of OOD examples (validation: 80, test: 364).

Training Details. We largely follow the architecture and training details described in Appendix A. We found a coefficient of one to work well for all the reconstruction losses. In this experiment, the reconstruction losses took longer to converge, thus we train the ISAP model for 50 epochs and save the model with the best validation performance on $\mathcal{L}_{\text{ELBO}}$.

C OOD Split Verification

To support the validity of our choice of OOD data splits (input past trajectory and map-based), we evaluate the CoverNet [38] baseline on the ID and OOD test sets using trajectory prediction metrics in Table 3. There is a significant drop in CoverNet performance for both the input past trajectory and map-based experiments when going from ID to OOD data. Thus, detecting these OOD examples would be important for safety critical applications.

D Entropy Visualization Results

In addition to the analysis provided in Section 5, we include visualizations of entropy histograms for both the input past trajectory and map-based experiments on ID and OOD test data in Fig. 4. We compare our ISAP approach to the larger ensemble ($N = 10$). To compute the entropy, we use the output categorical distribution for the ensemble and the categorical and Dirichlet distributions, capturing the aleatoric and epistemic uncertainty, respectively, for ISAP. In both experiments, ISAP provides a more clear distinction between ID and OOD data (individual peaks in the histograms) in terms of entropy than the ensemble, supporting our findings in Section 5. The ISAP entropy peaks are sharper for OOD data and higher in entropy value than those produced by the ensemble.

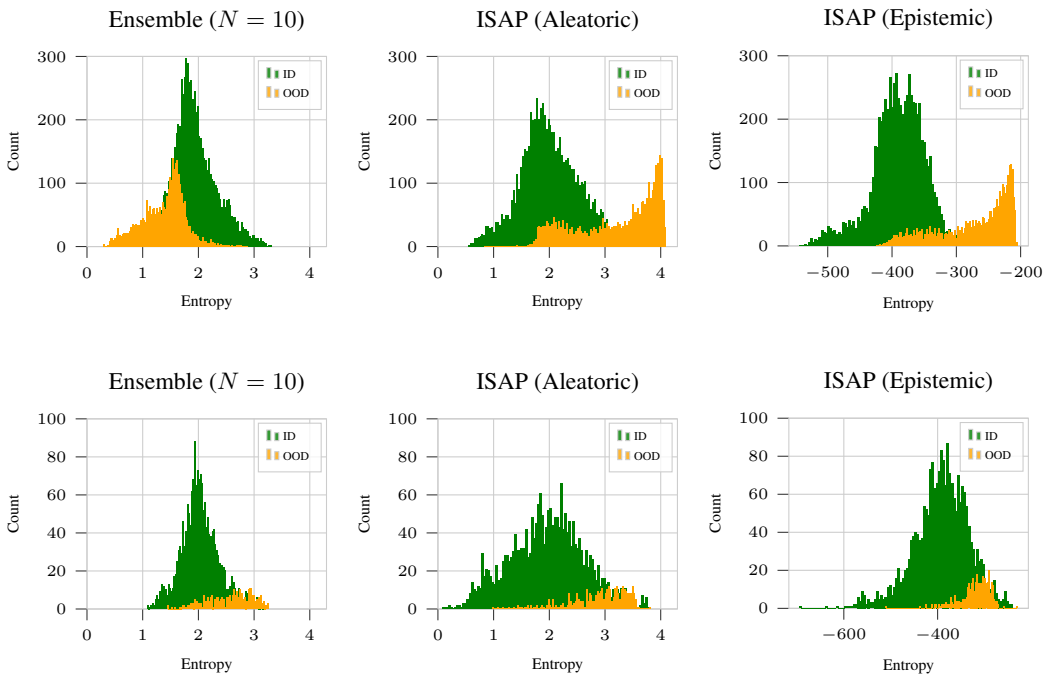

Figure 4: Entropy histograms for ISAP (ours) and the ensemble ($N = 10$). The first row shows the results for the input past trajectory experiment, while the second row shows those for the map-based experiment. All data is from the ID and OOD test sets. ISAP provides the clearest distinction (individual peaks) between ID and OOD data in terms of entropy.