HYPER: Learned Hybrid Trajectory Prediction via Factored Inference and Adaptive Sampling

Xin Huang1, Guy Rosman2, Igor Gilitschenski2, Ashkan Jasour1, Stephen G. McGill2, John J. Leonard1,2, Brian C. Williams1

Abstract—Modeling multi-modal high-level intent is important for ensuring diversity in trajectory prediction. Existing approaches explore the discrete nature of human intent before predicting continuous trajectories, to improve accuracy and support explainability. However, these approaches often assume the intent to remain fixed over the prediction horizon, which is problematic in practice, especially over longer horizons. To overcome this limitation, we introduce HYPER, a general and expressive hybrid prediction framework that models evolving human intent. By modeling traffic agents as a hybrid discrete-continuous system, our approach is capable of predicting discrete intent changes over time. We learn the probabilistic hybrid model via a maximum likelihood estimation problem and leverage neural proposal distributions to sample adaptively from the exponentially growing discrete space. The overall approach affords a better trade-off between accuracy and coverage. We train and validate our model on the Argoverse dataset, and demonstrate its effectiveness through comprehensive ablation studies and comparisons with state-of-the-art models.

I. INTRODUCTION

Predicting future trajectories of traffic agents is a key task for autonomous vehicles. This task is challenging due to multi-modal human intent. There is an inherent trade-off between accurately representing the distribution of trajectories and covering the diversity of potential intents [1], [2], [3].

Several recent works address the trade-off explicitly using a multi-stage approach [4], [5], [6], [7], [8]. First, they model high-level human intent as a discrete state and infer such intent to provide sufficient coverage over multi-modal human intent. Next, trajectories are generated conditioned on the intent. The models are trained to maximize the data likelihood to support prediction accuracy. They demonstrate great success in terms of prediction accuracy and coverage, and provide explainability in predicted trajectories. However, the existing approaches often use a simplified intent model that assumes the intent is fixed over time, to keep the prediction space reasonable. In practice, a traffic agent can change its intent (i.e. follow the lane, perform a lane change, and turn), especially over long horizons.

When accounting for the evolving discrete intent, the number of discrete modes grows exponentially in the prediction horizon [9]. This is studied in the context of factored inference, e.g. by merging and pruning mode hypotheses [10], [11] or sampling from the prediction space [12], [13]. In the domain of trajectory prediction, the exponentially growing discrete space can be mitigated by expanding the discrete predictions at a few selected points [14] or accounting for the most probable intent [15]; however, they may not provide sufficient accuracy and coverage in a multi-modal problem.

We propose an approach that better captures both accuracy and coverage by explicitly modeling discrete intent sequences. Our approach, HYbrid trajectory PrEdiction RoR (HYPER), uses a learned probabilistic hybrid automata model, as illustrated in Fig. 1(a), to jointly infer a sequence of high-level discrete modes when generating low-level trajectory predictions. We use neural proposal distributions [16] in the hybrid model and the farthest point sampling algorithm to obtain good coverage of trajectories with only a few samples, while preserving model accuracy. Our contributions are as follows: i) We formulate trajectory prediction as a general and expressive hybrid prediction problem allowing an evolving discrete intent, and learn a probabilistic hybrid automaton model as a deep neural network. ii) We leverage a learned proposal function to sample adaptively from an exponentially growing discrete space in the hybrid model to support both accuracy and coverage, and utilize a sample selection technique to further improve prediction performance given limited samples. iii) We train and validate our model on a naturalistic driving dataset and perform detailed experiments to test our hypothesis and demonstrate the effectiveness of our approach.

II. RELATED WORK

Multi-Modal Trajectory Prediction Trajectory prediction has been studied extensively in the past few years. To account for uncertainty and multi-modality in prediction space,
generative adversarial networks (GAN) [17] and variational autoencoders (VAE) [18], [19] are used to generate multiple trajectory predictions by sampling a latent space. Several works have attempted to improve coverage of the possible outcomes [1], [3], [20], [21], yet there is an inherent trade-off between accurately representing the trajectory distribution and covering a diverse set of intents [1], [2], [3]. To account for this trade-off between accuracy and coverage explicitly, hybrid models are proposed to classify discrete intent and generate continuous trajectories conditioned on the intent. The intent is defined by a variety of choices, including driving maneuvers [4], [22], [23], goal locations or waypoints [6], [8], [24], [25], and target lanes [7], [26], [27], etc. In these hybrid approaches, the intent is assumed to be fixed over time. In practice, however, the agent may change its intent over time, especially over a long horizon, or follow different intents to get to the same target location or lane. When accounting for evolving discrete intent, [14] leverages a support vector machine to infer discrete intent over specific decision points, and [15] proposes a discrete choice modeling approach to infer discrete anchors over time. Such models either expand discrete predictions at a few selected steps or predict the most-probable intent, to avoid dealing with the exponentially growing discrete space. In this work, we propose a general and expressive hybrid prediction framework that accounts for evolving intent by inferring a sequence of discrete modes over time, and predicting trajectories consistent with the mode sequence.

Factored Inference The discrete prediction space suffers from exponential growth as a function of the prediction horizon. This problem has been addressed in the context of factored inference, by approximating the intractable state space through pruning and sampling techniques. For instance, multiple model estimation algorithms estimate the possible operational modes for a system, and filter states from an exponential number of hypotheses, by merging and pruning hypotheses [10], [11]. Furthermore, [13] models the hybrid system through a hybrid Bayesian network, and proposes a sampling-based approximation algorithm to track hybrid states. In parallel, [28], [12], [29], [30] model hybrid systems through a probabilistic hybrid automaton (PHA) [31] or a switching linear dynamical systems (SLDS) [32], and apply efficient pruning, search, and sampling methods to maintain reasonable estimation performance. Existing factored inference methods often assume a linear system to obtain closed-form or tractable solutions, and do not fully utilize the relevant context information. In this work, we model a hybrid system as deep neural networks, which excel at modeling nonlinear agent dynamics and learning complex environment contexts such as map information.

Trajectory Sampling Many trajectory prediction methods [33], [17], [19] sample multiple predictions in parallel from a learned distribution. Our approach utilizes a sequential adaptive sampling technique to generate samples through a learned proposal distribution, conditioned on previously generated samples. This allows the model to provide more coverage with limited samples. The learned adaptive proposal function is inspired by a few ideas from sequential trajectory prediction sampling [34], sequential Monte Carlo [35], [16], [36], and adaptive sampling [37], [38], [39]. Compared to existing sequential Monte Carlo methods, our proposal function is used to sample per time step, as opposed to sampling a full trajectory at once, to avoid a large proposal state.

In many applications such as autonomous driving, only a small set of prediction samples can be afforded, as evaluating each sample for downstream tasks such as risk assessment is expensive [40]. To select a limited number of candidates from all predicted samples, [3], [21] leverage diverse sampling techniques to choose semantically meaningful samples from a latent space; [6] uses non-maximum suppression (NMS) to prune trajectories that are close to each other to improve coverage; [8] proposes task-specific sub-sampling techniques towards optimizing the evaluation metrics. Similar to [6], our approach selects samples directly over the predicted trajectories, offering better interpretability.

III. PROBLEM FORMULATION

In this section, we introduce the hybrid system model used for HYPER, followed by a formal problem statement.

A. Hybrid System Modeling

We model a traffic agent as a probabilistic hybrid automaton (PHA) [31]. Compared to a hidden Markov model [41], the transitions in PHA have an autonomous property [12], i.e. the discrete mode evolution depends on the continuous state. This property provides better model capacity in mimicking the behavior of traffic agents. The PHA is a tuple $H = \langle s, w, F, T, s_0, Z \rangle^1$, where $s = x \cup z$ denotes the hybrid state variables; $z$ denotes the discrete mode with a finite domain $Z$, and $x \in \mathbb{R}^n_x$ denotes continuous state variables; $w$ specifies the input/output variables, which consists of context variables $c$, continuous observation variables $o_c$, and discrete observation variables $o_d$; $F : Z \rightarrow F$ specifies the continuous evolution of the automaton for each discrete mode, in terms of a set of discrete-time difference equations $F$ over the variables $x$ and $c$; $T : Z \rightarrow T$ specifies the discrete evolution of the automaton for each discrete mode, as a finite set of transition probabilities drawn from $T$; $s_0$ denotes the initial hybrid state. The dependencies of hybrid state variables in a PHA are depicted as a graphical model in Fig. 1(a), where we omit the observation variables for simplicity. The state evolution is governed by the transition function $T$ and the dynamics function $F$, indicated by the three arrows going to $z_t$ and to $x_t$, respectively.

Similar to existing trajectory prediction approaches, we assume that the discrete labels are observable at training time, defined as maneuvers [4], [22], [23] or goal locations [6], [8]. These labels can be obtained through auto-labelling or unsupervised clustering over continuous trajectories. While it is challenging to perfectly label driver intent, we show in the experiments that our model is robust to imperfect labels, and defer learning with hidden intent for future work.

1We use the lowercase bold symbols to denote both the set of variables and the vector, and lowercase standard symbols for variable instantiations.
are normalized to an agent-centric frame with respect to the last observed position of the target agent, as in [42]. The observed path at each step is encoded through a multi-layer perceptron (MLP). The map is a static feature that is encoded through a model based on [42], which takes the map input as lane centerlines within 80 meters of the last observed agent position, and performs self-attention to pool the encoded states from all centerlines. Next, we run the encoded context state \( c_t \) through an LSTM network, which is commonly used in handling sequential data in trajectory prediction [34], [4], [3], to get the hidden context state vector \( h_t \) at the most recent observed step \( t = 0 \).

### B. Decoder

In the decoder, we generate a sequence of hybrid states, including discrete modes and continuous positions, using another LSTM network. At each time step, the LSTM unit takes as inputs the hidden state \( h_{t-1} \) and the hybrid state sample \( s_{t-1} \) from the previous step, and outputs a new hidden state \( h_t \) and an output state \( y_t \). The output LSTM state \( y_t \) is passed through a transition function \( T^s \), modeled as an MLP layer that produces a categorical distribution \( P_{T^s}(z_t) \) over discrete modes at time \( t \) as logits.

Existing factored inference algorithms [13], [12] sample from \( T \) to obtain discrete samples; however, naively sampling from \( T \) may take a large number of samples to sufficiently cover the prediction space. Therefore, we propose learning an additional \textit{proposal function} \( Q \) on the top of the transition function, to sample the discrete mode for the \textit{task} of achieving better prediction accuracy and coverage. The proposal function takes input from i) the distribution from the transition function \( T \), ii) the output LSTM state \( y_t \), and iii) the set of previously generated trajectory sequence samples \( S_t^{(P)} \), and outputs a categorical distribution over the next discrete modes as logits. The last input allows us to sample adaptively conditioned on previously generated samples. Each previous full trajectory sample is encoded through an MLP layer. The max pooling of all sample encodings is passed through an MLP layer before being fed into the proposal function. Compared to existing sequential Monte Carlo methods [35], [16], [36], our proposal function is used to sample per time step, as opposed to sampling a full trajectory at once, to avoid a large proposal state.

Given the output of \( Q \), we leverage a Gumbel-softmax sampler to sample the next mode \( z_t' \), whose real probability is obtained from \( T \). The mode is concatenated with \( y_t \) and fed into a dynamics function \( F \) as an MLP layer that outputs the distribution of the continuous state \( P_F(z_t|y_t) \). The distribution is parameterized as a Gaussian distribution with mean \( \mu_{z_t} \) and unit variance, which is chosen arbitrarily for stable training and is assumed in models such as [6]. We then sample a continuous state \( x_t' \) from the distribution. The predicted hybrid sample \( s_t' = (z_t', x_t') \) at time step \( t \) and the

\[ 2 \text{While the continuous state can be directly observed from perception systems, the discrete observation can be estimated from continuous observations, as discussed in Sec. III-A.} \]

\[ 3 \text{Code at github.com/ToyotaResearchInstitute/tristan.} \]
hidden state $h_t$ are used to generate the sample in the next step, until the prediction horizon is reached.

**Transition Function vs. Proposal Function:** Although both $T$ and $Q$ output a categorical distribution over modes, they serve different purposes. The transition function $T$ is part of the hybrid model defined in Sec. III-A, and it is used to compute the real probability of a sample. For instance, given an observation of a future hybrid state sequence, we can compute its likelihood by plugging it into the LSTM model (i.e. through $T$ and $F$). The likelihood is the summation of the discrete log-likelihood from $P_T$ and continuous log-likelihood from $P_F$, as in Eq. (1). It allows us to faithfully optimize the hybrid model, by maximizing the log-likelihood given the ground truth future observations, as defined in Sec. III-B.

On the other hand, the proposal function $Q$ determines which samples to generate, in order to improve prediction coverage in an exponentially growing space. It does not represent the true sample weight, which is determined by the transition function $T$. The weighted sample set allows us to cover the prediction distribution efficiently with only a few samples, which is advantageous to existing sampling-based methods that require a large number of samples to approximate the probability distribution. In Sec. V, we show a use case of sample weights to quantify prediction accuracy as negative log-likelihood.

**C. Learned Proposal Distributions**

To train the proposal function for accuracy and coverage, we generate $K$ trajectory sequences $\{s^{(k)} = (z^{(k)}, x^{(k)})\}_{k=1}^K$ sequentially from the decoder, and compute the min-of-$K$ L2 loss $\mathcal{L}_Q = \min_{k \in K} ||x^{(k)} - o_s||_2^2$.

There exist a few other options to learn the proposal function for coverage, such as maximizing entropy [44]. In this paper, we focus on the task of improving the diversity of the continuous trajectories when guaranteeing prediction accuracy, and choose the min-of-$K$ L2 loss (or variety loss [17]) that is widely used in the multi-modal trajectory prediction literature. While it is possible to train the model with only the min-of-$K$ L2 loss to favor towards prediction coverage, as in [17], [3], it leads to a diluted probability density function compared to the ground truth [2]. Therefore, we choose to improve prediction coverage while ensuring accuracy, by introducing the data likelihood loss in Eq. (1).

As a result, we can leverage the proposal distribution to generate representative samples, while obtaining the real probability of these samples from the transition function. To encourage the proposal distribution to be close to the transition distribution and the smoothness of the trajectory, we add a regularization loss $\mathcal{L}_{\text{reg}}$ penalizing i) the L2 difference between the two distribution logits and ii) the L2 loss over the second derivative of the predicted trajectory.

**D. Trajectory Sample Selection**

In many autonomous vehicle applications, we can only afford a small set of prediction samples, due to the non-trivial computational complexity of evaluating these samples for risk assessment [40]. To further improve coverage and boost prediction performance with a limited budget on samples, we propose to use the farthest point sampling (FPS) algorithm [45]. The algorithm selects trajectories that are far away from each other from samples generated from the proposal distribution, while maintaining their probabilities through the learned hybrid model. The algorithm works by selecting the next sample farthest away from the previously selected samples, in terms of the distance between end locations, with the first sample selected with the highest likelihood. FPS is able to capture the majority of distinct options thanks to its 2-optimal coverage property [45], as we show in Sec. V-B.2 on how it captures diverse samples with the proposal distribution.

**E. Model Training and Inference**

In training time, we train the entire model end-to-end,

$$\mathcal{L} = -\mathcal{L}_{\text{MLE}} + \alpha \mathcal{L}_Q + \beta \mathcal{L}_{\text{reg}},$$

where the MLE term (c.f. Eq. (1)) is negated as a loss to minimize, and $\alpha$ and $\beta$ are the loss coefficients.

At inference time, we i) sequentially call the hybrid model $M$ times with the proposal function to generate $M$ hybrid trajectory sequences, ii) compute their likelihoods based on the probabilities from the transition function and the dynamics function, and iii) perform FPS to select the final $N$ trajectory samples, and normalize the probabilities of each sample so that they sum up to 1.

**V. EXPERIMENTAL RESULTS**

In this section, we introduce the dataset and the model details, followed by a series of experiments demonstrating the effectiveness of our approach compared to baselines.

**A. Dataset and Model Details**

We train and validate HYPER on Argoverse v1.1 [46], a widely used benchmark for single agent trajectory prediction. The data contains 324,557 segments of agent trajectories, including two seconds of observed trajectories and three seconds of trajectories to predict, sampled at $10Hz$, as well as map information such as lane centerlines.

We augment the dataset offline with discrete mode labels over time, defined as stop, fast forward, slow forward, left turn, right turn, depending on the velocity and angular changes differentiated from the trajectories. Despite having a short prediction horizon, almost 40% of Argoverse data exhibits evolving intents with more than one label in three seconds, as we demonstrate in the experiments.

In the encoder, DynamicsNet is an MLP with 32 neurons; MapNet utilizes a similar structure as VectorNet [42]; the encoder LSTM has one layer with a hidden size of 32 and an output dimension of 32. In the decoder, the transition function and the proposal function use a two-layer MLP with (32, 5) neurons followed by a softmax layer; the dynamics function is a two-layer MLP with (32, 2) neurons; the sampler is a Gumbel-Softmax sampler [47] that produces
differentiable samples; the decoder LSTM has the same structure as the encoder LSTM. All MLPs are followed by ReLU and dropout layers with a rate of 0.1.

The loss coefficients $\alpha$ and $\beta$ in Eq. (2) are 1. The sizes of samples $K, M, N$ are 6, 50, and 6, respectively. The model is optimized using Adam [48] and trained on a single NVIDIA Tesla V100 GPU, with a batch size of 16 and a learning rate of 1e-3. It has 164K parameters and takes approximately 17 milliseconds to generate 50 predicted trajectories and 2 milliseconds to select 6 samples using FPS, making our model suitable in real-time systems.

We adopt the metrics used by the Argoverse benchmark to evaluate the prediction performance, including minimum average displacement error (minADE), minimum final displacement error (minFDE), a variant of minFDE that adds a probability penalization term (brier-minFDE), and miss rate (MR). All statistics are collected in the Argoverse validation dataset. For sampling-based methods (i.e. Gumbel-softmax), we run them five times and take the average. We annotate the method used in our model with *italics* font in the tables.

### B. Ablation Study

1) **Learned Adaptive Proposal Distribution:** We demonstrate the contribution of the learned proposal distribution by comparing different options of discrete intent proposal functions, including i) the learned transition function $T$ (i.e. setting $\alpha$ and $\beta$ to 0), ii) the non-adaptive proposal function that with no access to the other samples, and iii) our proposed proposal function $Q$ that samples adaptively by considering previously generated samples. The results are summarized in Table I, in which we obtain 6 samples without further sample selection, and compute the errors of the best sample. We observe that the proposal functions, learned to optimize the minimum errors, result in better metrics compared to the discrete function, especially through adaptive sampling.

2) **Trajectory Sample Selection:** We validate the effectiveness of our sample selection method, FPS, by comparing it with a few standard options, including i) a random sampler picking samples based on their weights, ii) a most-likely sampler that selects the top likely samples, similar to selecting the most-probable intent in [15] and best-$k$ enumeration in [29], iii) a sampler based on non-maximum suppression (NMS), as used in [6], which selects samples greedily by finding the next sample that is distant enough from existing samples given a threshold. For a fair comparison, the distance measure in NMS is the same as FPS based on final locations, and we empirically choose 2 different distance thresholds (2 meters and 4 meters) to select the next sample. If the number of valid NMS samples is smaller than $N$, we select the remaining samples randomly.

In the study, we first generate $M \in \{6, 30, 50\}$ samples using the proposal distribution, and select $N = 6$ samples. The results are summarized in Table II. When $M = N = 6$, no subsampling occurs. When $M > N$, a random sampler and a most-likely sampler do not improve the errors, as selecting only the most likely samples leads to worse errors, since trajectory prediction is a multi-modal problem. As $M$ grows, the most-likely sampler acts similar to a maximum likelihood estimator, and exhibits inferior results as the problem is multi-modal. NMS improves results but is limited by a fixed distance threshold: when the threshold is small (i.e. 2 meters), it fails to provide enough coverage in cases where the predicted samples are very far away; when the threshold is large (i.e. 4 meters), the number of valid samples can be smaller than $N$. On the other hand, FPS reduces the errors the most, by finding the 6 samples that provide both accuracy and coverage. When $M$ is larger than 50, the error reduction is small for both NMS and FPS.

### C. Quantitative Results

We compare our full model with a number of representative baselines, including i) DESIRE [18] that utilizes a conditional VAE model to generate trajectory prediction samples from a latent space; ii) DiversityGAN [3] that predicts diverse samples by learning a latent space in a GAN model such that the prediction samples with different semantic meanings are far away in that latent space. This baseline is similar to other diverse sampling works that improve coverage through a learned latent space [21]; iii) MultiPath [49] that learns the trajectory modalities as a set of anchors and predicts trajectories through anchor classification and offset regression; iv) TNT [6] that first infers discrete target locations and second predicts target-conditioned trajectories to support multi-modality. In addition, we introduce a few variants of our models to validate our hypothesis, including v) SingleMode that assumes a single mode and only samples from the continuous distribution; vi) HYPER-Linear that uses only linear layers in the decoder to simulate a linear dynamic system as in existing factored inference literature that assumes linear dynamics; vii) HYPER-NoMap that does not use the map as the input feature; viii) HYPER-Coverage that is trained with only the task-specific coverage loss. It shares the same spirit as [15] that optimizes for prediction coverage; ix) HYPER-Perturb that is trained on the same dataset, in which we randomly perturb 5% of discrete labels with a different mode chosen uniformly.

We use the metrics reported in [6], and present the comparison in Table III. HYPER outperforms all baselines that assume a fixed intent over time by using a fixed high-level intent state in the trajectory decoder (i-iv), ignore the discrete structure in the model (v), presume linear dynamics (vi), or ignore the map features (vii). In order to demonstrate its robustness with noisy discrete labels, we randomly perturbed 5% of discrete labels and observe that our model trained on the perturbed data achieves similar results (ix). We further improve the minADE metric with a variant,
Selection Method | 6 / 6 samples | 6 / 30 samples | 6 / 50 samples
--- | --- | --- | ---
Proposal + Random | 0.86 | 1.68 | 1.02
Proposal + Most-likely | 0.86 | 1.68 | 1.10
Proposal + NMS (2m) | 0.86 | 1.68 | 0.76
Proposal + NMS (4m) | 0.86 | 1.68 | 0.80
Proposal + FPS | 0.86 | 1.68 | 0.74

Table II: Accuracy over 3s for various sample selection methods. FPS achieves the best performance by selecting 6 samples from the proposal distributions.

TABLE III: Prediction performance over 3 seconds compared to baseline models. Our model balances between accuracy and coverage, with a variant (HYPER-Coverage) trained on the coverage task achieving the lowest minADE and MR.

TABLE IV: Compared to ManeuverLSTM, HYPER achieves better results in both discrete and continuous error metrics.

HYPER-Coverage, that is solely trained towards optimizing this metric, but sacrifices accuracy measured by the negative log-likelihood metric (NLL) [50]. Our method, on the other hand, balances between accuracy and coverage.

The work that is closest to ours in spirit is ManeuverLSTM (M-LSTM) [4]. It models driving modes explicitly as maneuvers labeled from trajectory data, and assumes the maneuver is fixed over time. For a fair comparison, we use the same model, training process, and definition of maneuvers as used in our model, except forcing each sample to have a fixed mode over the prediction horizons for ManeuverLSTM. We use five samples for a fair comparison given the number of maneuvers defined for ManeuverLSTM. In addition to the standard metrics, we introduce min-of-K discrete error rate (minDER) that measures the percentage of wrong predictions in discrete states for the best predicted sample, to quantify the discrete prediction accuracy and coverage. Table IV demonstrates that our model outperforms this baseline by a large margin, in both continuous and discrete metrics, by supporting evolving maneuver intent and utilizing a proposal function to explore the intent space.

D. Qualitative Results

In Fig. 3, we present a qualitative example to demonstrate the effectiveness of FPS. Fig. 3(a) shows the most likely examples selected based on the predicted likelihood, which fail to provide sufficient coverage. On the other hand, NMS (b) selects more diverse samples, but suffers from a fixed distant threshold. As a more robust and threshold-free alternative, FPS (c) finds diverse options more effectively.

In Fig. 4, we show a representative example that demonstrates the advantage of supporting evolving driving modes.

VI. Conclusion

In conclusion, we present a general and expressive hybrid prediction model that accounts for evolving discrete modes in the future trajectory. The model leverages learned proposal functions and the farthest point sampling algorithm to select a small number of accurate and diverse samples from an exponential space. The effectiveness of our model is validated in the Argoverse dataset, through both quantitative and qualitative experiments.