

Tractable Joint Prediction and Planning over Discrete Behavior Modes for Urban Driving

Adam Villafior¹, Brian Yang¹, Huangyuan Su¹, Katerina Fragkiadaki¹, John Dolan¹, Jeff Schneider¹

Abstract—Significant progress has been made in training multimodal trajectory forecasting models for autonomous driving. However, effectively integrating these models with downstream planners and model-based control approaches is still an open problem. Although these models have conventionally been evaluated for open-loop prediction, we show that they can be used to parameterize autoregressive closed-loop models without retraining. We consider recent trajectory prediction approaches which leverage learned anchor embeddings to predict multiple trajectories, finding that these anchor embeddings can parameterize discrete and distinct modes representing high-level driving behaviors. We propose to perform fully reactive closed-loop planning over these discrete latent modes, allowing us to tractably model the causal interactions between agents at each step. We validate our approach on a suite of more dynamic merging scenarios, finding that our approach avoids the *frozen robot problem* which is pervasive in conventional planners. Our approach also outperforms the previous state-of-the-art in CARLA on challenging dense traffic scenarios when evaluated at realistic speeds.

I. INTRODUCTION

Imagine a situation where a vehicle is exiting a crowded parking lot after a football game. Most drivers know that in bumper-to-bumper traffic, they should behave proactively by cautiously asserting themselves to make space for themselves and progress forward. While this is intuitive for human drivers, these highly interaction-dense scenarios are a major challenge for current autonomous vehicles.

In these situations, we require a robust model of how other agents might behave, and crucially, how they might respond to our own actions. To this end, significant progress has been made towards learning trajectory forecasting models from large datasets of urban driving logs [1], [2], [3], [4], [5]. State-of-the-art trajectory prediction models can capture the highly stochastic and multimodal distribution of outcomes in driving without needing to manually engineer complex human driving behaviors. In this work, we aim to explore fully leveraging these learned models for model-based planning and control.

Planning and prediction are usually treated as separate modules within a conventional autonomy stack. Usually a forecasting model will make predictions for all of the actors in the scene, and then plan open-loop against a set of possible open-loop trajectories for other agents in the scene. While this may ensure collision avoidance, it can cause the agent to behave overly conservatively, resulting in the *frozen robot problem* – if we are highly uncertain about nearby actors, then the only perfectly safe course of action is to do nothing,

which can result in deadlock. Ideally, we should perform fully reactive closed-loop prediction and planning, such that the ego-vehicle actions directly affect the predicted behaviors of other agents, and vice versa.

Unfortunately, performing fully reactive closed-loop planning over a learned forecasting model is often computationally intractable. The distribution over agent trajectories is highly multimodal and subject to change at every timestep, which means the space of possible outcomes grows exponentially over time. It is also impossible to enumerate and check every possible future, since the distribution is continuous. Existing work [6] performs fully reactive closed-loop planning by first parameterizing agent policies as continuous latents in a normalizing flow [7], and then optimizing the ego-agent policy with respect to some differentiable cost function, but these approaches have previously only been shown to work at small scale with one or two other interacting agents. We show later that our approach scales favorably on more challenging tasks.

In this work, we devise a novel approach for performing fully reactive closed-loop planning over multimodal trajectory prediction models. We consider recent approaches to trajectory prediction that leverage learned anchor embeddings [2], [4], [3] to predict a diverse set of trajectories. While these approaches have previously only been evaluated for open-loop prediction, to the best of our knowledge ours is the first work to show how these models can also be used for closed-loop planning for dense urban driving. Our contributions are as follows:

- **Closed-loop prediction with open-loop training.** We show how trajectory prediction models trained open-loop with learned anchor embeddings can be used to parameterize autoregressive closed-loop models without the need for retraining.
- **Fully reactive closed-loop planning over discrete latent modes.** We propose a novel planning approach which leverages discrete latent modes to do planning over a compact latent behavior space. Our planning approach scales linearly with the number of agents and the planning horizon, as opposed to naive search, which scales exponentially. This allows us to tractably model the causal interactions between agents when performing autoregressive rollouts; the predictions for other agents react to the planned ego-trajectory and vice versa. Similar to [6], this allows us to perform fully reactive closed-loop planning, but over discrete rather than continuous latent conditioned policies. We find that our discrete formulation ensures diverse trajectory proposals and

¹Carnegie Mellon University, Pittsburgh, PA, USA.
{avillaf1, brianyan, huangyus, kfragki2, jdolan, jeff4}@andrew.cmu.edu

improves downstream planning performance.

We validate our approach on a challenging suite of highly interactive urban driving scenarios, outperforming the demonstrator agent and several strong baselines. Our approach also beats the previous state-of-the-art in CARLA [8] on the Longest6 benchmark [9] when evaluated at realistic speeds.

Our code is available at <https://github.com/avillaflor/P2DBM>.

II. RELATED WORK

A. Trajectory forecasting models for driving

There has been a substantial amount of work towards training trajectory forecasting models for urban driving [10], [1], [5], [3], [4], [2]. However, none of these papers performs evaluations on closed-loop planning tasks beyond doing log-replay on offline driving logs for short time horizons. Most closely related to the model we use are Scene Transformer [3] and AutoBots [4], which both use vector-based input representations and large transformer models. Our model can support many state-of-the-art forecasting approaches, and only requires minimal modification to most in order to facilitate planning. Note that although all of these approaches are trained to perform open-loop prediction, our approach demonstrates how they can be adapted to perform fully reactive closed-loop prediction and planning without needing to substantially modify the training procedure.

B. Planning over learned forecasting models

There is a substantial amount of literature on planning over learned forecasting models for driving. Several prior works plan over imitation models, but do not perform closed-loop planning, instead optimizing a fixed ego-trajectory [11], [12], [13], [14], [15]. Another common approach is to not model the influence of the ego-agent on other agents [16], [17], [18]. As we will demonstrate, these approaches are unable to perform proactive planning when forced into close interactions with dense traffic. Model-based reinforcement learning approaches such as [19], [20] can theoretically leverage learned models and perform closed-loop prediction and planning for autonomous driving. However, many of these approaches have been restricted to small state spaces and have not been shown to perform well on more difficult closed-loop driving environments. In particular, we argue that rich vector-based state representations such as the one used in this work are key to effective prediction and planning. [21] models joint interactions between agents using a game-theoretic framework and evaluates planning performance in simulation, but does not use online fully reactive closed-loop planning like our approach.

The most similar work to ours in the literature is CfO [6], which also does fully reactive closed-loop planning for driving in CARLA. The tasks they consider are significantly simpler than the ones considered in this paper – they only consider interactions with one other vehicle (we consider up to 100). Their approach also relies on an autoregressive normalizing flow model which, to our knowledge, has not yet been scaled to larger-scale datasets, whereas our model

architecture is similar to other transformer-based forecasting models in the literature which have been deployed at scale.

C. Learning to drive in CARLA

Most competing approaches in CARLA [8] are based on imitation learning [22], [23], [24], [9], [25]. While most approaches focus on the visual imitation setting, our closest competitor is PlanT [26], which also does imitation learning, but operates in a similar setup to ours, where we assume the perception problem is solved. Imitation learning approaches rely on collecting optimal demonstrations, which can be challenging as we scale up to harder scenarios and more realistic data collection settings. We show that unlike imitation learning, our method is able to out-perform the demonstrator on our suite of challenging urban navigation tasks.

III. TRAJECTORY PREDICTION

A. Model formulation

In this section, we present our model architecture used for trajectory prediction. Let x_t^a denote the state of vehicle a at timestep t (we assume x_t^0 and $x_t^{1:A}$ are the vehicle states for the ego-vehicle and the rest of the vehicles respectively). We also assume that $t = 0$ is the current timestep and $x_{\leq T}$ refers to future vehicle states x_1, \dots, x_T (ignoring the current timestep). Our goal is to model the distribution over future vehicle states $P(x_{\leq T}|x_0, c)$ where c is some arbitrary conditioning information (omitted from now on for brevity). We can factorize this distribution autoregressively over time, as is common in conventional trajectory prediction approaches. In this work we only model marginal (as opposed to joint) distributions over agents, so we can additionally factorize our distribution as

$$P(x_{\leq T}|x_0) = \prod_{t=0}^T \prod_{a=0}^A P(x_{t+1}^a|x_t) \quad (1)$$

We propose a simple model formulation which captures multimodality and is well-suited for downstream reactive planning. We represent the high-level multimodal behavior of the agents with categorical latent variables $z_t^a \in [1, \dots, K]$ for each agent:

$$P(x_{\leq T}|x_0) = \prod_{t=0}^T \prod_{a=0}^A P(x_{t+1}^a|x_t, z_t^a)P(z_t^a|x_t) \quad (2)$$

Following [5], we implicitly characterize $P(x_{t+1}^a|x_t, z_t^a)$ using a deterministic mapping $x_{t+1}^a = f(x_t, z_t^a)$. Therefore, we only require our model to learn this 1-step prediction mapping $x_{t+1}^a = f(x_t, z_t^a)$, and estimate $P(z_t^a|x_t)$. In practice, we find that training our model to predict H steps open-loop into the future $x_{t+1:t+H}^a = f(x_t, z_t^a)$ is a useful auxiliary task that improves the training and generalization of the desired 1-step prediction mapping. Below we show how $f(x_t, z_t^a)$ and $P(z_t^a|x_t)$ can be parameterized by a transformer model.

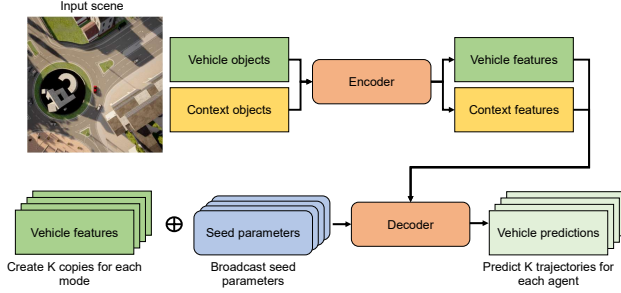


Fig. 1. Forecasting model architecture.

B. Network architecture

We design our architecture to predict each of the different modes for all agents independently, but all in one pass of the model. Thus, our model outputs a vector with shape $[K, A, H, 4]$ of potential future states of all agents in the scene, where K is the number of modes and A is the number of agents. This corresponds to $K \times A$ trajectories of length H , where each waypoint consists of position, heading, and speed. The high-level structure of our model is depicted in Figure 1.

Since the number of agents and context objects in the scene is not fixed, we adopt a flexible vector-based representation of the scene. We consider two types of entities: vehicles \mathcal{X} and scene context objects \mathcal{C} which consist of road points, traffic lights, pedestrians, stop signs, and goal waypoints. Each object is represented by a feature vector that contains the relevant raw information. For vehicles, this is relative pose, bounding box, speed, and current effective speed limit. For road points this is relative pose, lane width, whether the point is in an intersection, and whether one can change lanes to the left and right. For traffic lights this is relative pose, affected bounding box, and light state. For pedestrians this is relative pose, bounding box, and speed. For stop signs this is relative pose and affected bounding box. For goal waypoints this is relative pose and lane width. We assume access to all agents and context objects within 50m of the ego-vehicle up to a hard cap for each class of object. Each object is encoded using a MLP (one for each object type) to ensure all object features are the same dimension D . Then the entire set of input features is processed by an encoder to produce encoded vehicle features: $\mathcal{X}' = \text{Enc}(\mathcal{X}, \mathcal{C})$.

The encoder consists of stacked cross-attention layers where each vehicle cross-attends to both the vehicle features \mathcal{X} and the context features \mathcal{C} concatenated together. In practice, we find that only allowing each vehicle to attend to nearby map features improves performance, so we only attend to the closest 50 map features. Additionally, we use relative positional embeddings to maintain translational invariance. Prior work [27] has demonstrated that translational invariance is a useful inductive bias when performing multi-agent trajectory prediction. When feature u_i cross-attends to u_j , we compute query Q_{ij} , key K_{ij} , and value V_{ij} as

follows:

$$\begin{aligned} Q_{ij} &= W^q u_i \\ K_{ij} &= W^k (u_j + p_{ij}) \\ V_{ij} &= W^v (u_j + p_{ij}) \end{aligned}$$

where W^q, W^k, W^v are learned projection matrices and p_{ij} is a learned relative positional embedding. All features have an associated spatial position (x, y) as well as an orientation θ . To compute p_{ij} , we compute the pose (x_j, y_j, θ_j) of feature u_j transformed to the frame of feature u_i at pose (x_i, y_i, θ_i) to get a relative pose $(x_{ij}, y_{ij}, \theta_{ij})$. We then construct a relative pose feature by concatenating x_{ij} and y_{ij} , as well as $\sin(\theta_{ij})$ and $\cos(\theta_{ij})$. Finally, an MLP is used to encode this relative pose feature to obtain p_{ij} . Note that we do not include any absolute positional information in our features, so the network is only able to observe the relative poses of scene objects.

Following the encoder, our encoded vehicle features are of shape $[A, D]$. We expand these features to be of shape $[K, A, D]$ so that each unique combination of mode and actor corresponds to a single D -dimensional feature. Following [2], [3], [4], we introduce learnable anchor embeddings $M \in \mathbb{R}^{K \times D}$ for each of the K modes, which are broadcasted and added to the encoded vehicle features in order to distinguish the different modes. We learn two separate sets of anchor embeddings: one for the ego-vehicle only, and one for all other vehicles.

For each combination of agent and mode we get the query features $Q_k^a = \mathcal{X}'^a + M_k$ with appropriate broadcasting. Then, we use the decoder to produce K trajectory predictions for each agent: $Y_k = \text{Dec}(Q_k^a, \mathcal{X}')$. The decoder consists of two alternating operations: the query features Q_k^a cross-attend to both the encoded vehicle features \mathcal{X}' and map features \mathcal{C} , and then the query features Q_k^a self-attend along the K modes dimension. In other words, for each agent, the different mode features self-attend to each other. Note that during self-attention, only features corresponding to the same agent can attend to each other. This ensures that each agent future is decoded independently.

Finally, the processed query features are passed to a MLP. Since each unique agent and mode has its own feature, each D -dimensional feature is mapped to a trajectory of length H , where each waypoint consists of positions, orientations, and speeds. For each agent, we predict relative to that agent's coordinate frame, and also for each time step predict in-frame displacements rather than absolute positions, orientations, and speeds.

C. Training objectives

We train the model to predict the ground truth trajectories using a negative log-likelihood loss, where each prediction is parameterized by a Gaussian. Note that although we train our model to parameterize a Gaussian, we always take the mean when unrolling our model. In order to ensure our model can capture multimodal outcomes, we adopt a winner-takes-all objective [28] where only the closest of the K predictions will

backpropagate its loss. Since we are doing marginal prediction, our winner-takes-all objective takes the best prediction for each agent separately. Our model is also trained to output a logit for each mode to estimate the probability of that mode being used, and this is trained using a cross-entropy loss with a one-hot target for the closest of the K predictions.

IV. CLOSED-LOOP PLANNING

A. Planning with autoregressive rollouts

Our objective is to find the behaviors for the ego-vehicle that will maximize the discounted sum of rewards over T timesteps, where γ is a discount factor: $\sum_t^T \gamma^t R(x_t, c)$. Because the ego-vehicle’s decision in the trajectory is determined by the chosen z^0 , we can write our objective as

$$\operatorname{argmax}_{z^0} \mathbb{E} \left[\sum_t^T \gamma^t R(x_t, c) \right] \quad (3)$$

As illustrated in equation 2, we can sample $x_{\leq T}$ autoregressively using our transformer model. For each timestep, we just deterministically decode $x_{t+1}^a = f(x_t, z_t^a)$ for our latent sample z_t^a .

Given our specific transformer architecture, we can perform each autoregressive step of this procedure with just one forward pass of the model. We generate all the multimodal predictions at once by leveraging the anchor embeddings and pick the mode prediction that corresponds to $x_{t+1}^a = f(x_t, z_t^a)$. Note that even though we trained these models to make multi-step open-loop predictions, we only need to take the first step of the prediction, which is x_{t+1}^a . Then, we can repeat this procedure up to the decided horizon T in order to generate a sample of $x_{\leq T}$.

B. Evaluating ego-modes

Now in order to perform planning, we need to evaluate Equation 3 for specific values of z^0 . Since the ego latents $z^0 = [z_0^0, \dots, z_{T-1}^0]$ can vary between timesteps, z^0 can take on K^T possible values (K possible ego-modes for T timesteps), which is computationally expensive to fully enumerate.

Empirically, we find that keeping the latent modes consistent across time (i.e. $z_0 = \dots = z_{T-1}$) for both the ego-vehicle and surrounding vehicles leads to good performance, and thus we use this scheme in our experiments to reduce complexity. We hypothesize that this works because the learned anchor embeddings encourage the model to learn locally consistent modes since the embeddings are shared globally, i.e., they are not state-conditioned. Additionally, we find that conditioning on these different modes leads to meaningfully diverse behaviors over time. Thus by keeping the latent modes consistent across time, we only need to compare K ego latent modes to optimize Equation 3.

Specifically, we evaluate Equation 3 for any specific z^0 by generating N samples of $x_{\leq T}$ with the previously described procedure. At the first time step, for each surrounding agent we sample $z_0^a \sim P(\cdot|x_0)$ and for the ego-agent we set z_0^0 to the ego mode we are evaluating. Then, we continue to set $z_t = z_0$ at every subsequent timestep during autoregressive

generation. We estimate $\mathbb{E}[\sum_t^T \gamma^t R(x_t, c)]$ by taking the mean over these N samples. We do this for all K ego modes for z^0 , and pick the mode with the maximum expected sum of rewards. We use MPC and execute the first step by using it as a target for a PID controller, and continue to replan at every step.

This approach enables us to perform fully reactive closed-loop planning. Note that we are planning over latent modes rather than open-loop trajectories, which means we can effectively model the causal influence of the ego on other vehicles and vice versa. Additionally, this allows us to be robust to different responses from other agents, since we can evaluate each candidate ego latent z^0 under a variable number of latent samples.

C. Reward function

Our reward function is as follows:

$$\begin{aligned} R(x_t, c) &= \beta_{coll} R_{coll}(x_t, c) + \beta_{lane} R_{lane}(x_t, c) \\ &\quad + \beta_{speed} R_{speed}(x_t, c) + \beta_{light} R_{light}(x_t, c) \\ R_{coll}(x_t, c) &= -\mathbb{1}_{coll} \\ R_{lane}(x_t, c) &= 1 - \frac{|\text{lateral}(x_t, c)|}{0.5 \times (\text{lane width})} \\ R_{speed}(x_t, c) &= 1 - \frac{|\text{speed}(x_t) - (\text{speed limit})|}{(\text{speed limit})} \\ R_{light}(x_t, c) &= -\mathbb{1}_{red} \frac{\text{speed}(x_t)}{\text{speed limit}} \end{aligned}$$

where β_{coll} , β_{lane} , β_{speed} , and β_{light} control the relative weight assigned to each reward term. The collision term R_{coll} is -1 and leads to a termination if the ego collides with another object and 0 otherwise. The lane term R_{lane} penalizes deviation from the route, where $\text{lateral}(x_t, c)$ is the lateral distance of the ego from the closest route segment. The speed term R_{speed} encourages the ego to drive at the speed limit. The light term R_{light} penalizes the ego for having non-zero speed if the light is red; the penalty is scaled by the driving speed proportional to the speed limit.

V. RESULTS

A. Model training and hyperparameters

For all experiments, we train our method with 4 encoder layers consisting of a cross-attention layer followed by a feedforward layer, and 4 decoder layers consisting of a cross-attention layer, self-attention layer, and a feedforward layer. Following [29], we use Pre-Layer Normalization. For the merging scenarios, we use an embedding size of 128 for a model with approximately 1.9 million parameters. For the CARLA Longest6 benchmark, we use an embedding size of 256 for a model with approximately 7.4 million parameters. We train our model with AdamW [30] [31] with an initial learning rate of $2e-4$ with cosine annealing to 0 as we train for 100 epochs. We train our multimodal model with $K = 8$ latent modes and use $N = 8$ samples to estimate expected rewards during planning.

For both experiment settings, we use a visualization range of 50m and observe the closest vehicles by distance up to a

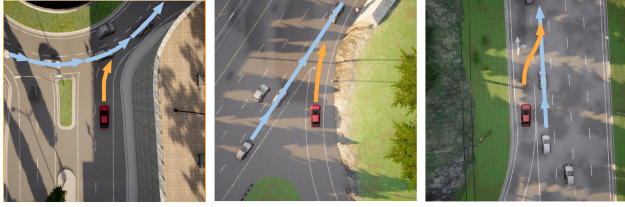


Fig. 2. Merge scenario visualization. The red car is the ego-vehicle, the orange arrow indicates the desired merging behavior, and the blue arrows indicate the flow of traffic in the target lane.

TABLE I
MERGE SCENARIOS

Approach	Success (%) \uparrow	Static (%) \downarrow	Crash (%) \downarrow
Ours	96.5 \pm 1.8	2.2 \pm 1.0	1.2 \pm 0.8
Open-Loop	78.3 \pm 2.1	20.5 \pm 2.3	1.2 \pm 0.2
Multimodal IL	50.2 \pm 1.0	46.5 \pm 1.5	3.3 \pm 0.6
Unimodal IL	51.8 \pm 1.6	46.3 \pm 2.	1.8 \pm 1.0
Continuous Latents	71.8 \pm 0.8	24.3 \pm 1.4	3.8 \pm 0.6
PlanT	67.2 \pm 1.2	30.7 \pm 0.9	2.2 \pm 0.3
Data Policy	68.0	12.5	19.5
Autopilot	85.5	1.5	13.0

cap of 100 vehicles. For the merging scenarios, we set the coefficients of the cost function as $\beta_{coll} = 20$, $\beta_{lane} = 0.1$, $\beta_{speed} = 1$, and $\beta_{light} = 0$, as there are no traffic lights in these scenarios. For the CARLA Longest6 benchmark, we set the coefficients of the cost function as $\beta_{coll} = 20$, $\beta_{lane} = 1$, $\beta_{speed} = 1$, and $\beta_{light} = 4$.

B. Merging scenarios

We design a suite of challenging urban navigation scenarios in order to evaluate whether our approach can safely perform proactive maneuvers in dense traffic. We identify 10 distinct merging scenarios that each involve a high degree of interaction with other cars in the midst of dense urban traffic. These other cars are controlled by CARLA’s internal traffic manager, which uses an adaptive controller with randomly initialized minimum safety distances and target speeds. In each scenario, the goal is point-to-point navigation where the goals are specified a priori. Three of these scenarios are visualized in Figure 2.

1) *Metrics and baselines*: An episode is considered a success if the ego-vehicle reaches the goal without committing any traffic infractions; otherwise, the episode is considered a failure. A common theme we found in prior CARLA driving benchmarks was that they rarely assessed the ego-agent’s assertiveness, instead relying solely on metrics related to collisions and other traffic infractions. As such, many of the top-performing approaches in CARLA tend to engage in overly conservative and unrealistic behaviors. To additionally assess the ability of each approach to be proactive, we introduce an additional infraction type for static episodes, i.e., situations where the car is stopped for long periods of time.

For each approach, we evaluate it on all 10 merging scenarios, each with 20 different random traffic initializations

for a total of 200 episodes. We separately report success rates, static rates, and crash rates for each approach. On these merge scenarios, we compare to several baselines:

- *Multimodal imitation*. Same transformer model as our approach, except we execute the mode prediction with the highest predicted probability.
- *Unimodal imitation*. Same transformer backbone as our approach, except we only use one mode and only make one prediction.
- *Ours w/ open-loop planning*. Same model as our proposed approach, except instead of unrolling our model autoregressively, we run our model once to generate marginal open-loop predictions for all cars. The planning procedure is the same as in the closed-loop case, except the trajectory predictions are fixed.
- *Ours w/ continuous latent variables*. We train a CVAE with Gaussian latent variables with our transformer backbone and do sampling-based planning for all agents (including the ego). The number of samples is equivalent to the total number of mode combinations expanded by our approach in order to ensure fair comparison.
- *PlanT*. We compare to PlanT [26], which is an imitation learning approach that is state-of-the-art on the CARLA Longest6 benchmark. We use the original code implementation of the model.
- *Data policy*. The configurable heuristic policy used to collect the dataset. We sample random configuration parameters (e.g. planning horizon, bounding box sizes, etc.) to generate diverse demonstrations.
- *Autopilot policy*. The best-performing configuration of the data collection policy.

To train all approaches, we collect a dataset of 100K samples using the data policy. We train our models on all scenarios and share them across baselines where possible. We report mean and standard error across 3 different model seeds.

2) *Performance analysis*: Table I shows the results on the merge scenario suite. Our approach outperforms the baselines and achieves the highest overall success rate. Notably our method is the only learned approach which actually outperforms the best-performing heuristic policy (autopilot). The open-loop variant of our approach performs worse than our closed-loop approach and has a higher static rate. The open-loop planner cannot predict that other vehicles will alter their current trajectory if the ego attempts to merge and therefore behaves much more conservatively, resulting in more static episodes. On the other hand, our closed-loop planner will predict that other vehicles will sometimes slow down to allow the ego-vehicle to merge. Figure 3 shows a qualitative comparison between the open and closed-loop planners. The imitation policies all have low crash rates due to excessive static episodes, and generally engage in extremely conservative behavior. This also suggests that the dominant modes in the dataset are highly cautious, even when we explicitly model multiple modes. However, our approach is able to identify and execute the high-performance modes in



Fig. 3. Qualitative example of proactive merging with closed-loop planning. The top row is from our proposed closed-loop planner and the bottom row is from the open-loop variant of our planner. Frames are in sequential order from left to right. The closed-loop planner merges proactively in front of other cars, causing the car behind to yield to the ego-agent. The open-loop planner does not predict that the ego will affect the behavior of other cars, so it does not attempt to merge.

the dataset.

Since our choice of discrete latents distinguishes our approach from previous work [6], we provide a more detailed comparison here between our approach and a CVAE baseline which uses continuous latents based on a Gaussian prior. We suspect the main reason for the disparity in performance is that our approach is better able to predict and sample distinct trajectories. Empirically we find that our discrete latent modes correspond to semantically different predictions that lead to a greater diversity of closed-loop predictions for both the ego-vehicle and the surrounding vehicles. On the other hand, taking arbitrary samples from a CVAE will not necessarily cover the relevant behavior modes, especially in a limited computation setting. Additionally, our approach explicitly enumerates distinct latent modes for the ego-vehicle independent of the predicted likelihood under the dataset, so it can more easily capture less common modes in the data.

C. Longest6 benchmark scenarios

1) *Benchmark details:* We also evaluate our approach on the CARLA Longest6 benchmark, which consists of driving on a difficult subset of the publicly available CARLA Leaderboard routes with added dense traffic. Unlike the previous scenario suite, these routes are much longer and consist of multiple scenarios defined according to the NHTSA typology.

We found that existing state-of-the-art approaches in CARLA Leaderboard are able to avoid many of the highly interactive scenarios by simply *driving significantly under the speed limit*. For instance, pedestrians on CARLA Leaderboard are regularly spawned to cross the road in front of the ego-vehicle at a constant speed. Many approaches avoid these scenarios by driving at a low enough speed such that the ego-vehicle will never be going fast enough to hit the spawned pedestrians. In general, driving at low speeds substantially

TABLE II
CARLA LONGEST6 SCENARIOS

Approach	DS \uparrow	RC \uparrow	IP \uparrow
Ours	51.0 \pm 1.2	89.2 \pm 1.7	0.582 \pm 0.010
Autopilot	44.6 \pm 1.6	84.0 \pm 1.1	0.546 \pm 0.040
PlanT	37.1 \pm 1.4	86.1 \pm 1.5	0.457 \pm 0.017
Multimodal IL	48.4 \pm 2.6	83.3 \pm 2.1	0.603 \pm 0.021

reduces the degree of interaction present in these scenarios.

To circumvent this loophole, we collect demonstration data with an autopilot policy driving at the speed limit. Specifically, we use the same autopilot to collect the data that was used in [26], but have the target speed be the current speed limit instead of a constant 4 m/s. This significantly increases the level of interaction required to solve these routes. In order to facilitate fair comparison with existing approaches, we retrained a state-of-the-art approach PlanT [26] using our new dataset.

2) *Metrics and baselines:* We adopt the main metric used by the CARLA Leaderboard. Driving score (DS) is the primary metric used to compare and rank approaches in CARLA Leaderboard, and is the product of a route completion score (RC) and an infraction penalty (IP) score. Route completion is the percentage of the route that is completed. Infraction penalty is a multiplier which penalizes various infractions such as collisions, running red lights, etc.

For baselines, we compare to the autopilot policy, a multimodal imitation baseline trained using our transformer backbone, and PlanT [26]. We collect 440K samples using the modified autopilot policy and use this dataset to train all approaches. For each approach, we evaluate on all 36 routes each with 5 different random seeds for a total of 180 episodes and we report the mean over all episodes and standard error over the 5 seeds.

3) *Performance analysis:* Table II shows the results on the CARLA Longest6 scenarios evaluated at reasonable driving speeds. Our closed-loop planning approach outperforms the autopilot used to collect the dataset as well as PlanT, which is the current state-of-the-art on the original Longest6 benchmark. While the imitation learning baseline has a better infraction penalty score, it ultimately has a lower overall driving score because it often deadlocks earlier in the denser routes and avoids many of the challenging interactions. In contrast, our approach is able to act proactively and make much more progress in these dense scenarios and achieve a superior driving score.

VI. CONCLUSION

In conclusion, we propose a novel approach for performing closed-loop planning over multimodal trajectory prediction models. Our approach is able to excel on a challenging suite of dynamic merging scenarios that require proactive planning behaviors. Additionally, we show that our approach outperforms state-of-the-art approaches on CARLA Longest6 scenarios when evaluated at reasonable driving speeds.

REFERENCES

- [1] C. Tang and R. R. Salakhutdinov, "Multiple futures prediction," *Advances in neural information processing systems*, vol. 32, 2019.
- [2] B. Varadarajan, A. Hefny, A. Srivastava, K. S. Refaat, N. Nayakanti, A. Cornman, K. Chen, B. Douillard, C. P. Lam, D. Anguelov *et al.*, "Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 7814–7821.
- [3] J. Ngiam, B. Caine, V. Vasudevan, Z. Zhang, H.-T. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal *et al.*, "Scene transformer: A unified architecture for predicting multiple agent trajectories," *arXiv preprint arXiv:2106.08417*, 2021.
- [4] R. Girgis, F. Golemo, F. Codevilla, J. A. D'Souza, M. Weiss, S. E. Kahou, F. Heide, and C. Pal, "Autobots: Latent variable sequential set transformers," *arXiv preprint arXiv:2104.00563*, vol. 3, no. 4, 2021.
- [5] S. Casas, C. Gulino, S. Suo, K. Luo, R. Liao, and R. Urtasun, "Implicit latent variable model for scene-consistent motion forecasting," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 2020, pp. 624–641.
- [6] N. Rhinehart, J. He, C. Packer, M. A. Wright, R. McAllister, J. E. Gonzalez, and S. Levine, "Contingencies from observations: Tractable contingency planning with learned behavior models," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 663–13 669.
- [7] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.
- [8] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.
- [9] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.
- [10] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine, "Precog: Prediction conditioned on goals in visual multi-agent settings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2821–2830.
- [11] N. Rhinehart, R. McAllister, and S. Levine, "Deep imitative models for flexible inference, planning, and control," *arXiv preprint arXiv:1810.06544*, 2018.
- [12] W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas, and R. Urtasun, "End-to-end interpretable neural motion planner," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8660–8669.
- [13] H. Song, W. Ding, Y. Chen, S. Shen, M. Y. Wang, and Q. Chen, "Pip: Planning-informed trajectory prediction for autonomous driving," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020, pp. 598–614.
- [14] J. Liu, W. Zeng, R. Urtasun, and E. Yumer, "Deep structured reactive planning," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4897–4904.
- [15] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 853–17 862.
- [16] J. Hardy and M. Campbell, "Contingency planning over probabilistic obstacle predictions for autonomous road vehicles," *IEEE Transactions on Robotics*, vol. 29, no. 4, pp. 913–929, 2013.
- [17] W. Zhan, C. Liu, C.-Y. Chan, and M. Tomizuka, "A non-conservatively defensive strategy for urban autonomous driving," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2016, pp. 459–464.
- [18] A. Cui, S. Casas, A. Sadat, R. Liao, and R. Urtasun, "Lookout: Diverse multi-future prediction and planning for self-driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 107–16 116.
- [19] J. Wu, Z. Huang, and C. Lv, "Uncertainty-aware model-based reinforcement learning with application to autonomous driving," *arXiv preprint arXiv:2106.12194*, 2021.
- [20] A. Kamenev, L. Wang, O. B. Bohan, I. Kulkarni, B. Kartal, A. Molchanov, S. Birchfield, D. Nistér, and N. Smolyanskiy, "Predictionnet: Real-time joint probabilistic traffic prediction for planning, control, and simulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8936–8942.
- [21] Z. Huang, H. Liu, and C. Lv, "Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving," *arXiv preprint arXiv:2303.05760*, 2023.
- [22] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Conference on Robot Learning*. PMLR, 2020, pp. 66–75.
- [23] D. Chen, V. Koltun, and P. Krähenbühl, "Learning to drive from a world on rails," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 590–15 599.
- [24] D. Chen and P. Krähenbühl, "Learning from all vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 222–17 231.
- [25] K. Chitta, A. Prakash, and A. Geiger, "Neat: Neural attention fields for end-to-end autonomous driving," in *International Conference on Computer Vision (ICCV)*, 2021.
- [26] K. Renz, K. Chitta, O.-B. Mercea, A. Koepke, Z. Akata, and A. Geiger, "Plant: Explainable planning transformers via object-level representations," *arXiv preprint arXiv:2210.14222*, 2022.
- [27] A. Cui, S. Casas, K. Wong, S. Suo, and R. Urtasun, "Gorela: Go relative for viewpoint-invariant motion forecasting," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 7801–7807.
- [28] S. Narayanan, R. Moslemi, F. Pittaluga, B. Liu, and M. Chandraker, "Divide-and-conquer for lane-aware diverse trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 799–15 808.
- [29] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, "On layer normalization in the transformer architecture," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 524–10 533.
- [30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.