2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

A Generalizable Physics-guided Causal Model for Trajectory Prediction in Autonomous Driving

Zhenyu Zong[1], Yuchen Wang[1], Haohong Lin[2], Lu Gan[3] and Huajie Shao[1]

***Abstract*— Trajectory prediction for traffic agents is critical for safe autonomous driving. However, achieving effective zero-shot generalization in previously unseen domains remains a significant challenge. Motivated by the consistent nature of kinematics across diverse domains, we aim to incorporate domain-invariant knowledge to enhance zero-shot trajectory prediction capabilities. The key challenges include: 1) effectively extracting domain-invariant scene representations, and 2) integrating invariant features with kinematic models to enable generalized predictions. To address these challenges, we propose a novel generalizable Physics-guided Causal Model (PCM), which comprises two core components: a *Disentangled Scene Encoder*, which adopts intervention-based disentanglement to extract domain-invariant features from scenes, and a *CausalODE Decoder*, which employs a causal attention mechanism to effectively integrate kinematic models with meaningful contextual information. Extensive experiments on real-world autonomous driving datasets demonstrate our method's superior zero-shot generalization performance in unseen cities, significantly outperforming competitive baselines. The source code is released at `https://github.com/ZY-Zong/Physics-guided-Causal-Model`.**

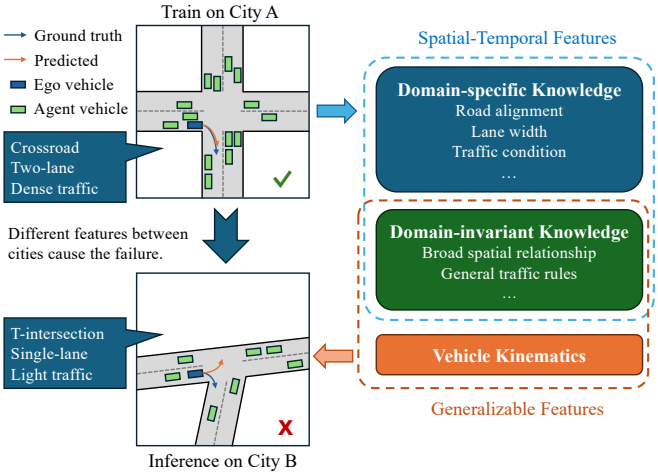

Fig. 1. Illustration of two right turn scenarios in two different cities. The model trained on city A can make correct prediction, but fail in unseen city B with different lane width, traffic condition and road alignment. To solve this problem, our method leverages two generalizable knowledge: **domain-invariant knowledge** separated from the spatial-temporal features and **vehicle kinematics**, to enhance domain generalization ability.

I. Introduction

Trajectory prediction aims to forecast the future paths of dynamic agents in autonomous driving scenarios [1], [2], [3], [4], which is critical for ensuring driving safety. Recent studies [5], [6] have primarily developed machine learning (ML) approaches for end-to-end trajectory prediction by leveraging multiple data modalities, including images, LiDAR, polylines, and waypoints. However, purely data-driven ML is prone to spurious correlation [7], thus struggling to generalize to unseen domains.

To address this problem, existing works have developed various domain generalization techniques, such as transfer learning [8], [9] and trajectory tokenization for next-token prediction [10]. While these methods contribute to enhancing domain generalization, several challenges remain. Transfer learning approaches are often computationally expensive and struggle with zero-shot trajectory prediction. Additionally, trajectory tokenization techniques require massive data to effectively bridge the domain gap, making them less feasible in data-scarce or rapidly changing environments. Moreover, these purely data-driven ML approaches may not comply with physical laws. To deal with these issues, recent studies [11], [12], [13] have developed physics-guided ML that incorporates system dynamics to enhance the robustness and generalizability of trajectory prediction. However, existing approaches often struggle to capture meaningful contextual information in dynamic and complex environments. In addition, none of them focuses on zero-shot prediction across two different domains.

To bridge this gap, we propose a generalizable Physics-guided Causal Model (PCM) that integrates domain-invariant features with vehicle kinematics to enhance zero-shot trajectory prediction in unseen domains. As shown in Fig. 1, models trained in City A often fail to generalize to scenarios in an unseen City B due to discrepancies in road alignments, lane widths, and traffic conditions. These characteristics represent domain-specific knowledge derived from spatial-temporal distributions. In contrast, broad spatial patterns such as road topology, and general traffic regulations are typically consistent across different urban environments. Motivated by this observation, we hypothesize that domain-invariant knowledge remains stable across cities, and that predicted trajectories should additionally conform to vehicle kinematic constraints to ensure physical plausibility and real-world deployability. The key challenges we are addressing include: (i) effectively extracting domain-invariant scene representations from map polylines, and (ii) integrating domain-invariant representations with kinematics to enhance the zero-shot

[1]Zhenyu Zong, Yuchen Wang and Huajie Shao are with Department of Computer Science, William & Mary, Williamsburg, VA 23185, USA {zzong, ywang142, hshao}@wm.edu
[2]Haohong Lin is with SafeAI Lab, College of Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA haohongl@andrew.cmu.edu
[3]Lu Gan is with School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA lgan@gatech.edu

generalizability of trajectory prediction.

To overcome the first challenge, we develop a *Disentangled Scene Encoder* to learn domain-invariant features through intervention-based disentanglement. To address the second challenge, we introduce a *CausalODE Decoder* that devises a causal attention mechanism to integrate invariant features and kinematics-guided predictions. Specifically, to comply with physical laws, we incorporate a two-wheel kinematic model [14] into neural ordinary differential equation (ODE) [15] to learn vehicle dynamics and then use it to initialize a trajectory query for the decoder. Besides, inspired by causal representation learning [7], [16], we design a causal attention mechanism based on the Transformer decoder [17] to fuse domain-invariant features and kinematics for capturing the interactions between them, thereby enhancing zero-shot generalization capability.

To evaluate the performance of our method, we implement two zero-shot experiments on real-world driving datasets: i) trained on nuPlan [18] and evaluated on nuScenes [19], and ii) trained on WOMD [20] and evaluated on nuPlan. Both of them show that our method significantly outperforms baselines in zero-shot trajectory predictions, suggesting its superior generalization capability.

Our contributions include: 1) we propose a novel generalizable Physics-guided Causal Model (PCM) that can enhance zero-shot trajectory prediction performance in unseen domains; 2) we introduce a Disentangled Scene Encoder to learn domain-invariant scene representations and a CausalODE Decoder to integrate learned invariant representations with kinematics to significantly enhance model's generalization capability; and 3) we present extensive experiments to validate the superior zero-shot prediction performance of our approach using real-world trajectory prediction datasets.

II. Related Works

Domain Generalization of Trajectory Prediction. Enhancing domain generalization of trajectory prediction remains an open research question. Existing techniques leverages transfer learning [21] together with random masks [8], [9] to tackle this challenge. However, these methods require high pretraining costs and finetuning models on new domains. In addition, they struggle with out-of-distribution (OOD) scenarios [22], [23]. SMART [10] leverages discrete representation to bridge different domains. However, it requires massive data to construct a generalizable token vocabulary. In addition, the discrepancy between the true value and its discretized representation can accumulate, which is harmful over long trajectories. A most recent study [24] has proposed Adaptive Prediction Ensemble (APE) that combines a rule-based expert and a learning-based expert to improve OOD generalization across datasets. However, since constant velocity rule-based expert struggle with predicting curved trajectories accurately, APE has to select from two inferior predictions when its learning-based expert cannot understand unseen turning scenarios. Different from existing works, we propose a new method that combines invariant scene features and kinematics to enhance the zero-shot generalization ability.

Physics-guided ML for Trajectory Prediction. Recent studies [13], [25], [11], [12] have explored physics-guided ML methods for generalized representation learning in trajectory prediction. For example, Geng et al. [13] proposed a physics-informed hybrid model that combines a data-driven Transformer with the Intelligent Driver Model (IDM) to predict final trajectories. Tischmann et al. [25] incorporated a physics-informed loss based on IDM to guide neural network training. Other works [11], [12] embedded physical principles directly into machine learning architectures to enhance model generalization. While these methods demonstrate improved performance through the integration of physics, their evaluations are limited to within the same dataset. Moreover, they often fail to capture generalized contextual information from dynamic and complex environments, which may lead to poor predictions in unseen domains. To tackle this issue, we introduce a Disentangled Scene Encoder for extracting domain-invariant contextual features and a causal attention decoder for fusing these features with kinematics.

Causal Representation Learning in Autonomous Driving. Some research also employed causal representation learning based on Structural Causal Models (SCMs) [7], [26], [16], [27] to improve generalization capability of trajectory prediction. For instance, CausalHTP [26] utilized causal graphs and counterfactual interventions to avoid environment biases that mislead prediction results. CaDeT [7] and CaST [16] proposed disentanglement approaches to separate spurious parts from causal relations and then de-confound spurious causality. However, these methods either pre-define the causal structure or require causal annotations like [28]. To tackle these issues, a recent work, FUSION [27], leverages safety-aware causal Transformer to learn causality of Reinforcement Learning (RL) inputs. Liu et al. [29] adopted sim-to-real causal transfer to learn multi-agent causal interactions without real-world causal annotations. CCDiff [30] automatically identify and inject causal structures into the diffusion model. Unlike previous works, we introduce a causal intervention-based technique to disentangle domain-invariant features from scenes without needing causal annotations and pre-defined causal structures. These features are then integrated with vehicle kinematics through the attention block to learn causal interactions between context features and physics.

III. Proposed Method

A. Problem statement

The objective is to predict the future trajectory of the ego agent in a multi-agent autonomous driving scenario consisting of $n-1$ neighboring agents. Let $\mathcal{D}_H = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$ represent the observed trajectories of these agents from time steps t_0 to t_H in history, where $\boldsymbol{X}_i$ denotes the observed trajectories of agent i, namely, $\boldsymbol{X}_i = [\boldsymbol{s}_i(t_0), \ldots, \boldsymbol{s}_i(t_H)]^\top$. For agent i, $\boldsymbol{s}_i(t_j) = [x_i(t_j), y_i(t_j), \theta_i(t_j), v_i(t_j)]^\top$ represents its state, including

the position, heading, and velocity at time step j. Let $\boldsymbol{M}$ denotes map polylines, which is represented by coordinate positions, directions, and lane types. Given environmental context $\boldsymbol{M}$ and observed trajectories $\mathcal{D}_H$, the goal of this work is to predict future trajectory of the ego agent e in prediction horizon T, denoted by $\boldsymbol{Y}_e = [\boldsymbol{s}_e(t_{H+1}), \ldots, \boldsymbol{s}_e(t_{H+T})]^\top$.

B. Overall architecture

To achieve domain generalization of the predictor, we develop a new physics-guided causal model to enhance zero-shot generalization capability, as shown in Fig. 2. The core idea is to disentangle domain-invariant features from scenes and then integrate them with kinematics for trajectory prediction. The proposed method comprises two main components: (i) Disentangled Scene Encoder, aiming to learn domain-invariant representations from scenes; and (ii) CausalODE Decoder that captures the interactions between invariant features and vehicle dynamics.

C. Disentangled scene encoder

First, we adopt a disentangled scene encoder to learn domain-invariant features from map polylines, and then feed them into the scene encoder. By extracting invariant features, the scene encoder is able to generate meaningful context representations devoid of domain-specific information, e.g. different layout of unimportant map polylines in different cities. As a result, the trajectory decoder would rely exclusively on domain-invariant features to perform zero-shot prediction.

As shown in Fig. 2, we take observed trajectories of agents $\boldsymbol{X}$ and map polylines $\boldsymbol{M}$ as inputs. Both inputs are respectively encoded into feature embeddings $\boldsymbol{Z}_X$ and $\boldsymbol{Z}_M$ using the multilayer perceptron (MLP). As illustrated in Fig. 2 (a), in order to learn invariant features from map polylines, we learn disentangled scene representation with the intervention mechanism that separates $\boldsymbol{Z}_M$ into domain-invariant and domain-variant groups, denoted by $\boldsymbol{Z}_M^+$ and $\boldsymbol{Z}_M^-$ respectively. We set the first k percent of latent features to be domain-invariant and the remaining to be domain-variant. Inspired by causal invariance in Structural Causal Model (SCM) [7], disentangling can be achieved by cutting the causal relation between $\boldsymbol{Z}_M^+$ and $\boldsymbol{Z}_M^-$ with the intervention denoted by $do(\boldsymbol{Z}_M^-)$. The core idea is to replace $\boldsymbol{Z}_M^-$ with an intervention set to simulate revoking domain-variant features. In this work, assuming latent features follow multivariate Gaussian distribution, we construct an intervention set with data sampled from standard Gaussian distribution.

After intervention, both original and intervened map features are concatenated with the vehicle trajectory representation $\boldsymbol{Z}_X$ to obtain fused modality embeddings, respectively. Both embeddings are then fed into a transformer-based scene encoder $\boldsymbol{g}_{\theta_1}$ [31] to obtain meaningful context information $\boldsymbol{C}, \tilde{\boldsymbol{C}} \in \mathbb{R}^{d \times h}$ respectively, where d is the spatial latent dimension size and h is the hidden latent dimension size.

D. CausalODE decoder

To further enhance the generalizability of trajectory prediction, we introduce a CausalODE Decoder that integrates domain-invariant features with kinematics. It consists of two parts: 1) a physics-guided neural ODE that serves as query based on vehicle dynamics, and 2) a causal attention module based on Transformer decoder [6] that captures interactions among context features and vehicle dynamics.

Physics-guided neural ODE: We aim to integrate kinematics into the neural ODE framework to learn the underlying governing equations. The main idea of neural ODE is to use deep neural networks (DNN) to approximate the vector field as $\frac{\mathrm{d}\mathbf{s}(t)}{\mathrm{d}t} = \Phi(\mathbf{s}(t), t, \phi)$, where Φ is the vector field parameterized by DNN and $\boldsymbol{s}(t)$ is system states at time t.

Recall that the vehicle state $\boldsymbol{s}(t)$ contains four attributes, i.e., (x, y, θ, v), which represents the position of vehicle wheel center, heading, and velocity respectively. Following prior works [14], [32], we use the bicycle kinematic model to model vehicle dynamics as follows:

$$\Phi(\boldsymbol{s}(t), t, \phi) = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} v\cos\theta \\ v\sin\theta \\ \frac{v\tan\delta}{L} \\ a \end{bmatrix}. \tag{1}$$

δ, a, L represent steering angle, acceleration, and wheelbase length of the vehicle. These unknown dynamics terms corresponds to the third and fourth rows in Equation (1), and are learned via an MLP. Specifically, for each mode i, the terms $(\dot{\theta}_i, \dot{v}_i)$ are predicted by an MLP as $(\dot{\theta}_i, \dot{v}_i) = \mathrm{MLP}_i(x_i, y_i, \theta_i, v_i)$. These learned components vary across vehicles and over time, capturing individual behavior and temporal dynamics.

Given the above vehicle dynamics, and let the trajectory's initial state be the current observed state $\boldsymbol{s}(t_H)$, trajectory prediction over a time horizon T can be formulated as an initial value problem (IVP), and then solved by:

$$\begin{aligned} \boldsymbol{s}(t) &= \boldsymbol{s}(t_H) + \int_{t_H}^{t} \Phi\big(\boldsymbol{s}(\tau),\, \tau,\, \phi\big)\, \mathrm{d}\tau, \\ \boldsymbol{s}(t_{H+j}) &= \mathrm{ODESolve}\big(\Phi,\, \boldsymbol{s}(t_H),\, t_H, t_{H+j}\big), \\ &\text{for } j = 1, \ldots, T. \end{aligned} \tag{2}$$

To capture uncertainty of unknown system coefficients in Eq. (1), we use m neural ODEs with different parameters to learn vehicle dynamics in complex driving scenarios such as left turn and right turn.

Causal attention module: Its inputs include outputs from neural ODEs, and context features generated by the scene encoder. To learn the causal relationships between them, we employ multi-head attention mechanism [6] to attend queries with context features. Specifically, for each prediction mode $j = 1, \ldots, m$ and spatial index $i = 1, \ldots, d$, the query, key and value for the decoder are:

$$\boldsymbol{Q}_j = \boldsymbol{S}_j \boldsymbol{W}_{\boldsymbol{Q}}, \boldsymbol{K}_i = \boldsymbol{C}_i \boldsymbol{W}_{\boldsymbol{K}}, \boldsymbol{V}_i = \boldsymbol{C}_i \boldsymbol{W}_{\boldsymbol{V}}, \tag{3}$$

where $\boldsymbol{S}$ represents the mode prediction along horizon T. $\boldsymbol{W}_{\boldsymbol{Q}}, \boldsymbol{W}_{\boldsymbol{K}}, \boldsymbol{W}_{\boldsymbol{V}}$ denotes the weight of query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and

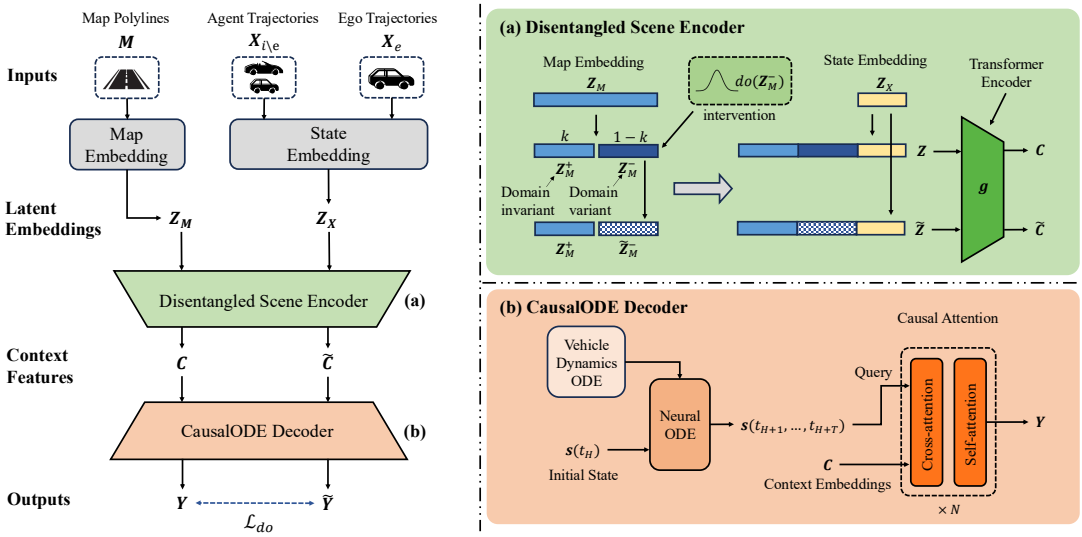

Fig. 2. Overall framework of the proposed method. It comprises two main parts: (a) a Disentangled Scene Encoder aiming to extract domain-invariant features; (b) a CausalODE decoder that integrates domain-invariant features with vehicle kinematics learned by neural ODE.

value $\boldsymbol{V}$, respectively. If we set the hidden feature dimension of each mode prediction as d_k. Then, the attention matrix $\boldsymbol{A}$ among input nodes are:

$$\boldsymbol{A} = \texttt{Softmax}(\boldsymbol{Q}\boldsymbol{K}^\top/\sqrt{d_k}). \tag{4}$$

Each value of attention matrix $\boldsymbol{A}_{i,j}$ indicates the impact of context feature $\boldsymbol{C}_i$ to each trajectory $\boldsymbol{S}_j$. We then compute each head's output by weighting the values, concatenating the vectors, and projecting back yields the physics query output. We follow each cross-attention block with a self-attention block to enhance the physics query internal relationship. After repeating this process for N layers, we we apply a final linear projection to produce the predicted trajectory $\boldsymbol{Y}$.

E. Objective function

The overall objective is composed of two loss functions: (i) intervention loss and (ii) Gaussian Mixture Model (GMM) loss commonly used in trajectory prediction [6], [33], [34]. We detail these two loss functions below.

Intervention loss: Recall that we feed both original and domain-invariant latent representations $\boldsymbol{Z}$ and $\tilde{\boldsymbol{Z}}$ through the encoder $\boldsymbol{g}_{\theta_1}$, resulting in context features $\boldsymbol{C} = \boldsymbol{g}_{\theta_1}(\boldsymbol{Z})$ and $\tilde{\boldsymbol{C}} = \boldsymbol{g}_{\theta_1}(\tilde{\boldsymbol{Z}})$. These two features are fed into the decoder, denoted by $\boldsymbol{f}_{\theta_2}$, to predict trajectories $\boldsymbol{Y}$ and $\tilde{\boldsymbol{Y}}$, respectively. To learn domain-invariant features, we force these two predicted trajectories to be close to each other. Mathematically, we have the following objective function:

$$\arg\min_{\theta_1,\theta_2} \mathcal{L}_{do}(\boldsymbol{Y},\tilde{\boldsymbol{Y}}) = \frac{1}{T}\sum_{j=H+1}^{H+T} \left\| \boldsymbol{f}_{\theta_2}(\boldsymbol{C}) - \boldsymbol{f}_{\theta_2}(\tilde{\boldsymbol{C}}) \right\|^2, \tag{5}$$

where H is the current time step and T is the prediction horizon. $\boldsymbol{g}_{\theta_1}$ and $\boldsymbol{f}_{\theta_2}$ denote the encoder and decoder, respectively.

GMM loss: During model training, we aim to find the nearest predicted trajectory compared with ground truth $\hat{\boldsymbol{Y}}$ and minimize their difference. Given m different modes of predictions, define $\hat{i}$ as the nearest index. This give the best predtion $\boldsymbol{Y}_{\hat{i}}$ with probability $p_{\hat{i}}$ Then the GMM loss [33] contains classification and regression part as below:

$$\mathcal{L}_{\text{GMM}}(\boldsymbol{Y},\hat{\boldsymbol{Y}}) = \mathbb{E}_{(x,y)}\left[\mathcal{L}_{\text{reg}}\left(\boldsymbol{Y}_{\hat{i}},\hat{\boldsymbol{Y}}\right) + \mathcal{L}_{\text{cls}}\left(p_{\hat{i}}\right)\right], \tag{6}$$

where $\mathcal{L}_{\text{reg}}$ denote negative log-likelihood of the ground-truth trajectory under the bivariate Gaussian distribution of the selected nearest mode, and $\mathcal{L}_{\text{cls}}$ denote the the cross entropy to maximize the probability of the selected trajectory $\boldsymbol{Y}_{\hat{i}}$.

Overall objective: The final objective is the sum of the intervention loss and two Gaussian Mixture Model (GMM) loss:

$$\mathcal{L} = \mathcal{L}_{GMM}(\boldsymbol{Y},\hat{\boldsymbol{Y}}) + \mathcal{L}_{GMM}(\tilde{\boldsymbol{Y}},\hat{\boldsymbol{Y}}) + \lambda\mathcal{L}_{do}(\boldsymbol{Y},\tilde{\boldsymbol{Y}}), \tag{7}$$

where λ is a hyperparameter that balances the third term.

IV. Experiments

A. Datasets

We evaluate our method using public dataset described below. In data preprocessing, we use ScenarioNet [35] and unitraj [36] to generated unified descriptions of three different datasets to achieve cross-dataset evaluation. Specifically, (1) **nuScenes** dataset [19] contains 850 train-validation scenarios and 150 test scenarios collected from Boston and Singapore. (2) **nuPlan** dataset [18] provides a broader coverage, with data from Boston, Singapore, Las Vegas, and Pittsburgh. Since nuScenes and nuPlan share city maps, we use only the Pittsburgh subset (35,058 scenes) for training in nuPlan to prevent overlap during cross-dataset evaluation, while the original nuPlan test split (20,756 scenes from all four cities) is used for testing. (3) **Waymo Open Motion Dataset (WOMD)** [20] offers a significantly larger scale, with 487K training scenes and 44K test scenes collected from six cities including San Francisco, Phoenix, Mountain View, Los Angeles, Detroit, and Seattle—all of which are unseen in nuPlan. To reduce computational costs, we randomly sample 10% of the training scenes from WOMD for training.

B. Implementation details

For feature embeddings, we use 1-layer and 2-layer MLP respectively to project agents trajectories and map polylines into hidden spaces with size of 240. We select the closest 384 map polylines around the ego agent, with up to 20 points for each polyline. For disentanglement, we set domain-invariant percentage to $k = 0.5$. We will also explore the effect of k on model performance in Section IV-F. The weight of intervention loss is $\lambda = 1$.

For scene encoder, we use a Perceiver-like [31] architecture with 1 cross-attention layer followed by 2 self-attention layers. For the decoder, we use 6 3-layer MLPs with hidden size of 200 to parameterize ODEs and solve IVP with Dopri5 solver [37]. The causal attention block consists of 8 causal attention layers.

For all experiments, models are trained 100 epochs with AdamW optimizer and OneCycleLR learning rate scheduler on 4 Nvidia RTX A5000 GPUs.

C. Metrics

We use four unified metrics in unitraj [36] for evaluation. For all metrics, we consider the top $k = 6$ prediction modes. Specifically, we use the following metrics:

- **Minimum Average / Final Displacement Error (minADE / minFDE)**: It computes the average / final time step of L2 distance error between ground truth and the closest trajectory among 6 modes of predictions:
- **Miss Rate (MR)**: The prediction is defined to miss when final trajectory displacement exceeding 2 meters. Miss rate is the ratio of miss samples at prediction horizon T.
- **Brier minimum Final Displacement Error (brier-minFDE)**: It accounts for probability of the best predicted trajectory p in minFDE by adding the penalty value $(1 - p)^2$ to the L2 distance.

D. Baselines

To assess the performance of our method, we not only compare it against competitive trajectory prediction baselines, but also some advanced methods with generalization capabilities. Specifically, **Wayformer** [6] fuses multimodal input features with self-attention encoder to generate spatial-temporal relationship, which is attended to predict trajectories through cross-attention decoder. **Autobot** [38] uses multi-head self-attention modules to learn social and temporal interactions, allowing scene-consistent multi-agent trajectory predictions. **MTR** [34] integrates global intention localization and local movement refinement with a unified transformer, which ensures the model to learn agent interactions and predict scene-compliant predictions. **G2LTraj** [39] simultaneously generates local intermediate trajectory predictions between global key steps with spatial-temporal constraints. **APE** [24] selects either learning-based or rule-based prediction expert with the highest ranking score for trajectory prediction. **Forecast-MAE** [9] randomly masks both agent trajectories and map elements, which are used to pretrain a scene-level encoder. **RMP** [8] randomly masks historical agent trajectories and pretrains an agent-centric motion encoder that learns motion dynamics and temporal continuity. **SMART** [10] models agent trajectories and vectorized map as discrete sequence tokens for next-token prediction via a decoder-only transformer.

TABLE I

NUPLAN-TO-NUSCENES: PERFORMANCE COMPARISON BETWEEN OUR METHOD AND BASELINES FOR ZERO-SHOT EVALUATION ON NUSCENES, TRAINED ON NUPLAN DATASET. RESULTS ARE AVERAGED OVER THREE RANDOM SEEDS. **BOLD** REPRESENTS THE BEST RESULTS.

Method	minADE_6	minFDE_6	brier -minFDE	Miss Rate
Wayformer [6]	1.036	2.621	3.299	0.469
AutoBot [38]	1.197	2.782	3.430	0.465
G2LTraj [39]	1.188	2.754	3.419	0.467
MTR [34]	1.360	3.236	3.877	0.600
Forecast-MAE [9]	1.159	2.673	3.974	0.602
RMP [8]	1.485	3.460	4.032	0.653
SMART [10]	1.812	3.587	4.513	0.695
APE [24]	3.655	9.214	9.214	0.889
Ours	**0.897**	**2.205**	**2.857**	**0.398**

TABLE II

WOMD-TO-NUPLAN: PERFORMANCE COMPARISON BETWEEN OUR METHOD AND BASELINES FOR ZERO-SHOT EVALUATION ON NUPLAN, TRAINED ON WOMD DATASET. RESULTS ARE AVERAGED OVER THREE RANDOM SEEDS. **BOLD** REPRESENTS THE BEST RESULTS.

Method	minADE_6	minFDE_6	brier -minFDE	Miss Rate
Wayformer [6]	0.745	2.034	2.635	0.317
AutoBot [38]	0.851	2.308	2.934	0.340
G2LTraj [39]	0.872	2.324	2.938	0.388
MTR [34]	0.791	2.125	2.621	0.394
Forecast-MAE [9]	0.936	2.531	2.985	0.407
RMP [8]	1.375	2.832	3.247	0.419
SMART [10]	1.614	4.082	4.626	0.492
APE [24]	3.122	8.373	8.373	0.766
Ours	**0.720**	**1.951**	**2.562**	**0.315**

E. Comparison against the state-of-the-art

We compare the zero-shot generalization performance of our method with baselines. Specifically, we train models on one dataset with 21 history time steps and zero-shot evaluate them on another with 60 prediction time steps. The performance are evaluated on the top $k = 6$ prediction modes with four unified metrics **minADE**, **minFDE**, **brier-minFDE** and **miss rate** in unitraj [36].

Evaluation on nuScenes and trained on nuPlan. As shown in Table I, our method outperforms the second best baseline by a margin of 0.139, 0.416, 0.442, and 0.067 in terms of minADE, minFDE, brier-minFDE and miss rate respectively. We attribute the improvement to physics knowledge and domain-invariant features. It reveals that our method can learn these two generalized knowledge so that the model maintains the performance well in unseen scenarios. The state-of-art generalizable trajectory prediction method, APE, fails to achieve good performance in our

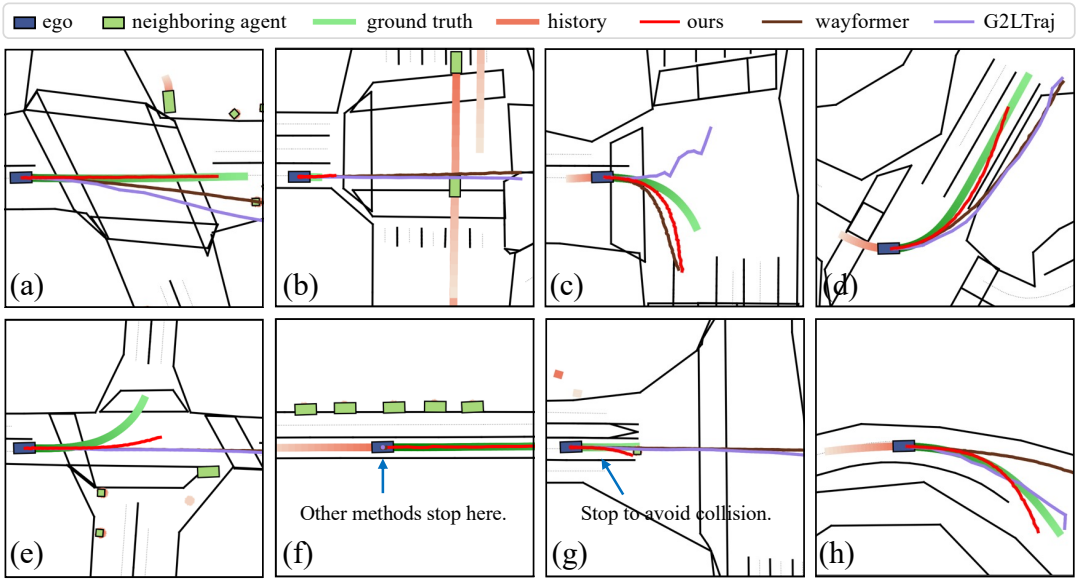

Fig. 3. The trajectory prediction visualization for top-three methods on **8** different nuScenes scenarios. (a) Our method drives normally, while others crash into pedestrians. (b) Our method stops before the pavement in the crossroad with heavy traffic, while two other predictions crash into other vehicles. (c) Our method turns right smoothly, while G2LTraj turns left. G2LTraj and Wayformer's predictions cannot follow vehicle kinematics well in this scenario. (d) Our method turns left without driving out of the road. (e) Our method correctly turns left at the crossroad. (f) Our method drives normally with vehicles parking aside the road, while other methods stop. (g) Our method stops to avoid collision to the pedestrian ahead. (h) Our method turns right smoothly, indicating it follows vehicle kinematics.

experiments, indicating it may not fully understand the road conditions in unseen domains. As a result, its routing function struggles to select the best prediction expert for trajectory prediction. In addition, our method outperforms Forecast-MAE, RMP and SMART, indicating our method's superior zero-shot generalization ability.

Evaluation on nuPlan and trained on WOMD. Compared to nuScenes and nuPlan, WOMD is an even larger-scale dataset which contains more complicated driving data by covering more turning and fewer straight scenarios. As shown in Table II, our method still achieves the best performance in complex scenarios with smaller minADE, minFDE, brier-minFDE and miss rate values of 0.025, 0.083, 0.059 and 0.002 respectively. It is noticeable that baselines like wayformer and MTR, are able to achieve promising trajectory prediction performance in unseen datasets. This indicates that, with sufficient training data, the spatial-temporal relationships learned by their models may contain generalizable information. Nevertheless, our method improves it further by learning vehicle dynamics and domain-invariant features. Forecast-MAE and RMP fail to deliver strong performance because randomly masking out data within a single dataset does not guarantee the learning of generalizable features that can effectively transfer to other datasets. This risks overfitting to dataset-specific patterns while ignoring transferable motion semantics. Similarly, SMART underperforms due to its tokenization strategy, which may lead to the loss of fine-grained motion nuances critical for precise trajectory forecasting.

F. Ablation study

Impact of key components. We also investigate the impact of two important components: (a) disentangled scene encoder with intervention loss, and (b) CausalODE decoder, on trajectory prediction performance using nuScenes dataset, as shown in Table III. We can observe that the performance drops 0.157, 0.514, 0.543 and 0.1 for minADE, minFDE, berierFDE and miss rate, respectively, when removing physics-guided ODE. This reveals learning physical dynamics can boost the generalization. Similarly, removing disentanglement and intervention loss from our method gives worse performance, which shows the effectiveness of disentanglement to generate domain-invariant features. Finally, the result without both components shows that fusion of physical dynamics and domain-invariants features can attend their relations well and boost model generalization capability.

Impact of domain-invariant percentage. Recall that we control the percentage of domain-invariant features in the disentangled scene encoder as illustrated in Table III with the hyperparameter k. We investigate the impact of k on the nuScenes dataset to assess its impact on model performance. As shown in Table VII, the performance degrades when k is set either too small or too large. When k is too small, the scene encoder fails to learn latent representations since most map embeddings are revoked. Conversely, when k is too large, the domain-variant components lack sufficient latent dimensions and are captured within the domain-invariant feature groups. In our experiments, we choose k = 0.5, which gives the best performance.

Parameter sensitivity. We conduct experiments on the nuPlan-to-nuScenes setting to analyze the sensitivity of parameter λ with values sampled from $\{0, 0.25, 0.5, 0.75, 1\}$. As shown in Fig. 4, minADE remains relatively stable when $\lambda > 0$, but increases significantly when $\lambda = 0$. Overall, the proposed method get the best performance with $\lambda = 1$, so we set $\lambda = 1$ in our experiments.

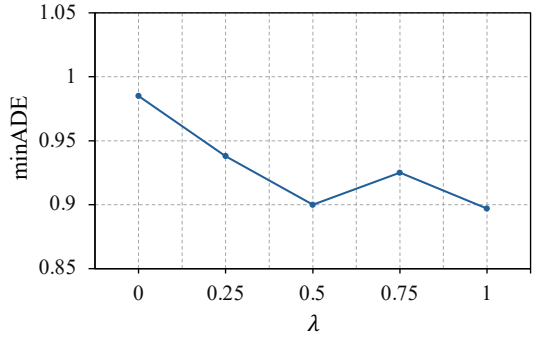

Fig. 4. Parameter sensitivity of λ.

TABLE III

ABLATION STUDIES OF THE PROPOSED METHOD. EXPERIMENTS ARE TRAINED ON NUPLAN AND EVALUATED ON NUSCENES.

(a)	(b)	minADE_6	minFED_6	brier-minFDE	Miss Rate
×	×	1.484	3.907	4.567	0.667
×	✓	1.031	2.652	3.309	0.400
✓	×	1.010	2.526	3.196	0.453
Ours		**0.853**	**2.012**	**2.653**	**0.353**

TABLE IV

IMPACT OF DOMAIN-INVARIANT PERCENTAGE k. EXPERIMENTS ARE TRAINED ON NUPLAN AND EVALUATED ON NUSCENES.

k	minADE_6	minFED_6	brier-minFDE	Miss Rate
0.2	1.115	2.756	3.381	0.533
0.4	0.999	2.573	3.252	0.480
0.6	0.972	2.439	3.105	0.493
0.8	1.012	2.560	3.228	0.487
0.5 (Ours)	**0.853**	**2.012**	**2.653**	**0.353**

G. Qualitative examples

In Fig. 3, we visualize the best predicted trajectories of top three methods in eight unseen scenarios in nuScenes. In case (a), where the traffic condition at the intersection is relatively clear, our method executes a normal driving maneuver, whereas Wayformer and G2LTraj fail to detect pedestrians and collide with them. In case (b), cross-street traffic is actively moving. Our model is able to stop safely before the pedestrian crossing, while the baselines overlook the traffic context and crash into surrounding vehicles. In case (c), both Wayformer and our method correctly execute a right turn consistent with the ground truth, while G2LTraj erroneously turns left. Notably, our predicted trajectory is smoother than those from the baselines, highlighting the effectiveness of the vehicle kinematic constraints embedded in CausalODE. Similarly, the smooth right turn depicted in case (h) further supports our model's compliance with physical vehicle dynamics. Case (d) illustrates our model's ability to safely execute a left turn without veering off the road. In case (e), our method correctly turns left to avoid both pedestrians on the right and a vehicle ahead, suggesting a strong understanding of spatial interactions in the scene. In case (f), despite parked vehicles along the roadside, our method proceeds normally, while baseline models halt unnecessarily. Finally, in case (g), our model appropriately stops to avoid a pedestrian directly ahead, demonstrating context-aware behavior.

In summary, these cases reveal that our method is generalizable to unseen scenarios by learning domain-invariant knowledge and vehicle kinematics.

TABLE V

INFERENCE LATENCY COMPARISON BETWEEN OUR METHOD AND BASELINES. WE MEASURE ALL THE MODEL'S ONE MODE PREDICTION TIME ON ONE RTX A5000 GPU (BATCH SIZE=1), AVERAGED OVER THE WHOLE NUSCENSE TEST DATASET.

Method	Wayformer	AutoBot	G2LTraj	MTR	Ours
Latency (ms)	13.93	**10.05**	10.16	65.00	37.23
Method	F-MAE	RMP	SMART	APE	
Latency (ms)	21.82	12.46	11.75	42.38	

H. Inference time

To guarantee safety in autonomous driving, the inference speed of the model should satisfy real-time requirements. In Table V, we investigate the inference speed of our method. Though vehicle kinematics increase the processing time, our method takes 37.23 ms, which still satisfy real-time requirement [40] within 100 ms.

V. CONCLUSION

In this work, we proposed a novel generalizable physics-guided causal model to enable zero-shot trajectory prediction performance by learning domain-invariant features and physical dynamics. Our method is composed of two core components: the disentangled scene encoder to effectively extract domain-invariant parts of scene representations, and the CausalODE decoder to fuse domain-invariant features with physics-guided trajectory queries. Extensive experimental results on real-world driving datasets demonstrated the superior zero-shot generalizability of our method over baselines.

Despite these promising results, our current implementation relies on a simplified bicycle model, which may not fully capture the complex dynamics of real-world vehicles. In future work, we plan to incorporate more sophisticated kinematic models to further enhance the physical fidelity.

ACKNOWLEDGMENTS

Research reported in this paper was sponsored in part by NSF CPS 2311086, NSF CIRC 716152, NSF RITEL 2506890, NAIRR 250288, and Faculty Research Grant at William & Mary 141446.

REFERENCES

[1] S. Afshar, N. Deo, A. Bhagat, T. Chakraborty, Y. Shao, B. R. Buddharaju, A. Deshpande, and H. C. Motional, "Pbp: Path-based trajectory prediction for autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12 927–12 934.

[2] B. Lange, J. Li, and M. J. Kochenderfer, "Scene informer: Anchor-based occlusion inference and trajectory prediction in partially observable environments," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14 138–14 145.

[3] Z. Huang, X. Mo, and C. Lv, "Multi-modal motion prediction with transformer-based neural network for autonomous driving," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2605–2611.

[4] L. Ye, Z. Zhou, and J. Wang, "Improving the generalizability of trajectory prediction models with frenét-based domain normalization," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2023, p. 11562–11568.

[5] L. Dal'Col, M. Oliveira, and V. Santos, "Joint perception and prediction for autonomous driving: A survey," *arXiv preprint arXiv:2412.14088*, 2024.

[6] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2980–2987.

[7] M. Pourkeshavarz, J. Zhang, and A. Rasouli, "Cadet: A causal disentanglement approach for robust trajectory prediction in autonomous driving," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 14 874–14 884.

[8] Y. Yang, Q. Zhang, T. Gilles, N. Batool, and J. Folkesson, "Rmp: A random mask pretrain framework for motion prediction," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 3717–3723.

[9] J. Cheng, X. Mei, and M. Liu, "Forecast-mae: Self-supervised pre-training for motion forecasting with masked autoencoders," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8679–8689.

[10] W. Wu, X. Feng, Z. Gao, and Y. Kan, "Smart: Scalable multi-agent real-time motion generation via next-token prediction," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 114 048–114 071.

[11] R. Gan, H. Shi, P. Li, K. Wu, B. An, L. Li, J. Ma, C. Ma, and B. Ran, "Goal-based neural physics vehicle trajectory prediction model," *arXiv preprint arXiv:2409.15182*, 2024.

[12] Y. Mao, Y. Gu, L. Sha, H. Shao, Q. Wang, and T. Abdelzaher, "Phytaylor: Partially physics-knowledge-enhanced deep neural networks via nn editing," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[13] M. Geng, J. Li, Y. Xia, and X. M. Chen, "A physics-informed transformer model for vehicle trajectory prediction on highways," *Transportation research part C: emerging technologies*, vol. 154, p. 104272, 2023.

[14] R. Tumu, L. Lindemann, T. Nghiem, and R. Mangharam, "Physics constrained motion prediction with uncertainty quantification," in *2023 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2023, pp. 1–8.

[15] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.

[16] Y. Xia, Y. Liang, H. Wen, X. Liu, K. Wang, Z. Zhou, and R. Zimmermann, "Deciphering spatio-temporal graph forecasting: A causal lens and treatment," 2023. [Online]. Available: https://arxiv.org/abs/2309.13378

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] K. T. e. a. H. Caesar, J. Kabzan, "Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles," in *CVPR ADP3 workshop*, 2021.

[19] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020.

[20] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, Z. Yang, A. Chouard, P. Sun, J. Ngiam, V. Vasudevan, A. McCauley, J. Shlens, and D. Anguelov, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 9710–9719.

[21] L. Ullrich, A. McMaster, and K. Graichen, "Transfer learning study of motion transformer-based trajectory predictions," in *2024 35th IEEE Intelligent Vehicles Symposium (IV)*. Jeju Island, Korea: IEEE, 2024, pp. 110–117.

[22] A. Vettoruzzo, M.-R. Bouguelia, J. Vanschoren, T. Rögnvaldsson, and K. Santosh, "Advances and challenges in meta-learning: A technical review," 2023. [Online]. Available: https://arxiv.org/abs/2307.04722

[23] D. Nguyen, S. Gupta, T. Nguyen, S. Rana, P. Nguyen, T. Tran, K. Le, S. Ryan, and S. Venkatesh, "Knowledge distillation with distribution mismatch," in *Machine Learning and Knowledge Discovery in Databases. Research Track*, N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, and J. A. Lozano, Eds. Cham: Springer International Publishing, 2021, pp. 250–265.

[24] J. Li, J. Li, S. Bae, and D. Isele, "Adaptive prediction ensemble: Improving out-of-distribution generalization of motion forecasting," 2024. [Online]. Available: https://arxiv.org/abs/2407.09475

[25] P. Tischmann, R. Baumann, and A. S. Novo, "Physics informed deep learning for motion prediction in autonomous driving," in *AmEC 2024–Automotive meets Electronics & Control; 14. GMM Symposium*. VDE, 2024, pp. 7–12.

[26] G. Chen, J. Li, J. Lu, and J. Zhou, "Human trajectory prediction via counterfactual analysis," 2021. [Online]. Available: https://arxiv.org/abs/2107.14202

[27] H. Lin, W. Ding, Z. Liu, Y. Niu, J. Zhu, Y. Niu, and D. Zhao, "Safety-aware causal representation for trustworthy offline reinforcement learning in autonomous driving," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, p. 4639–4646, May 2024. [Online]. Available: http://dx.doi.org/10.1109/LRA.2024.3379805

[28] R. Roelofs, L. Sun, B. Caine, K. S. Refaat, B. Sapp, S. Ettinger, and W. Chai, "Causalagents: A robustness benchmark for motion forecasting using causal relationships," 2023. [Online]. Available: https://openreview.net/forum?id=9WdB5yVICCA

[29] A. Rahimi, P.-C. Luan, Y. Liu, F. Rajic, and A. Alahi, "Sim-to-real causal transfer: A metric learning approach to causally-aware interaction representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025 [forthcoming publication]*, 2025.

[30] H. Lin, X. Huang, T. Phan-Minh, D. S. Hayden, H. Zhang, D. Zhao, S. Srinivasa, E. M. Wolff, and H. Chen, "Causal composition diffusion model for closed-loop traffic generation," *arXiv preprint arXiv:2412.17920*, 2024.

[31] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *International conference on machine learning*. PMLR, 2021, pp. 4651–4664.

[32] P. Wang, M. Zhu, X. Zheng, H. Lu, H. Zhong, X. Chen, S. Shen, X. Wang, Y. Wang, and F.-Y. Wang, "Bevgpt: Generative pre-trained foundation model for autonomous driving prediction, decision-making, and planning," *IEEE Transactions on Intelligent Vehicles*, pp. 1–13, 2024.

[33] Y. Chai, B. Sapp, M. Bansal, and D. Anguelov, "Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction," *arXiv preprint arXiv:1910.05449*, 2019.

[34] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6531–6543, 2022.

[35] Q. Li, Z. Peng, L. Feng, Z. Liu, C. Duan, W. Mo, and B. Zhou, "Scenarionet: open-source platform for large-scale traffic scenario simulation and modeling," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[36] Y. Zhu, J. J. Yu, X. Zhao, X. Wei, and Y. Liang, "Unitraj: Universal human trajectory modeling from billion-scale worldwide traces," *arXiv preprint arXiv:2411.03859*, 2024.

[37] M. Lienen and S. Günnemann, "torchode: A parallel ODE solver for pytorch," in *The Symbiosis of Deep Learning and Differential Equations II, NeurIPS*, 2022. [Online]. Available: https://openreview.net/forum?id=uiKVKTiUYB0

[38] R. Girgis, F. Golemo, F. Codevilla, M. Weiss, J. A. D'Souza, S. E. Kahou, F. Heide, and C. Pal, "Latent variable sequential set transformers for joint multi-agent motion prediction," *arXiv preprint arXiv:2104.00563*, 2021.

[39] Z. Zhang, Z. Hua, M. Chen, W. Lu, B. Lin, D. Cai, and W. Wang, "G2ltraj: A global-to-local generation approach for trajectory prediction," *arXiv preprint arXiv:2404.19330*, 2024.

[40] Y. Luo, "Time constraints and fault tolerance in autonomous driving systems," *Tech. rep, Tech. Rep*, 2019.