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

HetroD: A High-Fidelity Drone Dataset and Benchmark for Autonomous Driving in Heterogeneous Traffic

Yu-Hsiang Chen[1,2] Wei-Jer Chang[2] Christian Kotulla[3] Thomas Keutgens[3] Steffen Runde[3]
Tobias Moers[3] Christoph Klas[3] Wei Zhan[2] Masayoshi Tomizuka[2] Yi-Ting Chen[1,†]

[1]National Yang Ming Chiao Tung University [2]UC Berkeley [3]fka GmbH

† Corresponding Author

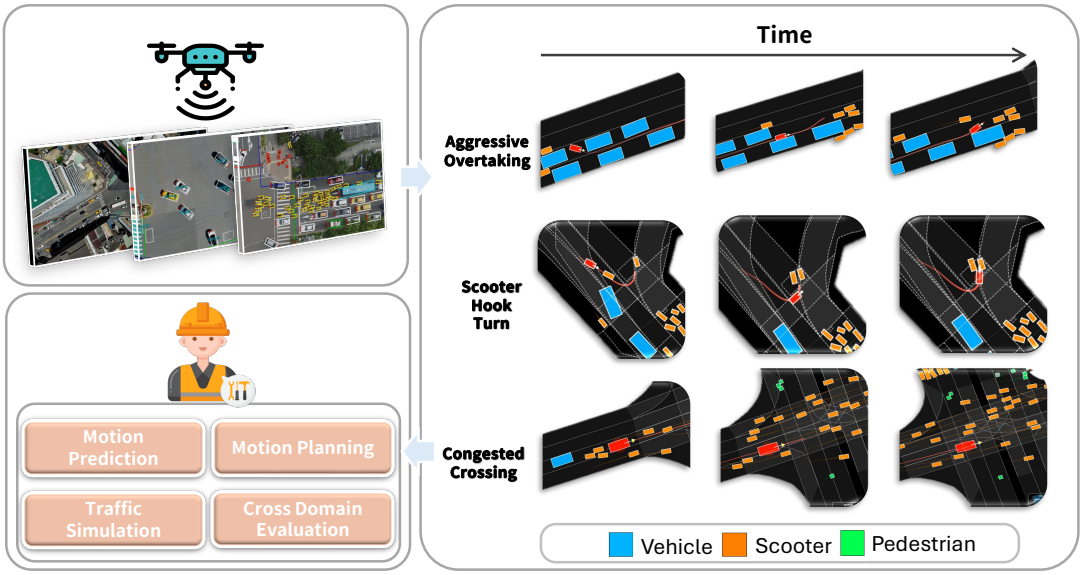

Fig. 1: *HetroD* is a high-fidelity, drone dataset that captures unstructured maneuvers such as hook turns, aggressive overtakes, queue cutting, and congested crossings among vehicles, scooters, and pedestrians in heterogeneous traffic environments. These maneuvers are critical for testing autonomous driving systems yet remain underexplored in the community. To address this, we construct a benchmark to evaluate existing methods in motion planning, motion prediction, traffic simulation, and conduct a thorough investigation of their generalization across datasets.

***Abstract*— We present HetroD, a dataset and benchmark for developing autonomous driving systems in heterogeneous environments. HetroD targets the critical challenge of navigating real-world heterogeneous traffic dominated by vulnerable road users (VRUs), including pedestrians, cyclists, and motorcyclists that interact with vehicles. These mixed agent types exhibit complex behaviors such as hook turns, lane splitting, and informal right-of-way negotiation. Such behaviors pose significant challenges for autonomous vehicles but remain underrepresented in existing datasets focused on structured, lane-disciplined traffic. To bridge the gap, we collect a large-scale drone-based dataset to provide a holistic observation of traffic scenes with centimeter-accurate annotations, HD maps, and traffic signal states. We further develop a modular toolkit for extracting per-agent scenarios to support downstream task development. In total, the dataset comprises over 65.4k high-fidelity agent trajectories, 70% of which are from VRUs. HetroD supports modeling of VRU behaviors in dense, heterogeneous traffic and provides standardized benchmarks for forecasting, planning, and simulation tasks. Evaluation results reveal that state-of-the-art prediction and planning models struggle with the challenges presented by our dataset: they fail to predict lateral VRU movements, cannot handle unstructured maneuvers, and exhibit limited performance in dense and multi-agent scenarios, highlighting the need for more robust approaches to heterogeneous traffic. See our project page for more examples: https://hetroddata.github.io/HetroD/**

I. Introduction

Navigating heterogeneous traffic remains one of the core challenges in the development of autonomous driving systems, particularly due to vulnerable road users (VRUs), including cyclists, pedestrians, and motorcyclists, who interact with vehicles in complex ways. In recent years, data-driven modeling has become the dominant approach for autonomous driving development, as it provides a scalable way to capture complex traffic interactions with less human effort than rule-based modeling. However, most publicly available datasets primarily capture lane-disciplined traffic and vehicle-to-vehicle interactions [8]. They include little data on VRUs (Table I) and heterogeneous interactions. As a result, they miss many real-world situations where

TABLE I: Comparison of Datasets on Interaction, Density & Diversity Metrics. We report key statistics across on-board and drone-view datasets. *Interaction Scale*[1] is the total number of interactions, computed per dataset by aggregating over all its scenarios and then normalized across datasets. *Heterogeneous Interaction Scale*[2] counts cross-type interactions with the same per-dataset aggregation and normalization. *Geographical Density*[3] represents the average number of agents per unit area A within an 8-second window. *VRUs*[4] denotes the proportion of VRUs among all traffic agents. All metrics except VRUs are min–max normalized to [0,1] across datasets where the metric is available; VRUs is reported in percent. Normalization is used for cross-dataset comparability rather than absolute interaction frequency. Boldface indicates the highest value among all datasets, while underlined values denote the highest among drone-view datasets.

Dataset	Platform	Tracks	Duration	Interaction Scale[1]	Heterogeneous Interaction Scale[2]	Geographical Density[3]	VRUs (%)[4]
NuScenes [1]	On-board	∼90k†	320h	0.675	0.549	—	20.1%
Waymo [2]	On-board	7.6M	574h	1.000	1.000	—	11.5%
Argoverse2 [3]	On-board	13.9M	763h	0.632	0.318	—	10.0%
NuPlan [4]	On-board	∼5M†	1282h	0.274	0.213	—	46.3%
INTERACTION [5]	Drone	40k	16.5h	0.132	—	0.011	—
inD [6]	Drone	13.5k	10h	0.071	0.185	0.023	39.4%
SinD [7]	Drone	13.2k	7.02h	0.099	0.324	0.016	62.1%
HetroD	Drone	<u>65.4k</u>	<u>17.5h</u>	<u>0.223</u>	<u>0.889</u>	**0.026**	**69.9%**

† Estimated values based on official statistics.
— Metric not available.
[1] $\mathcal{S}_{\text{inter}} = \sum_{\text{scenarios}} \mathcal{D}_{\text{inter}}$.
[2] $\mathcal{S}_{\text{het}} = \sum_{\text{scenarios}} \sum_{i,j} 1_{(\text{TTC}_{i,j}<2\,\text{s} \wedge \text{type}_i \neq \text{type}_j)}$.
[3] $\mathcal{D}_{\text{geo}} = N/A$, where N is the number of agents within an 8 s window and A is the corresponding area.
[4] VRUs $= 100 \times \frac{N_{\text{VRU}}}{N_{\text{VRU}}+N_{\text{Veh}}}$ (VRU: pedestrians, bicycles/cyclists, motorcycles, tricycles; Vehicles: cars, trucks, buses, vans).

different road users compete for space and negotiate right-of-way through subtle, culture-specific cues [9]. Therefore, downstream models and widely used simulators inherit these biases: they either hard-code simplified VRU templates [10] or merely replay recorded trajectories, such as Waymax [11] and NuPlan [4], which limits their ability to capture heterogeneous reactive dynamics. These limitations are further discussed in Section II.

This gap between current datasets and real-world scenes calls for data that captures the intricate interactions among mixed agents. Achieving such comprehensive data collection requires observation methods that overcome the occlusions and limited field-of-view inherent in on-board sensors. Drone-based observation provides a holistic scene coverage and temporal evolution of traffic participants, essential attributes for VRU interaction modeling.

We introduce *HetroD*, a drone-captured dataset collected across six topologically diverse, high-traffic urban locations in Taiwan. While the existing drone-based dataset SinD [7] marks an important first step in capturing heterogeneous traffic interactions, HetroD offers a substantially larger interaction scale, with up to twice the number of cross-agent interactions (Table I). In addition, HetroD involves a wide range of intricate maneuvers such as hook turns, lane splitting, and aggressive overtakes and offers topological diversity across six intersection archetypes. The dataset further includes centimeter-accurate HD maps, bounding boxes, and traffic signal states. Together, these traits position HetroD as a new testbed for developing autonomous driving systems in dense heterogeneous traffic.

Our contributions are summarized as follows:

- We construct a drone dataset with centimeter-level annotations of heterogeneous traffic. The dataset spans 17.5 hours and contains over 65.4k agent tracks, 70% from VRUs.
- We establish benchmarks with standardized evaluation protocols for motion prediction, planning, and cross-dataset evaluation.
- We show that state-of-the-art prediction and planning methods face clear limitations on HetroD: they struggle to predict lateral VRU movements, handle specific maneuvers, and sustain performance in dense multi-agent scenarios.

II. Related Work

Autonomous driving datasets vary by their sensing modalities and deployment context. We group related work into four categories: on-board, infrastructure-view, drone-view datasets, and unified development frameworks. Table I summarizes key characteristics of relevant datasets.

On-Board Sensor Datasets. Prior works [12], [13], [2], [1], [14], [15], [16], [3], [17], [18], [19] offer rich multimodal data but suffer from occlusions and limited VRU coverage in dense traffic [12], [2], [1]. While METEOR [20] pioneered heterogeneous traffic capture, its vehicle-centric data collection approach underrepresents VRU interactions. Furthermore, it lacks HD maps and comprehensive annotations needed for detailed traffic analysis.

Infrastructure-View Datasets. These datasets [21], [22], [23], [24], [25], [26], [27], [28], [29] use fixed cameras or V2X sensors to reduce occlusion, but often suffer from low spatial resolution that hampers small object detection [21], [25], inaccurate camera calibration affecting accurate localization [27], [28], or limited cross-agent type diversity [26], [23], limiting their utility for modeling heterogeneous agent behaviors.

Drone-View Datasets. Existing drone datasets [30], [31], [5], [22], [32], [33], [6], [34], [35], [7], [36], [37], [38], [39], [40] provide occlusion-free, global views and are ideal for interaction modeling and analysis. However, many are collected in lane-disciplined settings [31], [34] and exhibit fragmented VRU tracks due to the difficulty of annotating

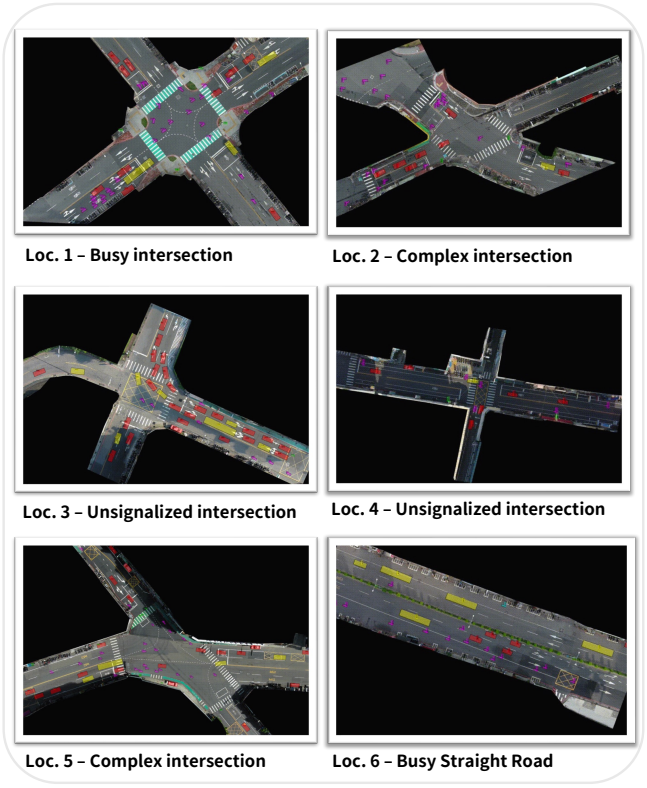

Fig. 2: Aerial views of the six recording locations in HetroD, capturing diverse urban traffic scenarios.

small objects [41], [40]. In addition, they also underrepresent unstructured maneuvers, such as informal yielding, weaving, or reverse flows. To the best of our knowledge, we are the first dataset that provides both per agent, centimeter accuracy in ground truth, and wide area coverage across diverse, heterogeneous urban environments for benchmarking autonomous vehicles in VRU-rich contexts.

Unified Development Toolkit. Recent development toolkits for tasks such as trajectory forecasting, scenario generation, and reinforcement learning [42], [43], [44], [45], [46], [47], [48] offer standardized interfaces to facilitate rapid development. However, existing drone datasets lack a suitable toolkit to enable collective development within the community. To address this, we make the first attempt to streamline existing toolkits such as ScenarioNet [43] and GPUDrive [47] (see Fig. 5) to be compatible with HetroD. We invite the community to tackle these critical challenges collectively.

III. The HetroD Dataset

HetroD is a large-scale drone-view dataset comprising 17.5 hours of ultra-high-resolution (5.4K) video, collected across six topologically and behaviorally distinct urban sites in Taiwan. As shown in Fig. 2, the dataset captures diverse traffic scenarios including busy signalized intersections (Locations 1, 2, and 5), unsignalized intersections (Locations 3 and 4), and a busy straight road segment (Location 6). It encompasses over 65.4k unique trajectories across these varied traffic environments—archetypes rarely represented together in existing datasets.

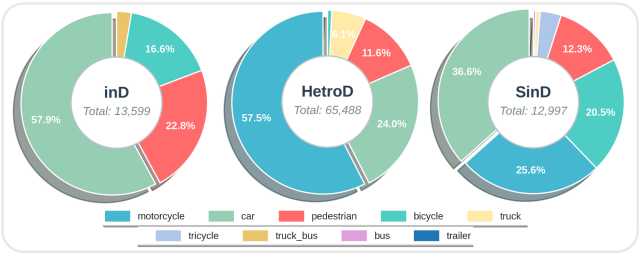

Fig. 3: Agent-type distribution of HetroD (center) compared against two prior datasets, inD and SinD.

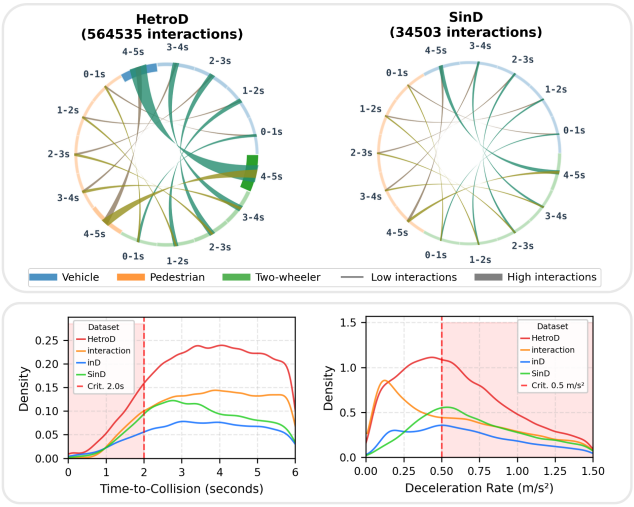

Fig. 4: Cross-type interaction patterns and TTC/DRAC distributions. In the chord diagrams (top), link thickness indicates the number of cross-type interactions between agent categories, with color denoting time-to-collision (TTC) [49] bands (0–1, 1–2, 2–3, 3–4, 4–5 s; lower indicates higher risk). The bottom panels show marginal distributions of TTC and deceleration rate to avoid crash (DRAC) [50]. HetroD exhibits denser and riskier cross-type interactions, particularly among vehicles, two-wheelers, and pedestrians with a clear shift toward shorter TTC and higher DRAC, highlighting the prevalence of complex VRU interactions in HetroD compared to other datasets.

A. Dataset Construction

HetroD is created from drone videos recorded continuously for approximately 20 minutes per flight at 25 or 30 FPS. Location 5 was recorded from 120 m above ground level (AGL), while all other locations were recorded from 100 m AGL. The videos are processed by an automated pipeline that removes distortion using the drone's intrinsic camera parameters and stabilizes the videos. Afterwards, visible road users are detected and classified using deep neural networks. The subsequent tracking step creates trajectories from the detections, while Kalman filter [51] predictions are used to fill in small gaps. The trajectories are refined afterwards using a series of post-processing steps to create smooth and precise trajectories and to remove false positives. The refined trajectories then undergo two quality assurance cycles. In each cycle, the results of automatic checks and manual inspection are reviewed and fixed. This yields highly accurate trajectory data without tracking errors. For each

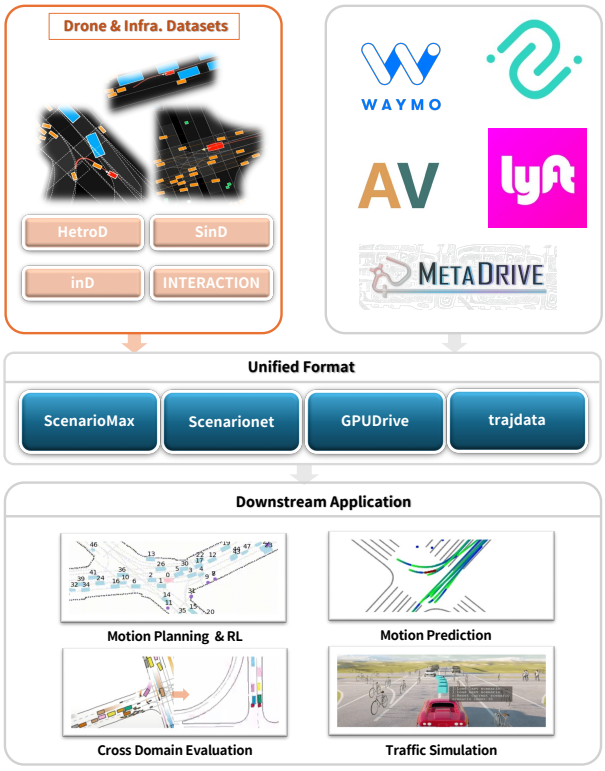

Fig. 5: We develop a unified development toolkit that converts a wide range of traffic scene datasets into standardized, agent-centric data formats [43], [44], [46], [47], enabling seamless comparisons across datasets for forecasting, planning, and simulation.

recording location, HD maps are created from an orthophoto of the respective scene. We provide maps in Lanelet2 [52] and OpenDRIVE [53] formats. The data, orthophoto and HD map are referenced against a common local metric frame $\mathcal{F}_{\text{location}}$ for each location. Traffic light states are derived from ground-based recordings using deep neural networks or from traffic light state diagrams for the location. In case of ground-based recordings, a mask is applied to ensure privacy of road users. Trajectory and traffic light data are synchronized in time to enable matching of traffic light states to the trajectory data.

B. Dataset Properties

To meaningfully quantify traffic complexity in dense, heterogeneous environments, we introduce four normalized metrics that capture spatial, behavioral, and interaction-level diversity (see Table I). These metrics (ranging from *interaction scale* to the presence of *VRUs*) are computed based on full drone-captured datasets. To ensure fair comparison across datasets of varying durations, all metrics are upsampled or downsampled to match the duration of HetroD (17.5 hours), and normalized accordingly. Scale-related metrics are computed using full-dataset coverage. Together, these metrics enable robust cross-platform and cross-dataset comparisons of traffic complexity. Our dataset analysis reveals two fundamental findings:

- **Dataset scale and interaction complexity:** HetroD contains the largest number of unique agent tracks and exhibits the highest levels of *interaction scale* and *heterogeneous interaction scale* among existing drone-view datasets.
- **Cultural and behavioral richness:** While SinD [7] offers a balanced distribution of agent types, HetroD presents a unique setting where *scooters outnumber vehicles*, reflecting traffic patterns not captured in prior datasets (Fig. 3). These agents demonstrate complex behaviors such as weaving, filtering, and informal negotiation, rarely modeled at scale. Risk indicators like TTC [49] and DRAC [50] reveal significantly higher latent conflict rates (Fig. 4).

C. Unified Development Toolkit

To facilitate community adoption, we develop a toolkit that converts HetroD data into formats compatible with popular autonomous driving frameworks (Fig. 5) including ScenarioNet [43], ScenarioMax [46], GPUDrive [47], and trajdata [44]. This enables researchers to leverage HetroD for motion prediction, planning, simulation, and cross-dataset evaluation without extensive preprocessing, making the dataset immediately usable within existing development workflows.

Leveraging this diversity, HetroD fills a long-standing gap in heterogeneous traffic modeling and unlocks two pivotal research axes:

- **High-fidelity heterogeneous traffic simulation:** From full-scene replay to reactive VRUs modeling.
- **VRUs motion prediction and cross-domain generalization:** Enabling out-of-distribution testing on rare, unstructured maneuvers.

IV. Evaluation

We construct a set of challenging per-agent scenarios from HetroD. Specifically, we sample agents exhibiting non-trivial behavior such as long traversals, abrupt heading changes, and dense interactions within multi-agent contexts. These selected agents are used to instantiate per-agent scenarios for evaluation.

A. Motion Forecasting

We evaluate the cross-dataset generalization of two state-of-the-art predictors, *MTR* [54] and *Wayformer* [55], on HetroD under the *UniTraj* [45] protocol. Models are trained on NuScenes, Waymo, SinD, and HetroD, and evaluated on each test set using the Brier-FDE metric [3], which measures probabilistic endpoint accuracy (lower is better). As shown in Table II, performance drops significantly, particularly when transferring from drone-view training data to on-board test sets. To further analyze the factors contributing to model performance, we follow the original MTR implementation, training with anchor files (trajectory endpoint cluster centers that guide the model's search space) generated from each training dataset and evaluate three distinct data-splitting regimes on HetroD, as summarized in Table III: (1) a random split, where training and testing samples are randomly

partitioned; (2) a map-based split, with training and testing conducted on geographically disjoint locations; and (3) a time-based split, where training uses data from busy hours (7-9 AM) and testing uses data from non-busy hours (11 AM-12 PM). Our findings are:

- **In-domain superiority:** Training and testing on the same dataset yields strongest performance but exhibits large cross-domain gaps.
- **Drone-view generalization advantage:** Models trained on drone-view data generalize better to other drone datasets than on-board models generalize to drone-view tests, owing to the near-complete, occlusion-free coverage of drone videos, whereas on-board datasets often yield fragmented or discontinuous tracks for non-ego agents due to limited field of view (FOV) and occlusions. Notably, even when MTR is provided with test-domain anchors, the performance gap remains substantial, suggesting that viewpoint geometry rather than anchor mismatch drives this distribution shift.
- **Map sensitivity:** Map shifts substantially harm in-domain accuracy (166% increase in error for MTR).
- **Temporal robustness:** Time shifts have minimal in-domain impact (5% decrease in error for MTR).
- **Sensitivity to anchor priors:** In MTR, the low in-domain Brier-FDE rises sharply with map changes, indicating strong dependence on the anchor prior.
- **Model trade-offs:** MTR outperforms Wayformer when trained on drone data but is more sensitive to map shifts, while Wayformer remains more consistent across on-board datasets.

These results reveal fundamental limits of current forecasting models in heterogeneous traffic and highlight the value of occlusion-free drone data, which provides complete scene coverage and enables new research opportunities on dense multi-agent interactions.

1) Scenario-Conditioned Evaluation: To analyze failure modes, we stratify evaluation by agent type and cross-type interaction risk. Agent types are *vehicles*, *two-wheelers (scooters & motorcycles)*, and *pedestrians*. Cross-type interaction risk is measured by the minimum time-to-collision (TTC) between the focal agent and agents of different types: *High* (TTC $< 2\,$s), *Moderate* ($2 \leq$ TTC $< 4\,$s), and *Low* (TTC $> 4\,$s). We also examine *scene-level heterogeneity* (low vs. high) using a combined score of local density and Shannon diversity [56] within a 10 m radius of the ego agent, and bin scenes accordingly.

Table IV shows that denser, more heterogeneous scenes and higher cross-type interaction risk yield larger errors (High $>$ Moderate $>$ Low across training sets). Across agent types, two-wheelers are generally the hardest to forecast, while pedestrians exhibit the smallest *absolute* Brier-FDE largely because they travel shorter distances. In terms of *predictability of behavior*, vehicles are typically easier to model due to more regular lane-following dynamics, whereas pedestrians' intentions can still be harder to anticipate despite their short displacements. Irregular two-wheeler maneuvers (e.g., weaving, filtering, abrupt lane changes) further increase difficulty and degrade performance. As heterogeneity and interaction risk increase, errors rise consistently. In high-density heterogeneous scenes, qualitative inspection indicates that MTR trained on Waymo tends to over-project and under-capture subtle multi-agent interactions, highlighting the challenge of forecasting complex, VRU-rich traffic.

TABLE II: Cross-dataset Brier-FDE (↓) for **MTR** [54] and **Wayformer** [55], two state-of-the-art motion prediction models. Rows correspond to training datasets, and columns to testing datasets. The **HetroD** row adopts the *same-map* setting from the ablation study. Models trained on Waymo or NuScenes generally over-predict due to heterogeneous traffic complexity.

		Test			
	Train	**NuScenes**	**Waymo***	**SinD**	**HetroD**
MTR	**NuScenes**	2.95	10.43	5.14	6.76
	Waymo	4.01	2.28	4.26	6.71
	SinD	16.07	26.34	2.06	3.30
	HetroD	21.39	26.49	3.71	0.44
		Test			
	Train	**NuScenes**	**Waymo***	**SinD**	**HetroD**
Wayformer	**NuScenes**	2.99	8.79	5.23	9.37
	Waymo	2.67	2.20	3.53	10.75
	SinD	8.23	13.40	1.96	9.23
	HetroD	19.57	25.28	8.06	0.75

Waymo* uses 30% of its original training data due to resource constraints.

TABLE III: Ablation on HetroD under different training/testing splits. We compare three regimes: *same-map*, *different-map*, and *different-time*. Values report Brier-FDE (↓) on the HetroD test set. Parentheses indicate percentage change relative to the in-domain baseline (*same-map*).

Setting	**MTR**	**Wayformer**
Same-map	**0.44**	**0.75**
Diff-map	1.17 (+166%)	1.53 (+104%)
Diff-time	0.42 (−5%)	0.76 (+1%)

2) Latent Scenario Embeddings: We extract the final decoder query embeddings from the Wayformer model trained on Waymo (chosen for its strong cross-domain performance), pool them per scenario, and project to 2D with t-SNE [57] for cross-dataset visualization. As shown in Fig. 6, randomly sampled scenarios indicate that Waymo and NuScenes share similar latent structure with substantial overlap, whereas HetroD scenarios occupy distinct regions, reflecting complex behaviors and marked differences in heterogeneity, density, and interaction patterns relative to popular on-board datasets.

We further annotate scenarios with their normalized Brier-FDE scores for qualitative analysis. Results show that HetroD scenarios are particularly challenging: the model frequently over-predicts in dense VRU interactions, failing to fully capture rich frontal interactions and nuanced intent among crowded agents. The small physical size of two-wheelers and pedestrians makes anticipating their lateral interactions with the ego agent more difficult. Moreover, unique and complex behaviors (e.g., irregular maneuvers or unexpected lane usage) remain extremely difficult to forecast and generalize.

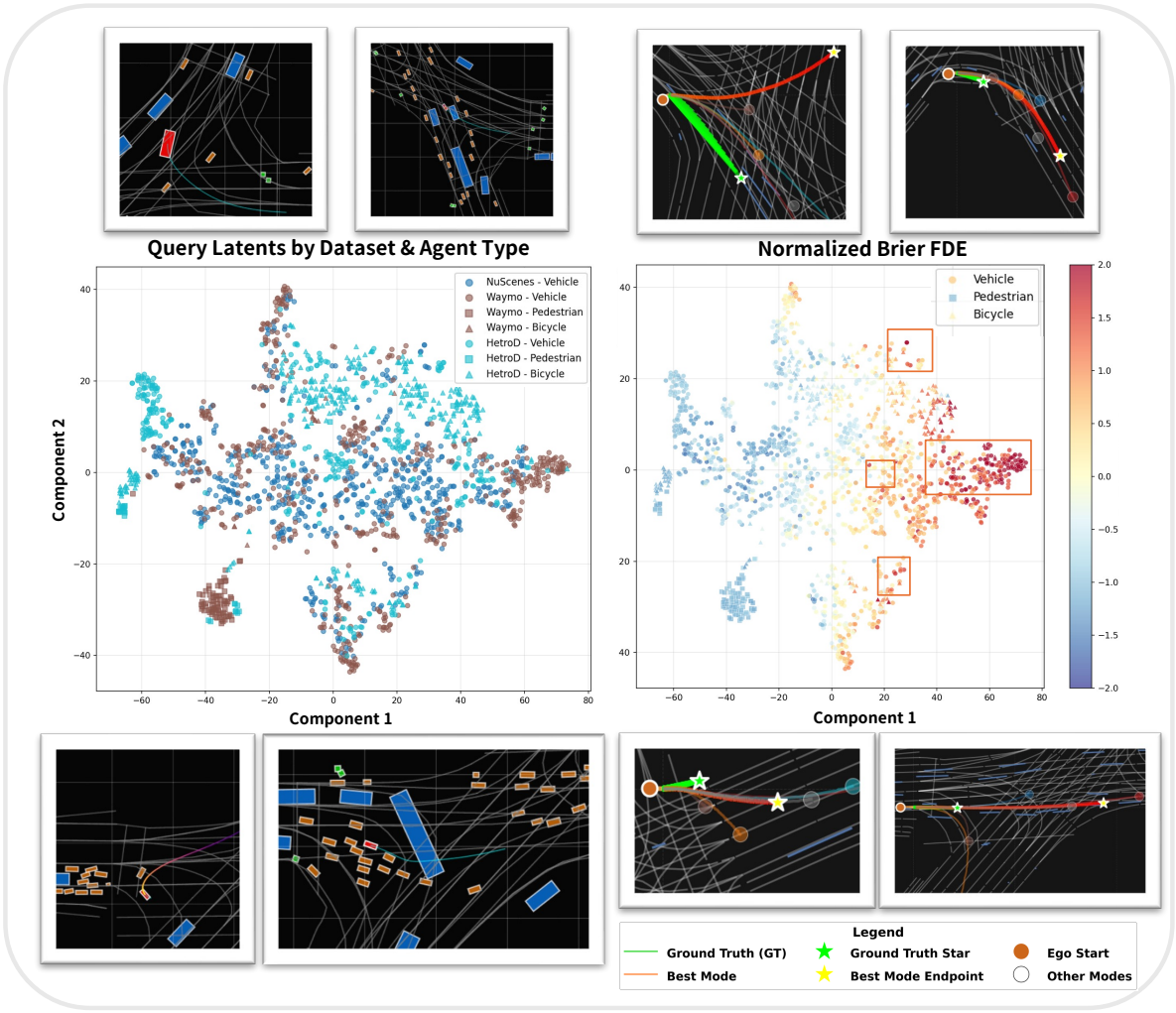

Fig. 6: Latent scenario embeddings (left) and error map (right). *Left:* Points represent scenarios colored by dataset × agent type. *Right:* Same embedding colored by normalized Brier-FDE (↓); warmer indicates higher error. Orange boxes highlight characteristic clusters. *Left insets:* Example HetroD scenarios with complex behaviors. *Right insets:* Model predictions in dense VRU scenarios.

TABLE IV: Scenario-conditioned Brier-FDE (↓) on HetroD test scenarios. MTR models trained on Waymo*, SinD, and HetroD are evaluated on stratified HetroD test cases grouped by agent type, cross-type TTC risk, and scene-level heterogeneity.

Scenario	MTR-Waymo*	MTR-SinD	MTR-HetroD
Agent type			
Vehicle	3.64	2.55	0.83
Two-wheeler	8.69	4.63	1.16
Pedestrian	2.85	1.17	0.26
Cross-type TTC risk			
High risk (TTC<2 s)	8.51	4.76	1.25
Moderate ($2 \leq \min \mathrm{TTC} < 4$ s)	8.48	4.67	1.19
Low (TTC>4 s)	7.87	4.00	0.90
Scene-level heterogeneity			
Low heterogeneity density	5.01	3.08	0.90
High heterogeneity density	8.04	4.25	1.06

Per-block % of filtered HetroD evaluation cases. Agent: Vehicles 39.4%, Two-wheelers 51.5%, Pedestrians 9.1%; TTC: High 26.5%, Moderate 8.7%, Low 64.8%; Scene heterogeneity: Low 66.3%, High 33.7%.

B. Motion Planning

We evaluate planner performance using the V-Max framework [46] under the NuPlan closed-loop, non-reactive scoring protocol. We use 4,420 HetroD vehicle scenarios and compare two rule-based planners: the Intelligent Driver Model (IDM) [58] and PDM-Closed [59], a strong planner from the NuPlan benchmark [4], [11]. To better reflect the challenges of VRU interactions, we augment the evaluation with a VRU-specific collision breakdown that separates *front* and *lateral* contacts involving two-wheelers and pedestrians, cases that standard forward-collision checks often miss in overtakes and unstructured flows.

As shown in Table V, both rule-based planners exhibit clear performance drops on HetroD compared to NuPlan: aggregate scores decrease, comfort deteriorates, and at-fault collisions rise. The collision breakdown in Table VI shows that failures on HetroD frequently involve *lateral* VRU interactions, indicating that planners optimized for structured, vehicle-centric settings struggle to anticipate side interactions. Fig. 7 illustrates representative failure modes, where PDM-Closed struggles with four common heterogeneous traffic scenarios: (1) busy straight road navigation requiring continuous multi-agent reasoning, (2) unprotected left turns with crossing VRUs, (3) narrow road negotiations demanding

TABLE V: Closed-loop, non-reactive planning results. To ensure a fair comparison, we disabled the off-road penalty in the NuPlan aggregate score because, in dense, high-flow, and very narrow-road scenes, these rule-based planners tend to rigidly follow map centerlines rather than adapting, making off-road violations disproportionately likely and obscuring interaction-induced failures.

Dataset	Planner	NuPlan Score ↑	TTC Within Bound ↑	Progress Ratio ↑	Multiple Lane Score ↑	Comfort ↑	At-Fault Collisions ↓
NuPlan	IDM	0.85	0.94	0.92	0.99	0.48	0.016
	PDM-Closed	0.83	0.97	0.91	0.99	0.31	0.006
HetroD	IDM	0.68	0.91	0.81	0.89	0.37	0.074
	PDM-Closed	0.70	0.95	0.78	0.97	0.21	0.040

TABLE VI: VRU collision breakdown on *HetroD*. At-fault collision rate (↓) decomposed into VRU *front* vs. *lateral* contacts for IDM and PDM-Closed.

Planner	At-Fault Collision Rate	VRU Front Collision Rate	VRU Lateral Collision Rate
IDM	0.074	0.008	0.031
PDM-Closed	0.040	0.004	0.022

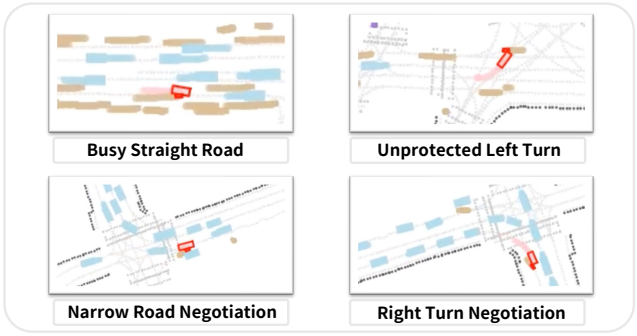

Fig. 7: PDM-Closed planner failures on HetroD. Red boxes indicate ego vehicle collisions. Colors denote traffic participants (blue: vehicles, brown: VRUs).

precise lateral spacing, and (4) right turn negotiations with surrounding VRUs. These qualitative examples reinforce our quantitative findings—rule-based planners lack the interaction awareness needed for dense, heterogeneous traffic, motivating planning objectives that explicitly account for lateral, multi-agent behaviors in unstructured environments.

V. Conclusions

In this work, we introduce HetroD, the first dataset and benchmark that addresses the need for heterogeneous traffic. It comprises 65.4k VRU-rich trajectories, HD maps, and traffic signal states. Our evaluation reveals fundamental limitations in current autonomous driving approaches when deployed in heterogeneous traffic scenarios. As traffic heterogeneity increases, state-of-the-art forecasting and planning methods exhibit significant performance degradation. We identify two critical failure modes: two-wheelers emerge as the most challenging agent type for accurate prediction, while rule-based planners demonstrate disproportionately high lateral VRU collision rates despite maintaining strong lane-keeping performance. Therefore, future work will focus on utilizing HetroD for VRU modeling and simulation, ultimately increasing the safety and compatibility of autonomous driving in heterogeneous traffic environments.

Acknowledgment

This work was supported in part by the National Science and Technology Council under Grants 113-2628-E-A49-022 and 114-2628-E-A49-007, the H&J Global Chair, the Ministry of Education, and the Yushan Fellow Program Administrative Support Grant. W.J. Chang was also supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2146752. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] 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.

[2] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in Perception for Autonomous Driving: Waymo Open Dataset," in *CVPR*, 2020.

[3] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, D. Ramanan, P. Carr, and J. Hays, "Argoverse 2: Next Generation Datasets for Self-driving Perception and Forecasting," in *NeurIPS*, 2021.

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

[5] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kümmerle, H. Königshof, C. Stiller, A. de La Fortelle, and M. Tomizuka, "INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps," *arXiv:1910.03088 [cs, eess]*, 2019.

[6] J. Bock, R. Krajewski, T. Moers, S. Runde, L. Vater, and L. Eckstein, "The inD Dataset: A Drone Dataset of Naturalistic Road User Trajectories at German Intersections," in *IV*, 2020.

[7] Y. Xu, W. Shao, J. Li, K. Yang, W. Wang, H. Huang, C. Lv, and H. Wang, "SIND: A Drone Dataset at Signalized Intersection in China," in *ITSC*, 2022.

[8] M. Liu, E. Yurtsever, J. Fossaert, X. Zhou, W. Zimmer, Y. Cui, B. L. Zagar, and A. C. Knoll, "A Survey on Autonomous Driving Datasets: Statistics, Annotation Quality, and a Future Outlook," *IEEE Trans. Intell. Veh.*, 2024.

[9] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "TraPHic: Trajectory Prediction in Dense and Heterogeneous Traffic Using Weighted Interactions," in *CVPR*, 2019.

[10] S. Malik, M. A. Khan, Aadam, H. El-Sayed, F. Iqbal, J. Khan, and O. Ullah, "CARLA+: An Evolution of the CARLA Simulator for Complex Environment Using a Probabilistic Graphical Model," *Drones*, 2023.

[11] C. Gulino, J. Fu, W. Luo, G. Tucker, E. Bronstein, Y. Lu, J. Harb, X. Pan, Y. Wang, X. Chen, J. D. Co-Reyes, R. Agarwal, R. Roelofs, Y. Lu, N. Montali, P. Mougin, Z. Yang, B. White, A. Faust, R. McAllister, D. Anguelov, and B. Sapp, "Waymax: An Accelerated, Data-Driven Simulator for Large-Scale Autonomous Driving Research," in *NeurIPS*, 2023.

[12] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *CVPR*, 2012.

[13] A. Patil, S. Malla, H. Gang, and Y.-T. Chen, "The H3D Dataset for Full-Surround 3D Multi-Object Detection and Tracking in Crowded Urban Scenes," in *ICRA*, 2019.

[14] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, L. Chen, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska, "One Thousand and One Hours: Self-driving Motion Prediction Dataset," in *CoRL*, 2020.

[15] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning," *CVPR*, 2018.

[16] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang, Y. Wang, and D. Yang, "PandaSet: Advanced Sensor Suite Dataset for Autonomous Driving," in *ITSC*, 2021.

[17] L. Gressenbuch, K. Esterle, T. Kessler, and M. Althoff, "MONA: The Munich Motion Dataset of Natural Driving," in *ITSC*, 2022.

[18] Y. Zhang, C. Wang, R. Yu, L. Wang, W. Quan, Y. Gao, and P. Li, "The AD4CHE Dataset and Its Application in Typical Congestion Scenarios of Traffic Jam Pilot Systems," *IEEE Trans. Intell. Veh.*, 2023.

[19] K. Chen, R. Ge, H. Qiu, R. Ai-Rfou, C. R. Qi, X. Zhou, Z. Yang, S. Ettinger, P. Sun, Z. Leng, M. Baniodeh, I. Bogun, W. Wang, M. Tan, and D. Anguelov, "WOMD-LiDAR: Raw Sensor Dataset Benchmark for Motion Forecasting," in *CORR*, 2023.

[20] R. Chandra, X. Wang, M. Mahajan, R. Kala, R. Palugulla, C. Naidu, A. Jain, and D. Manocha, "METEOR:A Dense, Heterogeneous, and Unstructured Traffic Dataset With Rare Behaviors," in *ICRA*, 2023.

[21] U. S. D. of Transportation Federal Highway Administration, "Next Generation Simulation (NGSIM) Vehicle Trajectories and Supporting Data," 2016.

[22] D. Yang, L. Li, K. A. Redmill, and Ü. Özgüner, "Top-view Trajectories: A Pedestrian Dataset of Vehicle-Crowd Interaction from Controlled Experiments and Crowded Campus," in *IV*, 2019.

[23] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, and Z. Nie, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *CVPR*, 2022.

[24] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An Open Benchmark Dataset and Fusion Pipeline for Perception with Vehicle-to-Vehicle Communication," in *ICRA*, 2022.

[25] H. Yu, W. Yang, H. Ruan, Z. Yang, Y. Tang, X. Gao, X. Hao, Y. Shi, Y. Pan, N. Sun, J. Song, J. Yuan, P. Luo, and Z. Nie, "V2X-Seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting," in *CVPR*, 2023.

[26] G. Kueppers, J.-P. Busch, L. Reiher, and L. Eckstein, "V2AIX: A Multi-Modal Real-World Dataset of ETSI ITS V2X Messages in Public Road Traffic," 2024.

[27] H. Xiang, Z. Zheng, X. Xia, R. Xu, L. Gao, Z. Zhou, X. Han, X. Ji, M. Li, Z. Meng *et al.*, "V2X-Real: a Large-Scale Dataset for Vehicle-to-Everything Cooperative Perception," *arXiv preprint arXiv:2403.16034*, 2024.

[28] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, and A. Knoll, "TUMTraf V2X Cooperative Perception Dataset," *arXiv preprint arXiv:2403.01316*, 2024.

[29] W. Zimmer, R. Greer, D. Lehmberg, M. Pavel, H. Caesar, X. Zhou, A. Ghita, M. Trivedi, R. Song, H. Cao *et al.*, "Towards Vision Zero: The Accid3nD Dataset," *arXiv preprint arXiv:2503.12095*, 2025.

[30] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes," in *ECCV*, 2016.

[31] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems," in *ITSC*, 2018.

[32] R. Krajewski, T. Moers, J. Bock, L. Vater, and L. Eckstein, "The rounD Dataset: A Drone Dataset of Road User Trajectories at Roundabouts in Germany," in *ITSC*, 2020.

[33] A. Breuer, J. Termöhlen, S. Homoceanu, and T. Fingscheidt, "openDD: A Large-Scale Roundabout Drone Dataset," in *ITSC*, 2020.

[34] P. Spannaus, P. Zechel, and K. Lenz, "AUTOMATUM DATA: Drone-based highway dataset for the development and validation of automated driving software for research and commercial applications," in *IV*, 2021.

[35] T. Moers, L. Vater, R. Krajewski, J. Bock, A. Zlocki, and L. Eckstein, "The exiD Dataset: A Real-World Trajectory Dataset of Highly Interactive Highway Scenarios in Germany," in *IV*, 2022.

[36] P. Tkachenko, N. Certad, G. Singer, C. Olaverri-Monreal, and L. del Re, "The JKU DORA Traffic Dataset," *IEEE Access*, 2022.

[37] A. Mukbil, Y. M. Yousif, S. Hossain, and J. P. Müller, "CTV-Dataset: A Shared Space Drone Dataset for Cyclist-Road User Interaction Derived from Campus Experiments," in *ITSC*, 2023.

[38] O. Zheng, M. Abdel-Aty, L. Yue, A. Abdelraouf, Z. Wang, and N. Mahmoud, "CitySim: A Drone-Based Vehicle Trajectory Dataset for Safety-Oriented Research and Digital Twins," *Transportation Research Record*, 2024.

[39] J. Meier, L. Scalerandi, O. Dhaouadi, J. Kaiser, A. Nikita, and D. Cremers, "CARLA Drone: Monocular 3D Object Detection from a Different Perspective," in *GCPR*, 2024.

[40] O. Dhaouadi, J. Meier, L. Wahl, J. Kaiser, L. Scalerandi, N. Wandelburg, Z. Zhuo, N. Berinpanathan, H. Banzhaf, and D. Cremers, "Highly Accurate and Diverse Traffic Data: The DeepScenario Open 3D Dataset," in *IV*, 2025.

[41] J. Andle, N. Soucy, S. Socolow, and S. Yasaei Sekeh, "The Stanford Drone Dataset Is More Complex Than We Think: An Analysis of Key Characteristics," *IEEE Trans. Intell. Veh.*, 2022.

[42] Q. Li, Z. Peng, L. Feng, Q. Zhang, Z. Xue, and B. Zhou, "Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning," *TPAMI*, 2022.

[43] 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," *NeurIPS*, 2023.

[44] B. Ivanovic, G. Song, I. Gilitschenski, and M. Pavone, "trajdata: A Unified Interface to Multiple Human Trajectory Datasets," in *NeurIPS*, 2023.

[45] L. Feng, M. Bahari, K. M. B. Amor, É. Zablocki, M. Cord, and A. Alahi, "UniTraj: A Unified Framework for Scalable Vehicle Trajectory Prediction," *arXiv preprint arXiv:2403.15098*, 2024.

[46] V. Charraut, T. Tournaire, W. Doulazmi, and T. Buhet, "V-Max: A Reinforcement Learning Framework for Autonomous Driving," 2025.

[47] S. Kazemkhani, A. Pandya, D. Cornelisse, B. Shacklett, and E. Vinitsky, "GPUDrive: Data-driven, multi-agent driving simulation at 1 million FPS," in *ICLR*, 2025.

[48] T. Westny, B. Olofsson, and E. Frisk, "Toward Unified Practices in Trajectory Prediction Research on Bird's-Eye-View Datasets," in *IV*, 2025.

[49] J. R. Ward, G. Agamennoni, S. Worrall, A. Bender, and E. Nebot, "Extending Time to Collision for probabilistic reasoning in general traffic scenarios," *Transportation Research Part C: Emerging Technologies*, 2015.

[50] J. Shen and G. Yang, "Crash Risk Assessment for Heterogeneity Traffic and Different Vehicle-Following Patterns Using Microscopic Traffic Flow Data," *Sustainability*, 2020.

[51] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of basic Engineering*, 1960.

[52] F. Poggenhans, J.-H. Pauls, J. Janosovits, S. Orf, M. Naumann, F. Kuhnt, and M. Mayr, "Lanelet2: A high-definition map framework for the future of automated driving," in *ITSC*, 2018.

[53] OpenDRIVE Initiative, "OpenDRIVE: Open Dynamic Road Information for Vehicle Environment," 2006.

[54] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," *NeurIPS*, 2022.

[55] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion Forecasting via Simple & Efficient Attention Networks," *ICRA*, 2022.

[56] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. University of Illinois Press, 1949.

[57] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, 2008.

[58] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical Review E*, 2000.

[59] D. Dauner, M. Hallgarten, A. Geiger, and K. Chitta, "Parting with Misconceptions about Learning-based Vehicle Motion Planning," in *CoRL*, 2023.