StyleDrive: Towards Driving-Style Aware Benchmarking of End-To-End Autonomous Driving

Ruiyang Hao[1,2], **Bowen Jing**[3], **Haibao Yu**[1,4], **Zaiqing Nie**[1*]
[1]AIR, Tsinghua University
[2]King's College London
[3]The University of Manchester
[4]The University of Hong Kong
zaiqing@air.tsinghua.edu.cn

Abstract

Personalization, while extensively studied in conventional autonomous driving pipelines, has been largely overlooked in the context of end-to-end autonomous driving (E2EAD), despite its critical role in fostering user trust, safety perception, and real-world adoption. A primary bottleneck is the absence of large-scale real-world datasets that systematically capture driving preferences, severely limiting the development and evaluation of personalized E2EAD models. In this work, we introduce the first large-scale real-world dataset explicitly curated for personalized E2EAD, integrating comprehensive scene topology with rich dynamic context derived from agent dynamics and semantics inferred via a fine-tuned vision-language model (VLM). We propose a hybrid annotation pipeline that combines behavioral analysis, rule-and-distribution-based heuristics, and subjective semantic modeling guided by VLM reasoning, with final refinement through human-in-the-loop verification. Building upon this dataset, we introduce the first standardized benchmark for systematically evaluating personalized E2EAD models. Empirical evaluations on state-of-the-art architectures demonstrate that incorporating personalized driving preferences significantly improves behavioral alignment with human demonstrations.

1 Introduction

As autonomous driving (AD) matures, personalization becomes key to real-world products. Tailoring vehicle behavior to individual user preferences is essential for enhancing user experience, building trust, and fostering long-term adoption (Hasenjäger and Wersing 2017). Traditional modular systems have extensively explored personalized strategies, enabling adaptations to driving preferences. However, these methods often rely on isolated, scenario-specific adaptations (Tian et al. 2022) or unrealistic human-in-the-loop (HITL) simulation (Ke et al. 2024), which limit to generalize to real-world environments. Furthermore, the fragmented nature of modular systems (Zhu et al. 2018; Cui et al. 2024; Kou et al. 2025) hinders scalability to massive real-world data. Due to these limitations, personalization remains largely underexplored in end-to-end autonomous driving (E2EAD), where perception, planning, and control are integrated in an unified architecture (Codevilla et al. 2018). This gap presents a significant barrier to realizing human-centric AD at scale.

This gap is particularly pressing. As E2EAD systems become increasingly capable and are deployed across diverse, open-world driving environments, aligning vehicle behavior with user preferences becomes essential for enhancing comfort, perceived safety, and long-term user acceptance (Speidel et al. 2019; Aledhari et al. 2023). Yet integrating personalization into the end-to-end paradigm introduces unique challenges. Unlike modular systems operating within narrowly defined scenarios, E2EAD requires preference generalization across a broad spectrum of complex, real-world scenarios. Meeting this demand necessitates large-scale datasets that not only cover diverse traffic conditions but also provide well-crafted driving style annotations.

To fill this critical gap, we introduce **the first large-scale real-world dataset purposefully designed for personalized E2EAD research**. Our dataset captures both objective behavioral patterns and subjective driving preferences across a wide range of driving scenarios. Static environmental features are firstly extracted from real-world road topologies, while dynamic context cues are inferred using a fine-tuned visual language model (VLM), enabling rich semantic scene understanding. Based on these features, we generate objective annotations through behavior distribution analysis and rule&distribution-based heuristics. To incorporate subjectivity, we further employ the VLM to label preferences considering static and dynamic scene semantics. The pipeline ends with a label fusion mechanism to ensure quality and consistency. Building upon this dataset, we establish **the first benchmark for evaluating personalized E2EAD models**. We conduct extensive experiments on multiple state-of-the-art architectures, with and without preference conditioning, and show that incorporating personalization significantly improves alignment with human-like driving behavior.

The motivation and overview of our work are illustrated in Fig. 1, aiming to bridge the gap between the growing momentum in E2EAD and the longstanding need for personalization in AD. In summary, our key contributions are:

- **A novel large-scale real-world dataset** for personalized end-to-end autonomous driving, annotated with both objective behaviors and subjective driving style preferences across a wide range of traffic scenarios.
- **A multi-stage annotation pipeline** that integrates be-

*Corresponding author.

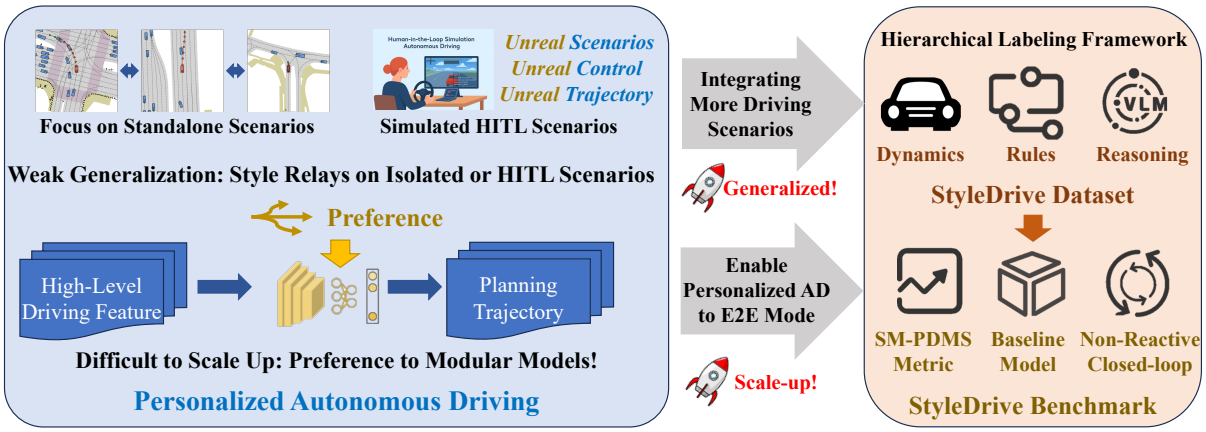

Figure 1: Motivation and Overview of StyleDrive.

havior feature analysis, rule&distribution-based heuristics, VLM reasoning, and human-in-the-loop validation to generate high-quality well-crafted preference labels.

- **The first benchmark for personalized E2EAD**, which enables standardized, quantitative comparison of preference-conditioned behavior across four different SOTA model architectures.
- **Comprehensive empirical results** demonstrating that incorporating user driving preferences significantly improves alignment with human-like behavior, underscoring the value of personalization in E2EAD.

2 Related Work

Personalized Autonomous Driving Personalization has long been recognized as a key factor to enhance user comfort, trust, and acceptance of AD (Liao et al. 2025b). In conventional modular pipelines, personalized strategies have been extensively explored, where the system can adapt to users' preferences in stand-alone scenarios, such as car following, ramp merge, and lane change (Zhao et al. 2022; Li et al. 2023; Liao et al. 2023). Other early attempts employ unrealistic human-in-the-loop (HITL) simulation (Ke et al. 2024). However, these approaches often suffer from poor generalization in dynamic, real-world environments.

Recent advances in large language models (LLMs) have introduced new possibilities for personalization in AD, enabling the encoding of user intent and preferences to influence downstream decision-making (Cui et al. 2024; Xu et al. 2024; Kou et al. 2025). Despite their potential, these methods remain constrained by modular frameworks and often depend on handcrafted features. MAVERIC (Schrum et al. 2024) represents a notable early attempt to enable end-to-end personalized driving. Nevertheless, in the absence of standardized datasets and benchmarks, these methods face challenges in reproducibility and scalability. Despite these advancements, existing work seldom addresses how to merge user preferences in an E2EAD manner. A critical bottleneck is the lack of large-scale, real-world datasets with well-crafted annotations of personalized driving styles.

End-to-End Autonomous Driving End-to-end autonomous driving represents a promising paradigm that maps raw sensor inputs directly to driving actions or planned trajectories. Early work (Codevilla et al. 2018) demonstrated the feasibility of E2EAD. Over the past decade, research has advanced toward more sophisticated architectures incorporating temporal reasoning, multimodal fusion, and Transformer-based methods (Prakash, Chitta, and Geiger 2021; Jia et al. 2023; Shao et al. 2023).

Despite this progress, most E2EAD models are optimized for average-level objectives, lacking the capacity to adapt to user-specific preferences. While recent architectures improve generalization and interpretability (Li et al. 2025a; Zheng et al. 2024; Jia et al. 2025), they remain limited in capturing individualized driving styles. Moreover, the absence of dedicated datasets and standardized benchmarks for evaluating personalization continues to hinder systematic advancement in this area. To address these limitations, we propose a framework for learning style-conditioned E2EAD models and establish a benchmark to evaluate their alignment with human driving behavior.

Benchmarking Autonomous Driving Benchmarking is fundamental to the development and evaluation of AD models. Typical AD datasets and benchmarks are summarized in Tab. 1. Early E2EAD benchmarks (Caesar et al. 2020; Peng et al. 2023) adopt non-interactive open-loop settings with real-world data, but fail to capture behavioral feedback. Recent works have introduced closed-loop benchmarks in simulation, such as Longest6 (Chitta et al. 2022), MetaDrive (Li et al. 2022), CARLA Series (Contributors 2024), which allow online policy evaluation but rely entirely on simulated scenarios. Semi-simulated platforms like NAVSIM (Dauner et al. 2024) strike a balance by simulating behavior feedback based on real-world scenes. However, existing benchmarks remain focused on task performance and ignore user preferences, cannot evaluate personalized driving behavior.

Efforts to benchmark personalized AD have led to two main dataset categories: human-in-the-loop (HITL) simulations and real-world driving logs. 1) HITL datasets are typically collected in controlled scenarios such as ramp merg-

Dataset	Reality	Scenarios	CL/OL	E2E	Style
nuScenes (Caesar et al. 2020)	Real	City	OL	✓	✗
OpenScene (Peng et al. 2023)	Real	City&Rural	OL	✓	✗
Longest6 (Chitta et al. 2022)	Sim	City	CL	✓	✗
CARLA (Contributors 2024)	Sim	City	CL	✓	✗
MetaDrive (Li et al. 2022)	Sim	City	CL	✓	✗
Bench2Drive (Jia et al. 2024)	Sim	City	CL	✓	✗
NAVSIM (Dauner et al. 2024)	Real	City&Rural	Semi-CL	✓	✗
HITL-RampMerging (Li et al. 2023)	HITL	Ramp-Merge	OL	✗	✓
HITL-CarFollowing (Zhao et al. 2022)	HITL	Car-Following	OL	✗	✓
HITL-LaneChange (Liao et al. 2023)	HITL	Lane-Change	OL	✗	✓
HITL-MultiScene (Ke et al. 2024)	HITL	City	OL	✗	✓
UAH (Romera, Bergasa, and Arroyo 2016)	Real	City&highway	OL	✗	✓
Brain4Cars (Jain et al. 2016)	Real	Lane-Change&Merge	OL	✗	✓
PDB (Wei et al. 2025)	Real	City	OL	✗	✓
StyleDrive (Ours)	Real	City&Rural	Semi-CL	✓	✓

Table 1: Comparison of driving datasets by data source, scenario coverage, evaluation protocol, end-to-end (E2E) support, and driving style annotations. **StyleDrive** is the only dataset that combines real-world data, diverse scenarios, semi-closed-loop (Semi-CL) evaluation, E2E learning capability, and structured driving style labels for personalized autonomous driving.
Abbreviations: OL = Open-Loop, CL = Closed-Loop, Semi-CL = Semi-Closed-Loop, HITL = Human-in-the-Loop, E2E = End-to-End driving model support, Style = With Style or Preference Annotation

ing (Li et al. 2023), car following (Zhao et al. 2022), lane changing (Liao et al. 2023), and multiple scenarios (Ke et al. 2024). While these setups enable preference modeling, they often rely on simplified simulation engines and curated environments, introducing a domain gap between real-world conditions and unrealistic scenarios. 2) Real-world datasets such as UAH-DriveSet (Romera, Bergasa, and Arroyo 2016), Brain4Cars (Jain et al. 2016), and PDB (Wei et al. 2025) provide behavioral annotations related to diving intent. These datasets have contributed substantially to understanding driver behavior and intent modeling. However, they are typically constrained to narrow driving contexts, such as highways or intersections. More importantly, they are not structured for end-to-end learning: most only provide high-level behavior labels without continuous perception-to-control supervision, and none include standardized evaluation protocols for comparing personalized driving policies.

To bridge the gap between E2EAD and personalized AD, we introduce the first benchmark tailored for personalized E2EAD. Built on real-world data, our benchmark supports both supervised learning and semi-closed-loop evaluation of preference-conditioned policies, enabling reproducible evaluation of personalized E2E across diverse scenarios.

3 StyleDrive Dataset

The StyleDrive dataset is sampled and constructed from the large-scale AD dataset OpenScene (Peng et al. 2023), and comprises nearly 30k driving scenarios annotated with style preferences. OpenScene features trajectories collected from over 16 drivers across diverse urban and suburban environments in both Singapore and the USA. This diversity offers a rich foundation for labeling user-specific tendencies in various scenarios (also illustrated in Fig. 4). Beyond the original OpenScene data, we introduce a unified framework for modeling and annotating personalized driving preferences.

Framework for Modeling and Annotation of Driving Preference

Framework Overview. To enable reliable and interpretable driving style analysis, we construct a hierarchical modeling and annotation framework, as shown in Fig. 2. We extract static environmental features from real-world road topology and dynamic motion features from driving logs, and infer dynamic contextual cues using a fine-tuned visual language model (VLM), enabling consistent and fine-grained scenario construction. From these constructed scenario features, we derive objective preference annotations via behavioral distribution analysis and rule-based heuristics. To capture the inherent subjectivity of driving style, we further leverage the VLM to generate subjective annotations by jointly modeling scene semantics and driver behavior. Final high-quality labels are obtained through a human-in-the-loop verification process that consolidates both perspectives.

Topology-Driven Static Scenario Taxonomy To facilitate structured analysis of driving behavior, we first categorize each driving clip into a coarse-grained traffic scenario type based on the topology of high-definition (HD) maps. This categorization provides a physically grounded and semantically interpretable foundation for subsequent scenario modeling. Specifically, we begin by enriching the topological structures based on the original nuPlan HD maps. We then align each clip data with detailed map elements, such as lane geometries, stop lines, crosswalks, and intersection layouts, to assign it to one of eleven primary traffic scenario types. Each type captures the static constraints and navigational features of the road environment, serving as a structural prior to interpreting the downstream driving behavior. The complete type list is shown in the left part of Fig. 3.

Dynamic Semantic Refinement via Vision-Language Model To enrich the traffic scenario classification with dy-

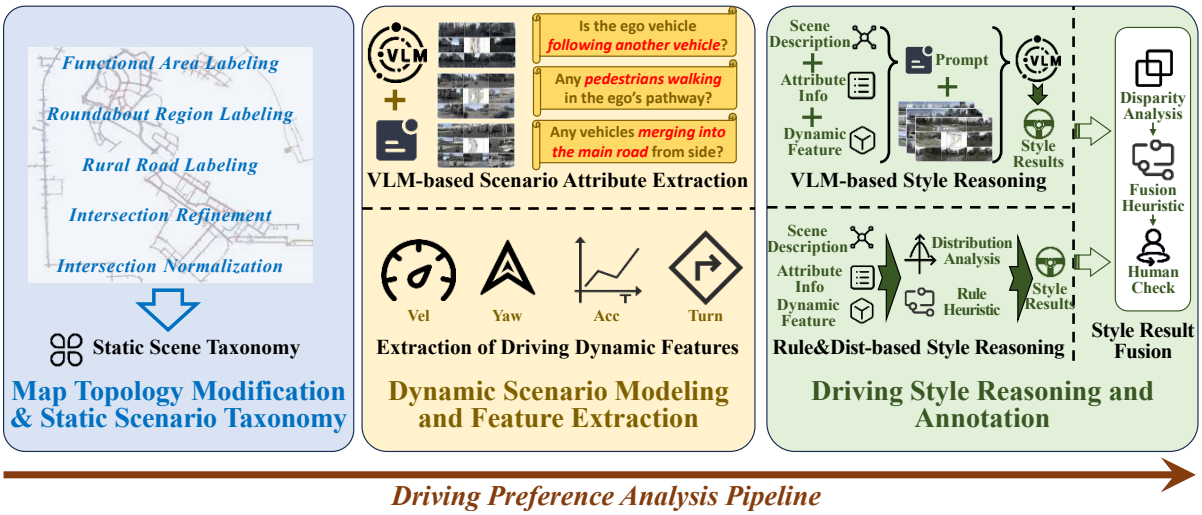

Figure 2: Framework for Modeling and Annotation of Driving Preference.

namic, context-sensitive information, we apply a fine-tuned vision-language model (VLM) to extract high-level semantic cues from driving videos. Since generic VLMs lack task-specific perception and reasoning capabilities for AD, we adapt the Video-LLaMA3 model (Zhang et al. 2025) via lightweight fine-tuning using the LingoQA dataset (Marcu et al. 2024) - a multimodal dataset for reasoning about road semantics.

For each traffic scenario type, we craft targeted prompts related to key driving events. These include questions such as whether a lead car is present in the same lane, whether the ego is merging, or whether pedestrians are visible in crosswalk zones. The model responses are then parsed into structured semantic attributes, such as the presence of lead vehicles, merging behavior, pedestrian involvement, and turning intent. This dynamic semantic layer augments the static topology with interaction-aware, temporally grounded cues, thereby enhancing downstream preference modeling via both rule&distribution-based and VLM-based reasoning.

Objective Preference Annotation via Rule&Distribution-based Heuristics Building on the enriched scene context, we generate objective driving style annotations using a set of interpretable, physically grounded heuristics informed by motion dynamics and semantic priors. We extract a set of behaviorally relevant physical ego motion features, such as speed, acceleration, yaw rate variation, and proximity to surrounding agents. Based on these ego motion features, structured scene semantics and static scenario type, we analyze the feature distribution and accordingly define scenario-specific rules calibrated from aggregated driving statistics. For example:

- Low speed and wide safety margins at intersections correspond to conservative tendencies;
- Sudden lane changes with low rear headway indicate aggressive behavior;
- Moderate-speed following and stable lane keeping fall into the normal category.

Grounded in population-level behavior distributions and further expert validation, these heuristics offer an explicit and robust foundation for objective style labeling.

Subjective Preference Annotation via VLM-Based Reasoning To complement rule&distribution-based methods, we leverage the contextual reasoning ability of our fine-tuned Video-LLama3 to generate subjective annotations that reflect human-like interpretations of driving style. Given the driving video, the structured scene semantics and corresponding extracted ego motion features, the model is prompted to answer behavioral questions such as:

- "Does the ego vehicle appear cautious given the movement of pedestrians?"
- "Is the vehicle merging assertively or yielding?"

These responses capture high-level intent and interaction cues that extend beyond the expressiveness of fixed heuristics. The multimodal attention of the VLM model enables it to model nuanced behavioral patterns, especially in borderline or semantically complex scenarios.

Multi-Source Fusion and Human-in-the-Loop Verification Before finalizing the annotation protocol, we conduct a human-in-the-loop analysis to assess consistency and divergence between rule&distribution-based and VLM-based labels. Manual inspection reveals the following patterns:

- Aggressive driving is often consistently detected by both sources. Even when disagreement occurs, each source frequently captures valid signals that the other misses. This complementary behavior suggests that aggressive tendencies are well-suited to permissive merging strategies.
- In contrast, conservative and normal styles are more easily conflated by VLM, particularly in low-speed or ambiguous interactions. This observation motivates the adoption of stricter criteria for conservative assignments.

These observations motivate a risk-aware fusion strategy: 1) If either rule-based or VLM-based reasoning labels a clip as aggressive, we annotate it as aggressive; 2) If both sources

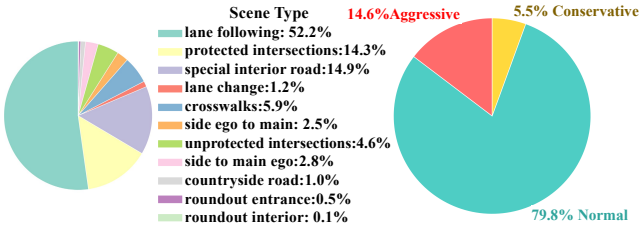

Figure 3: Dataset Statistics and Distribution Analysis.

agree that the clip reflects conservative behavior, we label it as conservative; 3) In other cases, the driving style is marked as normal. This strategy emphasizes consistency by being permissive for potential aggressive behavior while remaining strict about conservative assignments. It also ensures that the final label reflects both interpretability and robustness. Besides, final annotations are further verified in edge cases through targeted human review, ensuring label reliability.

Dataset Stats, Style Distribution and Visualization

To provide a global overview of the StyleDrive dataset, we present key statistics on both scenario composition and annotated driving preference. As shown in Fig. 3, we visualize the distribution across static/dynamic scene types and behavioral styles using two complementary pie charts.

Scene Composition. The left chart shows the distribution of traffic scenario types. Lane-follow and intersection scenarios dominate the dataset, reflecting their prevalence in real-world driving. Context-rich types such as merging, lane changes, and pedestrian interactions contribute to diversity.

Driving Style Distribution. The right chart illustrates the annotation outcomes for driving preferences. Normal behavior forms the majority, while aggressive and conservative styles are less common but sufficiently represented. This distribution mirrors typical driving patterns and supports balanced learning, especially for edge-case recognition.

Driving Style Visualization. Fig. 4 highlights the distribution of different driving styles. Notably, even under similar driving conditions, drivers exhibit substantial variations in their behavioral preferences.

4 StyleDrive Benchmark

To advance the development and evaluation of personalized E2EAD, we introduce the StyleDrive Benchmark, a non-reactive simulation-based evaluation framework for assessing driving preferences and performance in realistic traffic scenarios. This benchmark evaluates whether autonomous agents can generate behavior that aligns with target driving styles while ensuring safety and social compliance. Built upon the rich scenarios and structured annotations of the StyleDrive dataset, the benchmark defines a standardized testbed consisting of four components: simulation environment, the proposed SM-PDMS metric, baseline models, and benchmark results with performance analysis.

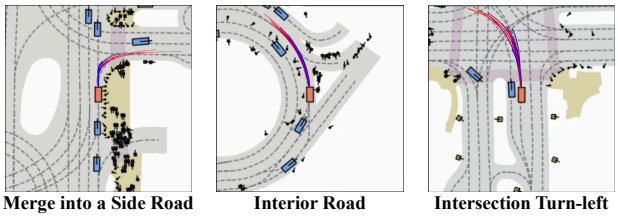

Figure 4: Visualization of driving style distribution in three typical scenarios. Each case is drawn from similar local scenes without pedestrians or leading cars, ensuring style differences arise primarily from drivers' own behavioral preferences. Red trajectories denote aggressive and blue ones denote conservative.

Simulation Environment. Our benchmark is built on a lightweight simulator NavSim (Dauner et al. 2024), which emphasizes non-reactive behavior simulation and grounded in real-world traffic scenarios. In our adaptation, the scenarios of StyleDrive are replayed, and style-conditioned E2EAD models learn policies in this playground.

Evaluation Metric: Style-Modulated PDMS

To evaluate whether a model ensures not only safety and feasibility but also alignment with a desired driving style, we propose the Style-Modulated Predictive Driver Model Score (SM-PDMS), an extension of the Predictive Driver Model Score (PDMS) introduced in the NavSim (Dauner et al. 2024). SM-PDMS enhances the original PDMS by incorporating a behavioral alignment component that quantifies the degree to which a policy conforms to specified style preferences. Sub-metrics related to traffic safety and road adherence are retained without modification, as they are considered invariant across driving styles. In contrast, driving style is primarily manifested through motion dynamics, such as following distance, speed, and angular velocity, which in turn affect style-sensitive sub-metrics including comfort (Comf.), ego progress (EP), and time-to-collision (TTC).

Applying the same metric configuration for style-sensitive sub-metrics across different styles is suboptimal for evaluating style-conditioned policies, as users with differing preferences inherently define "better" driving behavior in inconsistent ways. To remedy this, we modulate the style-relevant sub-metrics with annotated styles. Specifically:

- "EP": Ego progress objectives are calibrated to align with target cruising speeds and preferred following distances;
- "Comf.": Comfort-related thresholds are modulated to reflect varying sensitivities to physical disturbances;
- "TTC": TTC acceptability ranges are adjusted to capture heterogeneous risk tolerances across styles.

Further details on these modifications are provided in Supplementary. They ensure that SM-PDMS offers a systematic evaluation of both safety and style alignment.

Baseline Methods

To facilitate standardized comparison, we implement four representative style-conditioned baselines covering varied model complexities and paradigms, detailed as follows.

Models	NC ↑	DAC ↑	Style-Modulated Submetrics			SM-PDMS ↑
			TTC ↑	Comf. ↑	EP ↑	
AD-MLP (Zhai et al. 2023)	92.63	77.68	83.83	99.75	78.01	63.72
TransFuser (Chitta et al. 2022)	96.74	88.43	91.08	99.65	84.39	78.12
WoTE (Li et al. 2025b)	97.29	92.39	92.53	99.13	76.31	79.56
DiffusionDrive (Liao et al. 2025a)	96.66	91.45	90.63	99.73	80.39	79.33
AD-MLP-Style	92.38	73.23	83.14	**99.90**	78.55	60.02
TransFuser-Style	97.23	90.36	92.61	99.73	**84.95**	81.09
WoTE-Style	**97.58**	**93.44**	**93.70**	99.26	77.38	**81.38**
DiffusionDrive-Style	**97.81**	**93.45**	**92.81**	**99.85**	**84.84**	**84.10**
- DiffusionDrive-Style-A	97.38	93.20	92.01	99.62	84.01	83.04
- DiffusionDrive-Style-N	97.66	93.32	92.16	99.83	84.21	83.52
- DiffusionDrive-Style-C	98.23	93.59	94.98	99.87	81.36	83.90

Table 2: StyleDrive Benchmark Main Results. Style conditioning improves behavioral alignment, with higher SM-PDMS scores across model families. The ablation results (bottom) further confirm the effectiveness and learnability of style conditioning.

- **AD-MLP-Style**: Extends the AD-MLP baseline (Zhai et al. 2023) with a one-hot driving style vector concatenated to ego states. The combined input is passed through MLPs for style-aware trajectory regression.
- **TransFuser-Style**: Builds on TransFuser (Chitta et al. 2022), which fuses image and LiDAR features for planning. A one-hot style vector is concatenated with the trajectory query, fused via an MLP to restore dimensionality, and then fed into the trajectory prediction head.
- **DiffusionDrive-Style**: Modifies DiffusionDrive (Liao et al. 2025a) by injecting a one-hot style vector into the trajectory head. The vector is concatenated with agent features, fused via a fusion MLP, and followed by a regression MLP. This block runs twice for cascade refinement.
- **WoTE-Style**: Adapts the BEV World model (Li et al. 2025b), which forecasts future BEV states. Driving style is injected into the offset prediction head via concatenation and MLP fusion, following the same strategy as in DiffusionDrive-Style.

Training details are provided in Supplementary.

Main Results Analysis

Style Conditioning Improves Behavioral Alignment. The first two sections of Tab. 2 present the quantitative evaluation results of the proposed baseline models on the StyleDrive benchmark. Among the three model families - TransFuser, WoTE, and DiffusionDrive - the style-conditioned variants achieve higher SM-PDMS scores than their vanilla counterparts, clearly demonstrating the benefit of style conditioning. In contrast, the AD-MLP-Style variant slightly underperforms its vanilla version on SM-PDMS.

Overall, the improvements observed in TransFuser, WoTE, and DiffusionDrive, particularly on style-sensitive metrics like TTC, Comf., and EP, confirm the effectiveness of incorporating style information into planning. The consistent gains of style-conditioned models, especially on Comfort and EP, also provide strong evidence for the validity of our driving style annotations. At an individual level, DiffusionDrive-Style delivers the strongest performance, achieving the highest scores across most metrics. WoTE-Style and TransFuser-Style closely follow, and notably, both outperform the vanilla DiffusionDrive model, further highlighting the benefit of style conditioning.

The divergent trend in AD-MLP can be attributed to its simplicity: lacking perception, the model cannot effectively leverage style information. Although EP shows a slight improvement, indicating marginally better intent preservation, the drop in DAC ultimately leads to a lower overall score.

Ablation of Fixed Style Conditioning. The last section of Tab. 2 reports an ablation where the DiffusionDrive-Style model is evaluated by overriding the annotated style and instead enforcing a fixed style condition at test time. For example, DiffusionDrive-Style-A uses the aggressive style condition for all scenes, regardless of ground-truth condition. The results further confirm the effect and learnability of our style conditions: as the conditioned style shifts from aggressive to conservative, NC and DAC show clear improvements. And TTC and Comf. also increase, as SM-PDMS metrics do not penalize overly safe or smooth behaviors. EP peaks with normal style due to its dataset dominance. Besides, fixed-style conditioning underperforms ground-truth style, further validating the reliability of style labels.

Closeness to Human Demonstrations. To further assess the behavioral fidelity of style-conditioned models, we conduct an open-loop evaluation using L2 trajectory error against human driving demonstrations. As shown in Tab. 3, style-aware variants consistently exhibit lower prediction errors across all horizons, with DiffusionDrive-Style achieving the best performance. These results highlight that style conditioning not only enhances stylistic expressiveness but also improves direct consistency with human behavior.

Sensitivity Analysis of PDMS vs. SM-PDMS Tab. 4 presents a comparative analysis of the original PDMS and SM-PDMS architectures in terms of standard deviation and range of style-sensitive sub-metrics and final scores in three valid style-conditioned models. SM-PDMS exhibits higher standard deviation and range in final metric scores and style-sensitive sub-metrics such as EP and Comf across models, indicating improved sensitivity to driving style.

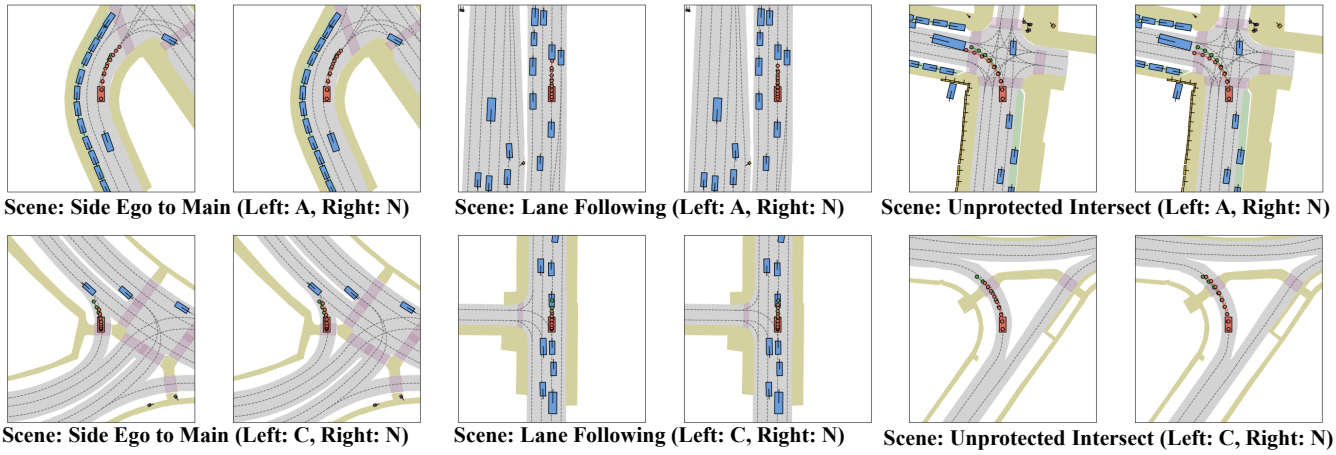

Figure 5: Qualitative illustration of DiffusionDrive-Style predictions under different style conditions across identical scenarios. Left: Aggressive (A) vs. Normal (N); Right: Conservative (C) vs. Normal (N). Red lines indicate the model's predicted trajectory under the given style condition; green lines denote the ground-truth human trajectory. Clear behavioral differences emerge with style variation, reflecting the model's ability to adapt its outputs to driving preferences.

Models	2s ↓	3s ↓	4s ↓	Avg ↓
WoTE	0.733	1.434	2.349	1.506
AD-MLP	0.503	1.262	2.383	1.382
TransFuser	0.431	0.963	1.701	1.032
DiffusionDrive	0.471	1.086	1.945	1.167
WoTE-Style	0.673	1.340	2.223	1.412
AD-MLP-Style	0.510	1.230	2.321	1.354
TransFuser-Style	0.424	0.937	1.656	1.006
DiffusionDrive-Style	**0.417**	**0.940**	**1.646**	**1.001**

Table 3: L2 Open-loop Evaluation Results. Style-conditioned models yield lower trajectory errors than vanilla models, demonstrating better alignment with human driving under preference-aware settings.

Attribute	Original PDMS		SM-PDMS	
	std.	range	std.	range
EP	1.657	3.28	**4.339**	**7.57**
TTC	0.643	1.16	0.575	1.08
Comf.	0.000	0.00	**0.312**	**0.59**
Scores	1.614	2.48	**1.660**	**3.01**

Table 4: Standard deviation (std.) and range of evaluation metrics across valid style-conditioned models under PDMS and SM-PDMS. SM-PDMS increases sensitivity to style-specific behaviors while maintaining overall consistency, whereas PDMS metric exhibits limited discriminative capacity, even with no variation in Comfortable submetric

Qualitative Case Study of Style Effects. To further illustrate the behavioral influence of style conditioning, Fig. 5 shows how our DiffusionDrive-Style model produces different trajectory predictions under aggressive, conservative, and normal style inputs across identical scenarios. Across scenarios, clear differences emerge in the model's trajectory choices. These visualizations confirm that the same policy network, when conditioned on distinct style vectors, can produce behaviorally diverse outputs aligned with human-like style variations. This evidences the controllability and expressiveness afforded by the style-conditioning mechanism.

5 Conclusion

This paper introduces StyleDrive, a novel large-scale dataset and benchmark tailored for advancing personalized E2EAD. By systematically integrating map topology analysis, fine-grained semantic context from vision-language models, and a hybrid annotation pipeline combining rule&distribution-based heuristics with subjective VLM-based reasoning, we establish a rich and interpretable foundation for driving style annotation. This dataset covers a broad spectrum of real-world traffic scenarios annotated with carefully designed style preference labels. To facilitate development and evaluation of personalized E2EAD, we further propose the StyleDrive Benchmark, a non-reactive simulation-based evaluation playground. Central to this benchmark is the Style-Modulated PDMS metric, which augments traditional safety and feasibility assessments with stylistic alignment measures calibrated to intended driver preferences. Extensive experiments across multiple model paradigms demonstrate the effect of style conditioning in enhancing behavioral alignment while preserving core driving competence.

Outlook. ***Coarse-to-Fine Style Labeling***: Current style annotation adopts a 3-level hierarchy, yet attempts at finer levels often result in ambiguous or overlapping. Further enhanced scene understanding/modeling may help resolve such issues. ***Model***: Beyond benchmark's baselines, future work should explore joint modeling of scene context and driving style preferences. ***Application***: Inferring styles from user profiles in real-world settings remains an open challenge with implications for commercial deployment.

Acknowledgments

This work is supported by the Wuxi Research Institute of Applied Technologies at Tsinghua University under Grant No. 20242001120.

References

Aledhari, M.; Rahouti, M.; Qadir, J.; Qolomany, B.; Guizani, M.; and Al-Fuqaha, A. 2023. Motion comfort optimization for autonomous vehicles: Concepts, methods, and techniques. *IEEE Internet of Things Journal*, 11(1): 378–402.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11621–11631.

Chitta, K.; Prakash, A.; Jaeger, B.; Yu, Z.; Renz, K.; and Geiger, A. 2022. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 12878–12895.

Codevilla, F.; Müller, M.; López, A.; Koltun, V.; and Dosovitskiy, A. 2018. End-to-end driving via conditional imitation learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 4693–4700. IEEE.

Contributors. 2024. CARLA Autonomous Driving Leaderboard.

Cui, C.; Yang, Z.; Zhou, Y.; Ma, Y.; Lu, J.; Li, L.; Chen, Y.; Panchal, J.; and Wang, Z. 2024. Personalized autonomous driving with large language models: Field experiments. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 20–27. IEEE.

Dauner, D.; Hallgarten, M.; Li, T.; Weng, X.; Huang, Z.; Yang, Z.; Li, H.; Gilitschenski, I.; Ivanovic, B.; Pavone, M.; et al. 2024. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Conference on Neural Information Processing Systems (NeurIPS)*, 37: 28706–28719.

Hasenjäger, M.; and Wersing, H. 2017. Personalization in advanced driver assistance systems and autonomous vehicles: A review. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 1–7. IEEE.

Jain, A.; Koppula, H. S.; Soh, S.; Raghavan, B.; Singh, A.; and Saxena, A. 2016. Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture. *arXiv preprint arXiv:1601.00740*.

Jia, X.; Wu, P.; Chen, L.; Xie, J.; He, C.; Yan, J.; and Li, H. 2023. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 21983–21994.

Jia, X.; Yang, Z.; Li, Q.; Zhang, Z.; and Yan, J. 2024. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Conference on Neural Information Processing Systems (NeurIPS)*, 37: 819–844.

Jia, X.; You, J.; Zhang, Z.; and Yan, J. 2025. DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving. In *International Conference on Learning Representations (ICLR)*.

Ke, Z.; Jiang, Y.; Wang, Y.; Cheng, H.; Li, J.; and Wang, J. 2024. D2E: An Autonomous Decision-Making Dataset involving Driver States and Human Evaluation of Driving Behavior. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2294–2301. IEEE.

Kou, G.; Jia, F.; Mao, W.; Liu, Y.; Zhao, Y.; Zhang, Z.; Yoshie, O.; Wang, T.; Li, Y.; and Zhang, X. 2025. PADriver: Towards Personalized Autonomous Driving. *arXiv preprint arXiv:2505.05240*.

Li, Q.; Peng, Z.; Feng, L.; Zhang, Q.; Xue, Z.; and Zhou, B. 2022. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3461–3475.

Li, S.; Wei, C.; Wu, G.; Barth, M. J.; Abdelraouf, A.; Gupta, R.; and Han, K. 2023. Personalized trajectory prediction for driving behavior modeling in ramp-merging scenarios. In *IEEE International Conference on Robotic Computing*, 1–4. IEEE.

Li, Y.; Fan, L.; He, J.; Wang, Y.; Chen, Y.; Zhang, Z.; and Tan, T. 2025a. Enhancing end-to-end autonomous driving with latent world model. In *International Conference on Learning Representations (ICLR)*.

Li, Y.; Wang, Y.; Liu, Y.; He, J.; Fan, L.; and Zhang, Z. 2025b. End-to-end driving with online trajectory evaluation via bev world model. *arXiv preprint arXiv:2504.01941*.

Liao, B.; Chen, S.; Yin, H.; Jiang, B.; Wang, C.; Yan, S.; Zhang, X.; Li, X.; Zhang, Y.; Zhang, Q.; et al. 2025a. DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12037–12047.

Liao, X.; Zhao, X.; Wang, Z.; Zhao, Z.; Han, K.; Gupta, R.; Barth, M. J.; and Wu, G. 2023. Driver digital twin for online prediction of personalized lane-change behavior. *IEEE Internet of Things Journal*, 10(15): 13235–13246.

Liao, X.; Zhao, Z.; Barth, M. J.; Abdelraouf, A.; Gupta, R.; Han, K.; Ma, J.; and Wu, G. 2025b. A review of personalization in driving behavior: Dataset, modeling, and validation. *IEEE Transactions on Intelligent Vehicles*, 10(2): 1241–1262.

Marcu, A.-M.; Chen, L.; Hünermann, J.; Karnsund, A.; Hanotte, B.; Chidananda, P.; Nair, S.; Badrinarayanan, V.; Kendall, A.; Shotton, J.; et al. 2024. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision (ECCV)*, 252–269. Springer.

Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. 2023. Openscene: 3d scene understanding with open vocabularies. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–824.

Prakash, A.; Chitta, K.; and Geiger, A. 2021. Multi-modal fusion transformer for end-to-end autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7077–7087.

Romera, E.; Bergasa, L. M.; and Arroyo, R. 2016. Need data for driver behaviour analysis? Presenting the public UAH-DriveSet. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 387–392. IEEE.

Schrum, M. L.; Sumner, E.; Gombolay, M. C.; and Best, A. 2024. Maveric: A data-driven approach to personalized autonomous driving. *IEEE Transactions on Robotics*, 40: 1952–1965.

Shao, H.; Wang, L.; Chen, R.; Li, H.; and Liu, Y. 2023. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning (CoRL)*, 726–737. PMLR.

Speidel, O.; Graf, M.; Phan-Huu, T.; and Dietmayer, K. 2019. Towards courteous behavior and trajectory planning for automated driving. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 3142–3148. IEEE.

Tian, H.; Wei, C.; Jiang, C.; Li, Z.; and Hu, J. 2022. Personalized lane change planning and control by imitation learning from drivers. *IEEE Transactions on Industrial Electronics*, 70(4): 3995–4006.

Wei, C.; Qin, Z.; Li, S.; Zhang, Z.; Zhao, X.; Abdelraouf, A.; Gupta, R.; Han, K.; Barth, M. J.; and Wu, G. 2025. PDB: Not All Drivers Are the Same–A Personalized Dataset for Understanding Driving Behavior. *arXiv preprint arXiv:2503.06477*.

Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K.-Y. K.; Li, Z.; and Zhao, H. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10): 8186–8193.

Zhai, J.-T.; Feng, Z.; Du, J.; Mao, Y.; Liu, J.-J.; Tan, Z.; Zhang, Y.; Ye, X.; and Wang, J. 2023. Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. *arXiv preprint arXiv:2305.10430*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *arXiv preprint arXiv:2501.13106*.

Zhao, Z.; Wang, Z.; Han, K.; Gupta, R.; Tiwari, P.; Wu, G.; and Barth, M. J. 2022. Personalized car following for autonomous driving with inverse reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2891–2897. IEEE.

Zheng, W.; Song, R.; Guo, X.; Zhang, C.; and Chen, L. 2024. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision (ECCV)*, 87–104. Springer.

Zhu, B.; Yan, S.; Zhao, J.; and Deng, W. 2018. Personalized lane-change assistance system with driver behavior identification. *IEEE Transactions on Vehicular Technology*, 67(11): 10293–10306.