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

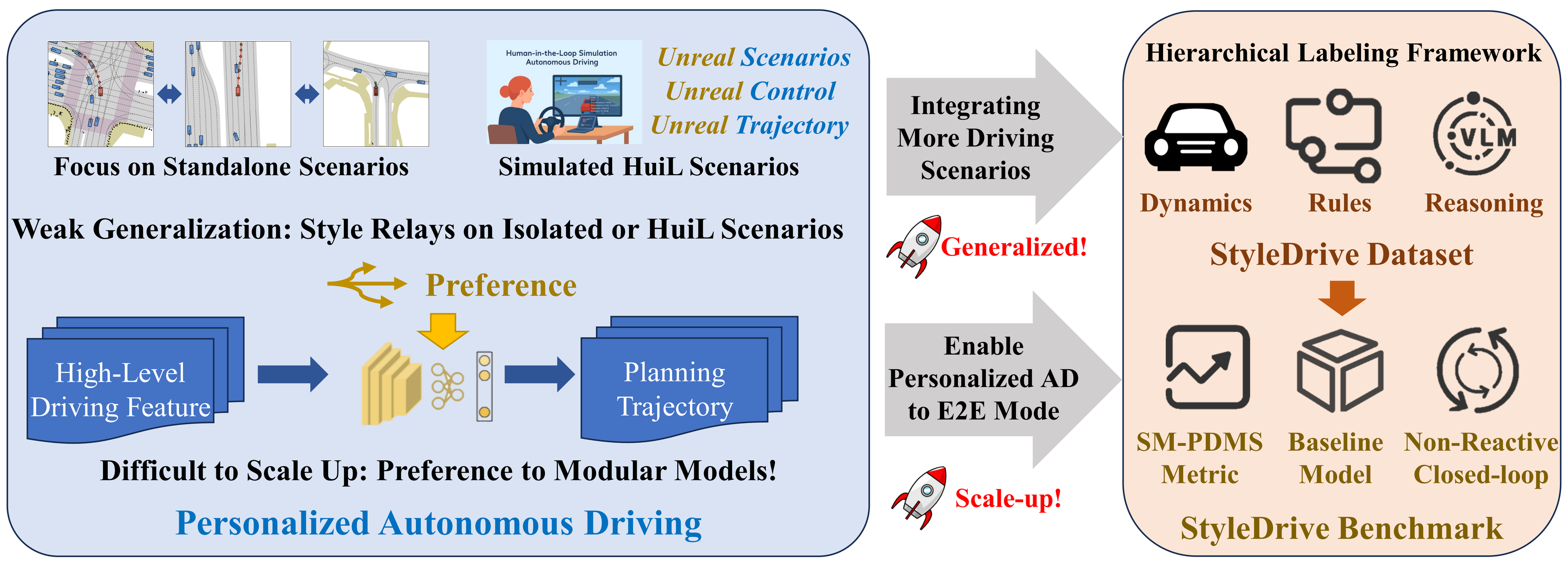
THU KCL UoM HKU

Ruiyang Hao[1,2], Bowen Jing[3], Haibao Yu[1,4], Zaiqing Nie[1,*]

GitHub Home

Association for the Advancement of Artificial Intelligence

Significance and Motivation of StyleDrive



To bridge the gap between personalized autonomous driving and E2EAD, we introduce the first dataset and benchmark tailored for personalized E2EAD.

Contribution1: A novel large-scale real-world dataset (30k Clips) for personalized E2EAD, annotated with driving style preferences across diverse traffic scenarios.

Contribution2: A multi-stage annotation pipeline combining rule-based analysis, visual language model (VLM) reasoning, and human-in-the-loop verification to ensure consistent and interpretable style labels.

Contribution3: The first benchmark for personalized E2EAD, enabling standardized and quantitative comparison of style-conditioned driving behavior across different model architectures.

Contribution4: Comprehensive empirical results showing that style-aware models better align with human behavior, demonstrating the value of personalization for improved autonomy.

StyleDrive Dataset

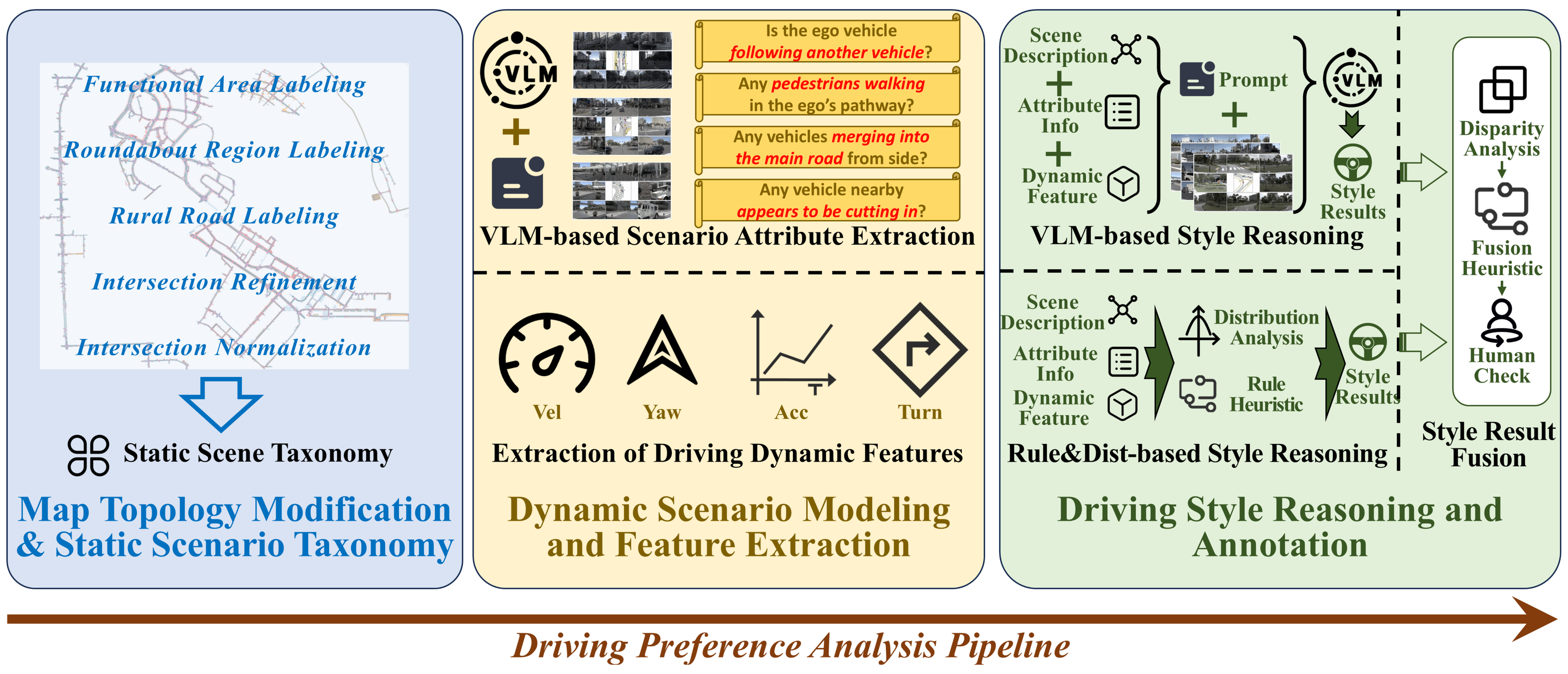
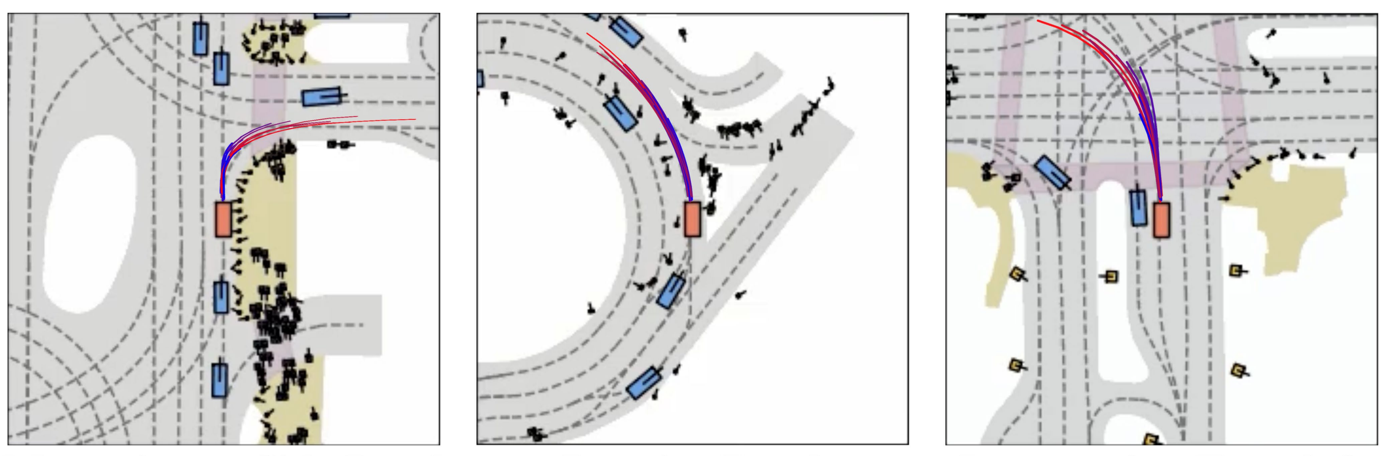
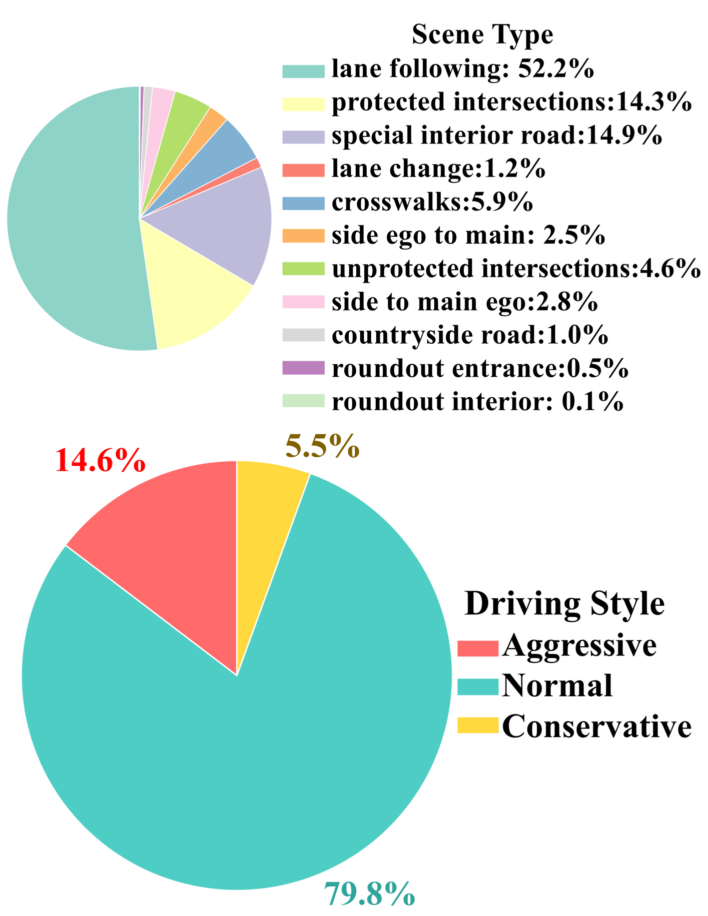


Illustration of the Hybrid Annotation Pipeline integrating Topology, Scene Semantics, Rule-based Reasoning, VLM-based Reasoning, and Human Verification. The final datasets consist of 30k driving scenarios labeled with driving styles.

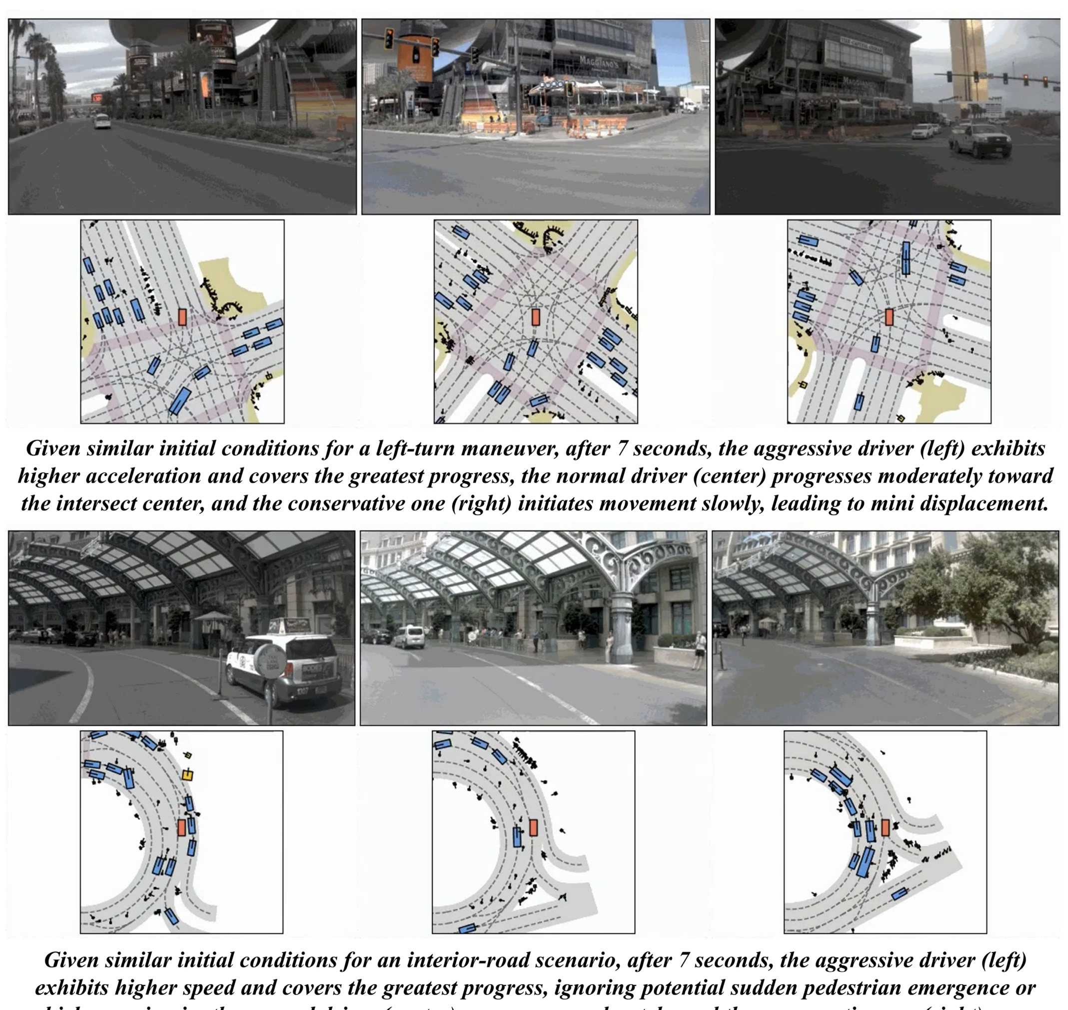


Merge into a Side Road | Interior Road | Intersection Turn-left

Visualization of Driving Style Distribution in 3 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.



Scenario Distribution and Style Distribution of StyleDrive Dataset



Given similar initial conditions for a left-turn maneuver, after 7 seconds, the aggressive driver (left) exhibits higher acceleration and covers the greatest progress, the normal driver (center) progresses moderately toward the intersect center, and the conservative one (right) initiates movement slowly, leading to mini displacement.

Given similar initial conditions for an interior-road scenario, after 7 seconds, the aggressive driver (left) exhibits higher speed and covers the greatest progress, ignoring potential sudden pedestrian emergence or vehicle merging in, the normal driver (center) progresses moderately, and the conservative one (right) moves slowly and cautiously in this complex scenario.

More Visualization of Different Driving Styles

StyleDrive Benchmark

Metric Design

Central to the benchmark is the **Style-Modulated Predictive Driver Model Score (SM-PDMS),** which integrates behavior alignment module to reflect driving style preferences.

- **Preserves safety-related metrics** across styles
- **Adjusts style-sensitive sub-scores based on annotated style tendency**
  - ✓ *Comfort* thresholds tuned to personal tolerance
  - ✓ *Ego progress* aligned with assertiveness
  - ✓ *TTC ranges* adjusted for risk tolerance

Benchmark Methods

We adapt 4 classic E2EAD architectures with driving style as condition input:

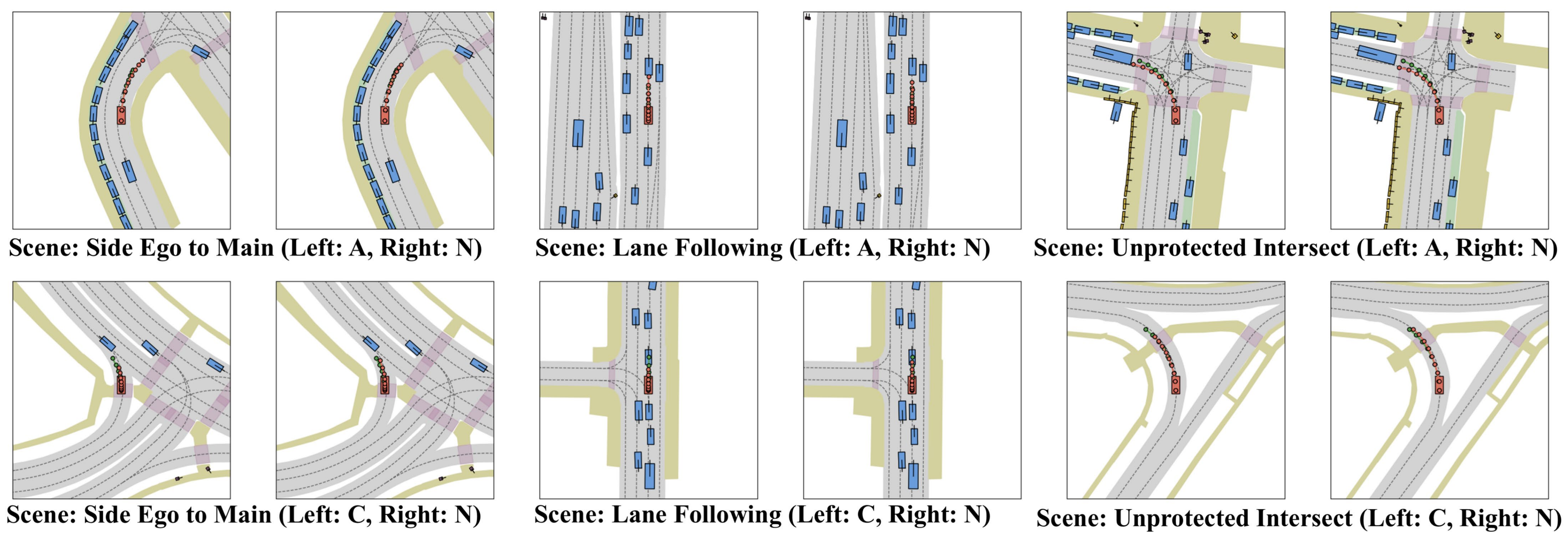
- **AD-MLP-Style**: *Classic MLP*; Concatenates the driving-style vector with ego features and uses an MLP to output style-conditioned trajectories.
- **TransFuser-Style**: *Image + LiDAR fusion model*; Injects the style encoding into the multimodal fusion network to enable style-controlled planning.
- **DiffusionDrive-Style**: *Diffusion-based planner*; Integrating style signals through a two-stage refinement process to generate more personalized trajectories.
- **WoTE-Style**: *BEV world model prediction* model; Incorporates driving-style conditions into the BEV world modeling to modulate trajectory offsets.

Main Results of StyleDrive Benchmark

Models	NC ↑	DAC ↑	Style-Modulated Submetrics TTC ↑	Comf. ↑	EP ↑	SM-PDMS ↑
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

Result Summary:

1. **Style Conditioning Effectively Improves Behavioral Alignment;**
2. **Ablation of Fixed Style Conditioning Verifies Style Controllability;**
3. **Style Conditioning enables Closeness to Human Demonstrations.**



Qualitative illustration of DiffusionDrive-Style predictions under different style conditions across identical scenarios. Top Row: Aggressive vs. Normal; Down Row: Conservative vs. Normal. *Red Traj*: the model's predicted trajectory under the given style condition; *Green Traj*: 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**

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 std.	range	SM-PDMS 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**

Metric Verification: SM-PDMS Metric enhances **Sensitivity to Style-Specific Behaviors** while preserving overall consistency, whereas Traditional PDMS metric shows limited discriminative power

Outlook:

1. **Dataset:** Coarse-to-Fine Style Labeling;
2. **Model:** Joint Modeling of Scene & Style Preferences;
3. **Real-world Application**: Inferring Driving Styles from User Profiles.