Drive My Way: Preference Alignment of Vision-Language-Action Model for Personalized Driving

Anonymous submission

Paper ID

Abstract

Human driving is inherently personal, shaped by long-term habits and short-term intentions. Existing end-to-end autonomous driving systems optimize for generic objectives or fixed modes, lacking the ability to adapt to individual preferences or natural language intent. We propose Drive My Way (DMW), a personalized Vision-Language-Action (VLA) framework that aligns with users' long-term driving habits and adapts to real-time instructions. DMW learns a user embedding from our personalized driving dataset and conditions the policy on this embedding, while language instructions provide short-term guidance. Closed-loop evaluation on Bench2Drive shows that DMW improves style adaptation, and user studies confirm that its behaviors are recognizable as each driver's own style.

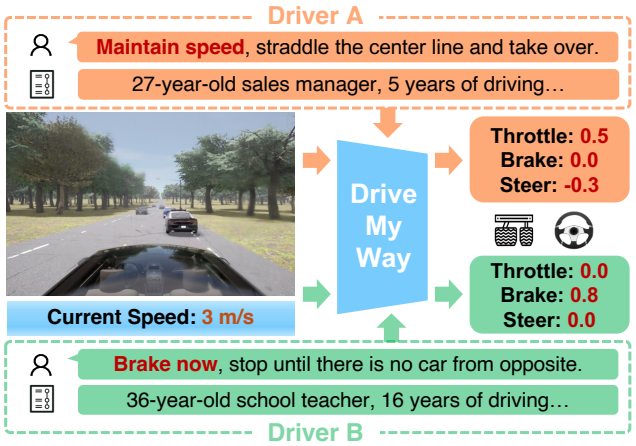

Figure 1. DMW achieves end-to-end personalized driving via long-term preference alignment and short-term style instruction adaptation.

1. Introduction

End-to-end autonomous driving has emerged as a powerful paradigm that directly maps multi-modal sensor inputs to trajectories or control actions [1–3]. Recent advances in unified architectures and foundation-model-based frameworks have demonstrated impressive performance on open-loop benchmarks through imitation of expert trajectories. However, these systems optimize generic safety and efficiency objectives and overlook the *individuality* of human driving: drivers differ markedly in how assertively they accelerate, brake, or overtake depending on purpose or emotional state (e.g., commuting, leisure, or emergency travel). Existing personalized systems offer only a few preset modes ("sport", "comfort") or manual parameter adjustments that cannot capture subtle, evolving preferences [4, 5], nor can they interpret intuitive natural-language cues such as "I'm tired" or "I'm late for work". Long-term, context-adaptive personalization is thus crucial for enhancing user trust, comfort, and satisfaction [6, 7].

Existing research on personalized autonomous driving falls into two categories. Data-driven methods [6, 8] extract predefined style labels from human demonstrations and learn style-conditioned policies using behavior cloning or inverse reinforcement learning. While effective at mimicking human-like styles, these approaches require large-scale datasets and scale poorly to a growing population of users with heterogeneous preferences. Multi-objective methods [9] mitigate this limitation through runtime adjustment of preference weights, but cannot handle real-time human interaction via natural language or adapt to user instructions on the fly. Language-driven approaches [10–12] leverage LLMs for instruction-based personalization, providing high-level cues for interpretable decisions, yet they have not been systematically investigated in complex, dynamic driving contexts such as merging, overtaking, or emergency braking, where trade-offs among safety, comfort, and efficiency must adapt to dynamic scenes.

We propose **DMW** (Drive My Way), a personalized end-to-end autonomous driving framework that integrates contextual user embeddings to align policy behaviors with individual driving preferences. DMW also enables real-

time human interaction and rapid response to natural language instructions via a style-aware reinforcement fine-tuning stage (Fig. 1). To enable both long-term personalization and short-term real-time adaptation, we construct the Personalized Driving Dataset (PDD) with thirty real drivers performing diverse traffic scenarios in CARLA [13], which provides a valuable dataset for developing and evaluating human-centered driving models. We conduct extensive evaluations on the Bench2Drive benchmark [14], complemented by personalization metrics and user studies, which demonstrate the effectiveness of DMW in aligning its driving behavior to individual user preferences while maintaining a balanced trade-off among safety, efficiency, and comfort. Our contributions are summarized as: (i) DMW, a personalized VLA framework with user-embedding conditioning and style-aware RL fine-tuning on natural language instructions; (ii) the Personalized Driving Dataset (PDD) collected from thirty real drivers in CARLA across twenty diverse scenarios; and (iii) a thorough closed-loop evaluation on Bench2Drive with personalization metrics and user studies, establishing strong gains over prior style-conditioned and preference-based baselines.

2. Personalized Driving Dataset

We construct a new **Personalized Driving Dataset (PDD)** (Fig. 2) capturing driving environments and personal context across diverse, safety-critical situations. Unlike prior datasets with synthetic instruction-action pairs [15] or retrospectively inferred style labels [6], PDD records real human behavior during complex real-time interactions.

We recruit thirty drivers with diverse backgrounds. Each completes a structured questionnaire covering demographics, driving history, and typical driving purposes, as commuting or leisure travel. These profiles provide a semantic context that connects personal background with long-term driving habits. Each participant then performs twenty standardized CARLA scenarios spanning overtaking, merging, intersections, and pedestrian/cut-in events, controlled via a Logitech G-series steering wheel and pedal setup to enable realistic-looking control inputs. PDD records ego-vehicle motion (throttle, brake, steering, and trajectory), scene perception (surrounding vehicles, pedestrians, cyclists, hazards, and handling interactions), and navigating pedestrian crossings or encountering vehicle cut-ins. All drivers operate the vehicle using a Logitech G-series steering wheel and pedal setup to enable realistic control inputs. PDD additionally records ego-vehicle motion states, scene perception of surrounding vehicles, cyclists, and road hazards, as well as traffic context such as signals, signs, route geometry, and speed limits. We also record an expert target speed from PDM-Lite [16], and the deviation between the human-controlled speed and this target serves as a dense descriptor of driver style and situational preference.

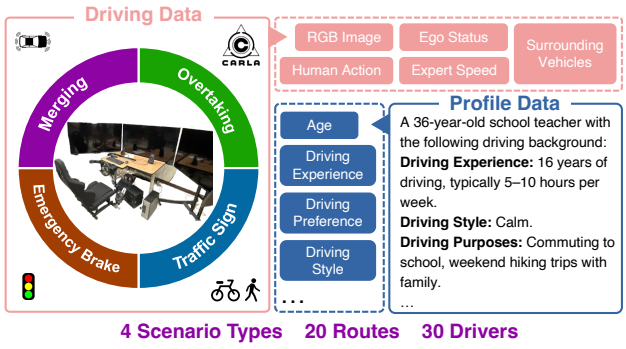

Figure 2. Overview of the Personalized Driving Dataset, consisting of driving data and structured driver profiles.

3. Method

Problem Setup. For each driver $m \in \mathcal{M} = \{1, \dots, M\}$, we collect a dataset $\mathcal{D}^m = \{(s_t^m, a_t^m)\}_{t=1}^{T^m}$ and a profile P^m summarizing background and habits. We cast personalized driving as an MDP with observation $o_t = (\mathbf{I}_t, \mathbf{q}_t, I_t, g_t, P^m)$, where $\mathbf{I}_t$ is the front-view image, $\mathbf{q}_t$ the ego state, I_t a language instruction, g_t the target, and action $\mathcal{A} \subset \mathbb{R}^3$ over throttle, brake, and steering. The goal is a policy $\pi_\theta(a_t \mid o_t)$ that maximizes expected return while conditioning on a user embedding z_p^m encoded from P^m.

VLA Backbone and Residual Decoder. As shown in Fig. 3, we build on SimLingo [17], an InternVL2-1B [18]-based VLA with InternViT-300M and Qwen2-0.5B. The backbone predicts temporal and geometric waypoints yielding a base action a_t^{base} via a PID controller. To allow personalized specialization, we introduce learnable *residual query* tokens fused with vision, language, navigation, user-embedding, and motion tokens in the LM. An MLP with a categorical head produces discrete residual adjustments in speed and steering, yielding $a_t = a_t^{\text{base}} + a_t^{\Delta}$. This preserves safe planning while expressing diverse styles.

User Embedding Learning. A long-term preference encoder $f_p(\cdot)$ maps profile P^m to z_p^m via DeBERTaV3 [19] and a projection head. A route processor $f_b(\cdot)$ (temporal encoder with self-attention) maps a length-k history $\xi_t^m = \{(\mathbf{I}_{t-k:t}^m, \mathbf{q}_{t-k:t}^m, a_{t-k:t}^m)\}$ to a behavior embedding $z_{b,t}^m$. We align profile and behavior via InfoNCE [20]:

$$\mathcal{L}_t^m = -\log \frac{\exp(\text{sim}(z_p^m, z_{b,t}^m)/\tau)}{\sum_{j=1}^{M} \exp(\text{sim}(z_p^j, z_{b,t}^m)/\tau)}. \quad (1)$$

Preference Alignment via GRPO. We condition the policy on z_p^m and fine-tune with GRPO [21]. For each target driver m, we augment with a least-similar peer $u = \arg\min_x \text{sim}(z_p^m, z_p^x)$; the augmented action is $\tilde{a}_t^m = (\bar{a}^m / \bar{a}^u) \cdot a_t^m$. The reward is $\mathcal{R}(s_t^m, a_t) = d(a_t, a_t^m)$ when conditioned on P^m and $d(a_t, \tilde{a}_t^m)$ when conditioned on P^u.

Style Instruction Construction. In complex and dy-

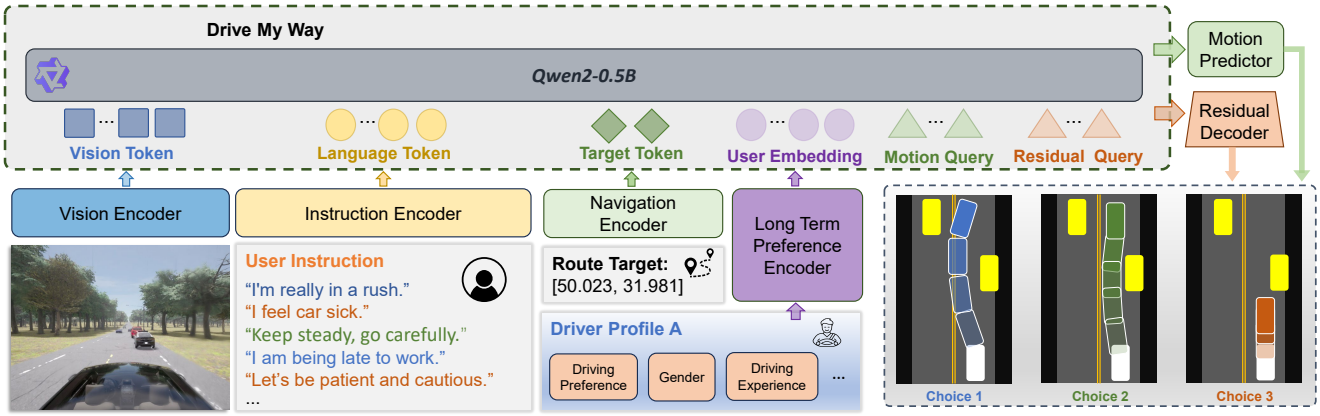

Figure 3. Overview of DMW. The VLA backbone takes front-view images, instructions, route target points, and user profile as inputs; the motion predictor outputs route and speed waypoints deriving the base action, while the residual decoder outputs a discrete residual producing the final personalized action.

namically evolving traffic scenarios, effective personalization requires the autonomous agent to infer not only a driver's long-term preference but also the short-term context-adaptive cues that vary across situations. To capture these nuances, we build a style instruction set spanning twenty distinct driving scenarios. Each instruction incorporates both style intent and scene-specific semantics, covering three driving styles expressed at three levels of directness, which encode the degrees of implicitness defined in prior work [10]. This design enables the model to interpret both explicit commands and subtle linguistic cues that reflect real-time user intent across diverse contexts.

Style-Aware Reward. To translate user intent into safe, personalized objectives, the driving reward combines safety, efficiency, and comfort:

$$\mathcal{R}(s_t, a_t) = w_s R_{\text{safety}} + w_e R_{\text{eff}} + w_c R_{\text{comf}}, \qquad (2)$$

with $R_{\text{safety}} = \mathbb{I}(\text{TTC}_t \geq \beta_{\text{safety}})$, $R_{\text{eff}} = \exp(-\alpha|v_t - v_{\text{pref}}|)$, and $R_{\text{comf}} = \mathbb{I}(|a_t^{\text{steer}}| < \beta_{\text{lat}} \wedge |a_t^{\text{acc}}| < \beta_{\text{long}})$. Weights and thresholds are style-dependent: an LLM (GPT-5) infers $(w_s, w_e, w_c, \beta_{\text{safety}}, v_{\text{pref}}, \beta_{\text{lat}}, \beta_{\text{long}})$ from the scenario description and instruction style $S \in \{\text{Conservative, Neutral, Aggressive}\}$, and an expert review refines them. Aggressive styles receive higher w_e and v_{pref}; conservative ones receive tighter TTC. This formulation provides a more distinct optimization objective for each style than relying on a generic combination, and the generated parameters are initialized within predefined upper bounds for each style to maintain balanced trade-offs among metrics, bridging the gap between subtle preferences in language and dynamic style rewards.

4. Experiments

Setup. We evaluate closed-loop on Bench2Drive [14], terminating on collision. We compare DMW with SimLingo [17], a StyleDrive-like [6] policy, and MORL-PD [9].

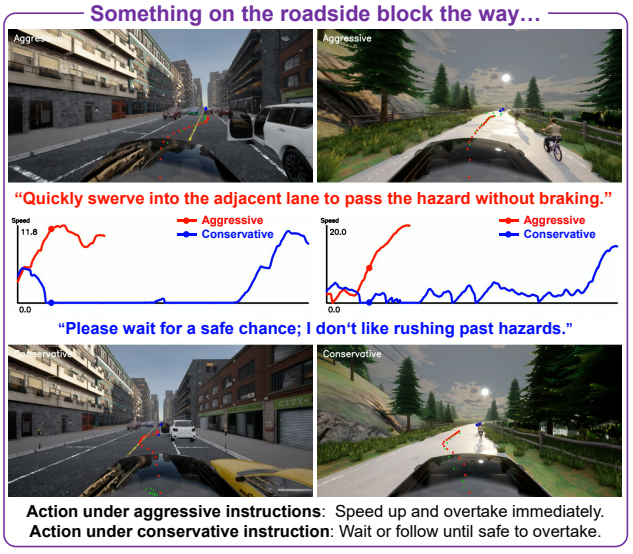

Figure 4. Short-term preference adaptation under aggressive and conservative instructions in safety-critical scenarios.

Metrics include DS, SR, Efficiency, Comfort, Speed, Acceleration, Lane Changes (LC), Headway, and Travel Time (TT). The Alignment Score (AS) is classification accuracy of roll-outs into driver clusters from historical logs; ten evaluators also rate similarity (1–10). We test on 25 ID and 5 OOD drivers. Training: AdamW (lr 1e-4), LoRA [22] on Qwen2-0.5B, 8×A6000, GRPO with 4 samples per input.

Adaptation to Style Instructions. Table 1 shows that DMW achieves an 18.77% efficiency gain under aggressive instructions (vs. 3.70% for SimLingo and 6.00% for StyleDrive), while conservative instructions yield the highest DS and SR with larger headway. Reduced acceleration and lane-change counts under conservative styles confirm distinct behavioral shifts. DMW-Vanilla with fixed reward weights already improves DS/SR but collapses style differ-

Table 1. **Bench2Drive closed-loop metrics under different style instructions.** We compare SimLingo and StyleDrive with our fine-tuning using fixed reward weights (DMW-Vanilla) and style-aware weights (**DMW**).

Method	Style	DS	SR	Efficiency	Comfort	Speed	Acceleration	LC	Headway	TT
SimLingo [17]	Aggressive	78.56	65.83	247.60	18.61	7.66	5.39	0.75	25.99	25.35
	Neutral	78.15	65.85	241.44	24.67	7.37	5.22	0.75	27.81	31.41
	Conservative	78.18	65.56	238.77	26.99	7.21	5.29	0.70	29.12	33.02
StyleDrive [6]	Aggressive	75.68	60.89	256.71	16.79	7.23	5.59	0.74	24.95	27.76
	Neutral	76.26	62.13	249.07	21.35	6.98	5.43	0.66	23.62	29.12
	Conservative	77.02	61.96	242.18	23.67	6.82	5.39	0.70	27.19	29.98
DMW-Vanilla	Aggressive	82.19	70.97	253.10	15.86	**7.86**	5.29	**0.78**	26.46	19.69
	Neutral	81.96	70.63	247.77	19.21	7.66	5.17	0.77	26.63	23.16
	Conservative	81.48	71.05	246.80	21.87	7.75	5.37	0.75	26.90	22.51
DMW	Aggressive	79.50	67.36	**281.56**	21.62	7.72	**6.01**	0.70	26.37	26.93
	Neutral	82.03	70.95	244.98	28.67	6.34	5.43	0.61	27.60	40.75
	Conservative	**82.72**	**71.56**	237.06	**34.62**	6.18	5.26	0.60	**30.05**	**47.38**

Table 2. **Driving metrics with and without style instructions, for each driver profile with a different long-term preference.** Note that for Alignment Score (AS) and Ratings, we compute their values regardless of the style since they only measure long-term alignment.

Driver	**Scenario**	**Style**	**DS**	**Speed**	**Efficiency**	**Acceleration**	**Headway**	**AS**	**Ratings**
D1	Emergency Brake	None ↓ Aggressive	95.10 → 84.02	8.06 → 9.58	150.94 → 167.12	6.31 → 6.39	18.94 → 17.11	1.00	8.8
D2			98.64 → 96.71	4.61 → 4.97	96.02 → 99.21	4.94 → 5.35	23.10 → 21.18	0.67	8.3
D3			97.88 → 95.93	5.39 → 6.34	109.83 → 121.06	5.63 → 6.16	21.61 → 19.22	1.00	8.0
D4			94.02 → 93.15	7.76 → 8.35	145.22 → 159.56	6.03 → 6.92	21.03 → 18.19	0.67	7.8
D1	Merging	None ↓ Conservative	90.05 → 94.38	9.32 → 8.76	270.84 → 260.10	5.15 → 4.68	96.44 → 101.16	0.67	8.4
D2			86.94 → 91.72	6.37 → 5.76	196.82 → 175.31	4.29 → 4.16	118.92 → 121.03	1.00	8.1
D3			96.85 → 95.94	6.98 → 6.20	205.41 → 195.77	4.46 → 4.24	111.26 → 114.05	0.67	7.6
D4			97.31 → 97.88	8.78 → 8.63	261.52 → 259.52	5.01 → 4.91	88.76 → 100.93	1.00	8.2
D1	Overtaking	None ↓ Neutral	97.56 → 97.94	7.70 → 7.94	220.19 → 223.13	7.28 → 7.37	14.69 → 14.53	1.00	9.0
D2			98.23 → 98.41	5.50 → 5.56	167.40 → 167.80	5.65 → 5.78	22.23 → 22.04	1.00	8.6
D3			97.14 → 96.62	6.34 → 6.13	191.02 → 183.84	6.18 → 6.01	20.79 → 22.19	0.67	7.5
D4			80.61 → 82.19	9.63 → 9.51	271.98 → 265.84	7.11 → 6.82	17.81 → 18.03	1.00	8.1
D1	Traffic Sign	None ↓ Conservative	90.82 → 96.67	9.08 → 8.49	408.74 → 401.28	7.12 → 6.37	31.44 → 32.52	1.00	8.7
D2			97.92 → 98.35	5.79 → 4.91	193.26 → 183.11	5.72 → 5.51	35.02 → 36.91	1.00	8.2
D3			89.31 → 92.02	6.88 → 5.37	237.05 → 231.46	5.96 → 5.91	34.54 → 36.51	1.00	8.1
D4			95.91 → 97.32	8.91 → 6.74	418.02 → 379.68	6.54 → 5.81	31.19 → 31.77	0.67	7.9

ences, validating the need for style-aware adaptation.

Long-term Preference Alignment. Table 2 reports per-driver metrics for two seen (D1, D2) and two unseen (D3, D4) drivers. D1/D4 show higher efficiency and acceleration, while D2/D3 show lower acceleration and a larger headway. Behaviors remain distinct across scenarios, indicating effective long-term alignment. Compared with MORL-PD, which has an AS of 0.50 in ID and 0.29 in OOD, DMW attains an AS of 0.92 in ID and 0.83 in OOD, and a consistency higher evaluator ratings, indicating that user embeddings capture profile semantics and generalizes to unseen drivers. In addition, consistent with Table 1, instructions can provide additional adaptations, shifting the driving towards the short-term preference.

Qualitative Analysis. Fig. 4 illustrates policy behavior in two safety-critical scenarios: a roadside vehicle with an opened door (left) and a group of bicyclists traveling along the road (right). Under the aggressive command, the policy quickly moves into the adjacent lane, asserting its position more decisively. In contrast, under the conservative command (blue), it slows down earlier and yields to oncoming traffic before proceeding. These results show that the same visual observation can lead to substantially different speed profiles, indicating that style-aware reward shaping captures distinct behavioral characteristics rather than collapsing to an averaged policy, consistent with results in Table 1.

5. Conclusion

We introduced a personalized VLA driving framework that captures long-term driver preferences via profile-sensitive user embeddings and adapts to short-term style instructions via style-aware reinforcement fine-tuning. Experiments on Bench2Drive, and user studies, show that DMW generates recognizable and personalized driving behaviors while preserving strong safety. We plan to further expand PDD to cover a broader and more diverse set of drivers.

References

[1] Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, Hao Ye, Zihao Sheng, Xin Zhao, Tuopu Wen, Zheng Fu, Sikai Chen, Kun Jiang, Diange Yang, Seongjin Choi, and Lijun Sun. A survey on vision-language-action models for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4524–4536, October 2025. 1

[2] Litian Gong, Fatemeh Bahrani, Yutai Zhou, Amin Banayeeanzade, Jiachen Li, and Erdem Bıyık. Autofocus-il: Vlm-based saliency maps for data-efficient visual imitation learning without extra human annotations. In *International Conference on Robotics and Automation (ICRA)*, 2026.

[3] Rui Zhao, Yuze Fan, Ziguo Chen, Fei Gao, and Zhenhai Gao. Diffe2e: Rethinking end-to-end driving with a hybrid diffusion-regression-classification policy. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[4] Genghua Kou, Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Ziheng Zhang, Osamu Yoshie, Tiancai Wang, Ying Li, and Xiangyu Zhang. Padriver: Towards personalized autonomous driving. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2025. 1

[5] Manabu Nakanoya, Junha Im, Hang Qiu, Sachin Katti, Marco Pavone, and Sandeep Chinchali. Personalized federated learning of driver prediction models for autonomous driving, 2021. 1

[6] Ruiyang Hao, Bowen Jing, Haibao Yu, and Zaiqing Nie. Styledrive: Towards driving-style aware benchmarking of end-to-end autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4627–4635, 2026. 1, 2, 3, 4

[7] Udita Ghosh, Dripta S. Raychaudhuri, Jiachen Li, Konstantinos Karydis, and Amit K. Roy-Chowdhury. Reducing oracle feedback with vision-language embeddings for preference-based rl. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2026. 1

[8] Anqing Jiang, Yu Gao, Yiru Wang, Zhigang Sun, Shuo Wang, Yuwen Heng, Hao Sun, Shichen Tang, Lijuan Zhu, Jinhao Chai, et al. Irl-vla: Training an vision-language-action policy via reward world model. *arXiv preprint arXiv:2508.06571*, 2025. 1

[9] Hendrik Surmann, Jorge De Heuvel, and Maren Bennewitz. Multi-objective reinforcement learning for adaptable personalized autonomous driving. In *2025 European Conference on Mobile Robots (ECMR)*, pages 1–8. IEEE, 2025. 1, 3

[10] Can Cui, Zichong Yang, Yupeng Zhou, Yunsheng Ma, Juanwu Lu, Lingxi Li, Yaobin Chen, Jitesh Panchal, and Ziran Wang. Personalized autonomous driving with large language models: Field experiments. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pages 20–27. IEEE, 2024. 1, 3

[11] Can Cui, Zichong Yang, Yupeng Zhou, Juntong Peng, Sung-Yeon Park, Cong Zhang, Yunsheng Ma, Xu Cao, Wenqian Ye, Yiheng Feng, et al. On-board vision-language models for personalized autonomous vehicle motion control: System design and real-world validation. *arXiv preprint arXiv:2411.11913*, 2024.

[12] Yunsheng Ma, Can Cui, Xu Cao, Wenqian Ye, Peiran Liu, Juanwu Lu, Amr Abdelraouf, Rohit Gupta, Kyungtae Han, Aniket Bera, James M. Rehg, and Ziran Wang. Lampilot: An open benchmark dataset for autonomous driving with language model programs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15141–15151, 2024. 1

[13] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 2

[14] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844, 2024. 2, 3

[15] Ziye Qin, Siyan Li, Chuheng Wei, Guoyuan Wu, Matthew J Barth, Amr Abdelraouf, Rohit Gupta, and Kyungtae Han. Investigating personalized driving behaviors in dilemma zones: Analysis and prediction of stop-or-go decisions. *IEEE Robotics and Automation Letters*, 2025. 2

[16] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, 2024. 2

[17] Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11993–12003, 2025. 2, 3, 4

[18] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2

[19] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTav3: Improving deBERTa using ELECTRA-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[20] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[21] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 3