
Supplementary Material: Prioritizing Perception-Guided Self-Supervision: A New Paradigm for Causal Modeling in End-to-End Autonomous Driving

Yi Huang^{12*}, Zhan Qu^{2*}, Lihui Jiang^{2†}, Bingbing Liu², Hongbo Zhang²,

¹The Chinese University of Hong Kong, Shenzhen

²Huawei Noah’s Ark Lab

yihuang11@link.cuhk.edu.cn

{quzhan, jianglihui1, liu.bingbing, zhanghongbo888}@huawei.com

This supplementary document provides additional qualitative results and experimental data that further substantiate the conclusions presented in the main paper. Due to the file-size limitations for supplementary uploads, the complete materials are provided via the following external link: Supplementary Materials. We also include an extended discussion on the real-world transferability of the proposed approach.

A Real-World Transferability

The proposed Perception-Guided Self-Supervision (PGS) framework is designed with real-world applicability as a central objective. This section discusses the transferability of our approach to real-world autonomous driving systems from both benchmark and algorithmic perspectives.

A.1 Diversity and Realism of the Benchmark

Although our experiments are conducted in simulation, the Bench2Drive (B2D) benchmark provides unique advantages for evaluating real-world readiness. Unlike open-loop datasets where surrounding agents remain static and independent of the ego vehicle, B2D employs a fully closed-loop evaluation protocol that enables bidirectional interactions between the ego vehicle and its environment. This interactive setup more accurately reflects the operational complexity of real-world driving.

In terms of diversity, B2D encompasses 44 interactive scene categories—such as *ParkingExit*, *HazardAtSideLane*, and *ParkedObstacleTwoWays*—alongside 23 distinct weather conditions, covering a wide spectrum of traffic patterns and perceptual challenges. Moreover, it explicitly includes scenario groups involving traffic lights and stop signs, which directly assess a model’s causal understanding of traffic signals. Such evaluation is critical for diagnosing causal confusion, a common real-world failure mode where the ego vehicle reacts to other agents instead of regulatory cues (e.g., red lights or stop signs).

In contrast, datasets like NAVSIM [5], derived from nuPlan, contain large-scale real-world data but lack interaction-driven evaluation. Their assessment remains semi-closed-loop at best, since surrounding agents do not respond to ego actions. Furthermore, NAVSIM does not explicitly test causal compliance, such as stopping at red lights independently of other agents’ behavior. These limitations constrain its ability to measure models’ causal reasoning, which is a primary focus of our work.

*Equal contribution.

†Corresponding author.

While B2D may exhibit a domain gap in rendering fidelity or sensor modality, its closed-loop protocol better captures the causal and interactive aspects essential for robust real-world deployment.

A.2 Transferability of the PGS Framework

From an algorithmic standpoint, the PGS framework is inherently simulation-agnostic. It does not rely on CARLA-specific APIs or handcrafted signals, but instead leverages intermediate outputs from perception modules—such as lane centerlines and multi-agent trajectory predictions—as auxiliary supervision. These signals are well-established and have been extensively validated across real-world datasets.

For instance, lane topology extraction networks (e.g., HDMapNet [8], CenterLineDet [14]) have achieved state-of-the-art results on nuScenes, while motion forecasting methods (e.g., HiVT [15], MultiPath++ [12], Scene Transformer [10]) perform robustly across nuScenes [1], Waymo Open [11], and Argoverse [2]. Consequently, the supervisory signals used in STPS, MTPS, and NTPS are directly transferable to real-world platforms equipped with HD maps and prediction modules. The planning head, trained to align with perception-driven representations, remains compatible with real-world perception stacks.

A.3 Empirical Evidence from Prior Work

Recent literature further supports the transferability of simulation-trained models to real-world domains. Two key trends have emerged:

- **Open-loop to closed-loop degradation.** Methods that achieve strong open-loop performance (e.g., VAD [7], UniAD [6] on nuScenes) often experience substantial degradation when evaluated in closed-loop settings, necessitating extensive architectural adaptation (e.g., VADv2 [3] for CARLA). This highlights the inadequacy of open-loop metrics in predicting real-world performance.
- **Closed-loop to real-world success.** Conversely, models validated under closed-loop simulation frequently demonstrate successful real-world transfer. Notably, TransFuser [4], originally developed for CARLA, has become a lightweight yet effective baseline in real-world evaluations such as NAVSIM. Its design principles have been adopted and extended in recent state-of-the-art frameworks including GoalFlow [13] and WoTE [9], the latter of which trained on B2D and achieved strong NAVSIM performance without retraining.

These observations collectively indicate that robust closed-loop simulation is a strong predictor of real-world readiness, especially when the model demonstrates sound causal reasoning and interaction awareness.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.
- [2] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019.
- [3] S. Chen, B. Jiang, H. Gao, B. Liao, Q. Xu, Q. Zhang, C. Huang, W. Liu, and X. Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024.
- [4] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022.
- [5] D. Dauner, M. Hallgarten, T. Li, X. Weng, Z. Huang, Z. Yang, H. Li, I. Gilitschenski, B. Ivanovic, M. Pavone, et al. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719, 2024.

- [6] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023.
- [7] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023.
- [8] Q. Li, Y. Wang, Y. Wang, and H. Zhao. Hdmapnet: An online hd map construction and evaluation framework. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4628–4634, 2022. doi: 10.1109/ICRA46639.2022.9812383.
- [9] Y. Li, Y. Wang, Y. Liu, J. He, L. Fan, and Z. Zhang. End-to-end driving with online trajectory evaluation via bev world model. *arXiv preprint arXiv:2504.01941*, 2025.
- [10] J. Ngiam, B. Caine, V. Vasudevan, Z. Zhang, H.-T. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal, et al. Scene transformer: A unified architecture for predicting multiple agent trajectories. *arXiv preprint arXiv:2106.08417*, 2021.
- [11] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.
- [12] B. Varadarajan, A. Hefny, A. Srivastava, K. S. Refaat, N. Nayakanti, A. Cornman, K. Chen, B. Douillard, C. P. Lam, D. Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022.
- [13] Z. Xing, X. Zhang, Y. Hu, B. Jiang, T. He, Q. Zhang, X. Long, and W. Yin. Goalflow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1602–1611, 2025.
- [14] Z. Xu, Y. Liu, Y. Sun, M. Liu, and L. Wang. Centerlinedet: Centerline graph detection for road lanes with vehicle-mounted sensors by transformer for hd map generation. *arXiv preprint arXiv:2209.07734*, 2022.
- [15] Z. Zhou, L. Ye, J. Wang, K. Wu, and K. Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8823–8833, 2022.