Open X-AV: Unifying End-to-End Autonomous Driving Datasets

Long Nguyen[1] Micha Fauth[1] Bernhard Jaeger[1] Daniel Dauner[1]
Maximilian Igl[2] Andreas Geiger[1] Kashyap Chitta[1,2]
[1]University of Tübingen, Tübingen AI Center [2]NVIDIA Research

Abstract

The fragmentation of existing autonomous vehicle (AV) datasets hinders the development of generalizable driving policies that can handle complex and infrequent events. To overcome this, we introduce the Open-X AV (OXAV) repository, an initiative designed to aggregate a wide variety of AV datasets and enable models to learn from these diverse sources. We propose a two-stage training workflow using OXAV: a pre-training phase using perception-focused data, followed by post-training on challenging planning-centric scenarios. Our method DiffusionLTF, a simple end-to-end policy trained on OXAV, ranked second in the 2025 Waymo vision-based end-to-end driving challenge, demonstrating the benefits of diverse, aggregated data.

1. Introduction

Large-scale, diverse datasets have been pivotal for creating capable and generalizable AI systems [1]. While Autonomous Vehicles (AVs) stand to greatly benefit from this, the current landscape of publicly available datasets is highly fragmented. Individual datasets typically prioritize a specific discipline, such as perception data with precise labels [2], synthetic data for simulation [14, 27], or auto-labeled data with large-scale annotations crucial for planning [9, 11]. This specialization, however, limits the ability of end-to-end systems trained on a single dataset to effectively handle complex, rare events [4].

In this work, we aim to bridge this gap by fostering cross-platform, multi-dataset learning. Mirroring similar efforts for robotics [7], we introduce the Open-X AV (OXAV) initiative, aimed at aggregating diverse AV datasets to facilitate research into generalizable autonomous driving models. Specifically, we propose a two-stage training workflow that capitalizes on the distinct strengths of varied data sources. (1) An initial pre-training phase on large-scale perception-focused splits, potentially incorporating vast amounts of synthetic data, and building a strong representational foundation. (2) A post-training phase focusing on curated planning-centric splits that expose the model to challenging and rarely observed driving scenarios.

Figure 1. **Overview of Open X-AV.** Four complementary datasets (top left: CARLA [12]; top middle: WOD-P [24]; top right: NAVSIM [11]; bottom: WOD-E2E [9]), ranging from heavily annotated synthetic data to minimally labeled but challenging real-world scenarios, supporting our multi-stage training methodology.

The OXAV repository, in its initial release, supports joint training on four distinct datasets shown in Fig. 1, intentionally selected to embody this perception-planning complementarity. To validate the efficacy of such aggregated data, we train simple end-to-end driving models on OXAV to participate in the Waymo Vision-based End-to-End Driving Challenge. Our models achieve a high ranking on the official leaderboard, despite using only a ResNet34 backbone [13] and requiring only a single day of training on an A100 GPU. We find the Waymo Open Dataset-Perception split [24] to be particularly effective in the pre-training phase. This result provides initial evidence for the benefits of cross-dataset learning in end-to-end AV development.

2. Open X-AV

2.1. Task

The Waymo Vision-based E2E Driving Challenge is an open-loop benchmark for autonomous driving stacks that involves predicting future vehicle trajectories given sensory and vehicle motion inputs, including: (1) camera images providing 360° environmental perception, (2) historical vehicle states, and (3) a discrete navigational command. Unlike prior open-loop benchmarks, the dataset used for this challenge was curated to contain long-tail driving scenar-

Dataset	Input Modality		Output Modality					Size	Curation
	RGB	LiDAR	Trajectories	BEV Semantics	Bounding Boxes	Semantics	Depth	(Hours)	
CARLA [28]	✓	✓	✓	✓	✓	✓	✓	42	Hand-crafted scenarios
NAVSIM [11]	✓	✓	✓	✓	✓			30*	Constant velocity filtering
WOD-P [24]	✓	✓	✓	✓	✓			4	Geographic coverage
WOD-E2E [9]	✓		✓					12	Long-tail sampling

Table 1. **Datasets.** CARLA, NAVSIM, and WOD-P provide diverse input and output modalities, whereas WOD-E2E provides curated, challenging scenarios. *We support the `navtrain` split, which contains around 25% of the 120 hours available in OpenScene [8].

ios. Participating teams must predict 5-second future waypoints in bird's-eye-view coordinates which are scored with the new Rater Feedback Score described in Section 3. Our work explores cross-dataset learning to address this task.

2.2. Datasets

Our approach currently leverages four diverse datasets for cross-dataset learning, as summarized in Table 1.

Waymo Open Dataset - Vision-Based E2E (WOD-E2E) [9]: is the main dataset considered in the challenge evaluation. WOD-E2E comprises 4,021 curated 20-second segments that specifically target long-tail events occurring with less than 0.003% frequency in daily driving.

CARLA [12]: is a simulator that can provide synthetic training data generated using the camera calibration parameters from WOD-E2E. We use PDM-Lite [22] as our expert policy to collect driving demonstrations across Town01-10, 12, and 13, following the procedure of [27, 28]. In addition to trajectory labels, we also collect semantic segmentation, depth maps, bird's eye view (BEV) maps, and bounding boxes as auxiliary supervision.

NAVSIM [11]: is constructed from OpenScene [8], a redistribution of the nuPlan dataset [15] containing 120 hours of real-world driving data sampled at 2Hz. We utilize navtrain, a filtered subset of 103k challenging samples that removes trivial driving scenarios where simple baselines achieve high performance. NAVSIM provides BEV semantic maps and bounding boxes as additional supervision.

Waymo Open Dataset - Perception (WOD-P) [24]: provides 3D bounding boxes and HD maps. We convert the latter to BEV semantic segmentation masks to match the format of [5]. The dataset also benefits from having similar camera calibration parameters to WOD-E2E.

To this end, all datasets provide 3 front-facing RGB camera images and planning trajectories. Auxiliary modalities differ across datasets (see Table 1). To partially mitigate the differences of camera calibration when unifying data for training, we crop images to align focal sizes. Camera height differences are mitigated by padding images with constant values. However, differences in other camera parameters, distortion coefficients, and mounting positions remain, which may affect cross-dataset generalization.

2.3. Baselines

To handle the multimodality of the challenging test data, we consider several baseline methods in our experiments.

Latent TransFuser (LTF): serves as our first baseline architecture. The original TransFuser model fuses a LiDAR BEV representation with a RGB perspective image [5, 14, 21], whereas LTF simply replaces the LiDAR input with a constant. This effectively transforms the LiDAR branch into learnable queries for the transformer-based sensor fusion, enabling vision-only end-to-end driving.

In addition to the 3 concatenated front-facing camera images, LTF incorporates vehicle status inputs including navigational commands. Our version for the WOD-E2E challenge also included past vehicle speeds, and past vehicle positions to provide temporal context and directional guidance. These status inputs are processed through dedicated embedding layers and concatenated with the BEV tokens before serving as context tokens in the transformer decoder.

Diffusion Latent TransFuser (DiffusionLTF): extends the LTF baseline by incorporating a diffusion-based trajectory generation head. Following conceptual ideas of DiffusionDrive [17] and SmartRefine [26], we employ a truncated diffusion schedule which we find effective for diverse trajectory generation while maintaining fast inference. The architecture adopts an encoder-decoder transformer framework where each waypoint is treated as an individual query, enabling fine-grained reasoning for trajectory refinement.

As shown in Fig. 2, this approach uses a discrete vocabulary of representative driving patterns derived from the training data. During training, we apply Gaussian noise to these trajectory prototypes and identify the nearest corrupted trajectory to the ground truth as the initialization for the denoising process. The model learns to reconstruct clean trajectories by minimizing the error between denoised and ground-truth trajectory.

To select a denoised proposal, we use a classification head to score each candidate, optimized with a cross-entropy loss. Let $\hat{T}$ be the ground truth, T_i denotes the i-th denoised candidate and $||\hat{T} - T_i||$ is the denoising error of a candidate, the target distribution is defined as:

$$y_i = \frac{\exp(-||\hat{T} - T_i||)}{\sum_{j=1}^{k} \exp(-||\hat{T} - T_j||)} \quad (1)$$

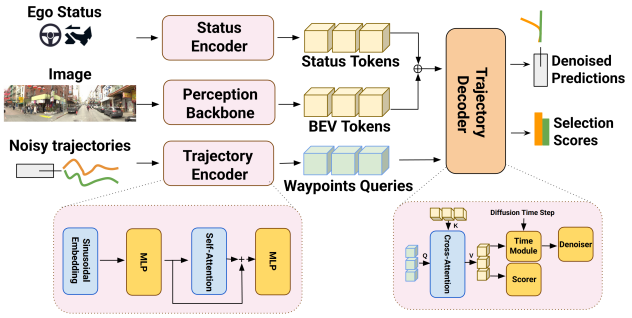

Figure 2. **DiffusionLTF architecture.**

3. Experiments

In context of the challenge, the WOD-E2E dataset splits into 2,037 training segments with ground truth trajectories, 479 validation segments containing both ground truth trajectories and single-frame rater feedback annotations, and 1,505 test segments without any annotations for competition evaluation. For submission, participants submit 1,505 predicted trajectories for the final frame of each test segment. In our experiments, we report the model's performance on the validation set, if not indicated otherwise.

The challenge employs the **Rater Feedback Score (RFS)** that evaluates predicted trajectories against human expert judgment. Three expert raters annotate reference trajectories with quality scores from 0 (poor driving) to 10 (excellent driving). Scoring relies on trust regions around acceptable driving behaviors, defined by lateral and longitudinal thresholds at evaluation timestamps T=3s and T=5s. To receive an expert's score, predicted trajectories must fall within both trust regions of an annotated trajectory; otherwise, they receive exponentially decreasing penalties based on distance to the nearest reference, with a minimum floor of 4. As a secondary metric, we also report **Final Displacement Error (FDE)**, measured as the L2 distance between the predicted and ground truth trajectory endpoints at T=5s.

Implementation Details: Our LTF architecture employs ResNet18 for the LiDAR branch and ResNet34 for the image branch (for more details, see [5] and [14]). The image inputs are resized to size 768×288. Training follows a cosine learning rate schedule [18] with gradient norm clipping at 1.5, comprising 128k gradient steps for pre-training followed by 72k steps for post-training, with AdamW [19] and a batch size of 64. Since auxiliary annotations differ between datasets, we learn auxiliary labels with independent heads for each dataset. DiffusionLTF employs a transformer trajectory encoder (2 layers) and decoder (6 layers) using 256-dimensional tokens [25]. We use DDIM sampling [23] with a squared cosine noise schedule [20]. The trajectory vocabulary is clustered with furthest point sampling on WOD-E2E. As for the denoising objective, we sum up average and final error with Smooth-L1 loss function.

	FDE ↓	RFS ↑
Planning decoder		
LTF [5]	5.77	7.91
DiffusionLTF	5.80	7.88
Proposal ensembling		
N=1	5.65	7.95
N=2	5.65	7.96
N=5	**5.63**	**7.97**
N=10	**5.63**	**7.97**
Logged Trajectories	0.00	8.10

Table 2. **Model architecture ablation on WOD-E2E validation set.** The base model is our best DiffusionLTF checkpoint. While adding the diffusion decoder to LTF alone does not improve the score, it enables ensembling which leads to an overall boost.

3.1. Model Analysis

For our final test-set submissions, we train both the LTF deterministic waypoint decoder and DiffusionLTF waypoint decoder with a shared backbone. A performance overview of both planners can be found in Table 2. We see that a hybrid ensemble approach significantly outperforms individual model components. In particular, we first draw 10 proposals from DiffusionLTF. These 10 proposals might belong to different modes and are grouped accordingly. Within each mode group, we average the trajectories and sum their corresponding selection logits. The averaged trajectory of the mode group with the highest combined logit score is averaged, again, with the LTF deterministic planning output to produce the final prediction.

3.2. Dataset Mixture Analysis

As a next step, we examine the effects of mixing diverse data sources during training, investigating how different dataset combinations impact model performance across both perception pre-training and post-training phases. To ensure fair comparison, all experiments maintain identical total gradient steps while systematically varying data composition. During pre-training, we balance batches with equal proportions from all included datasets, while post-training batches maintain at least 30% WOD-E2E data to preserve target domain representation. Given the stochastic nature of training, we generate a single seed for pre-training and three seeds for post-training, reporting metrics from the checkpoint achieving the highest RFS score per seed.

Pre-training mixture: In this experiment (Table 3), we pre-train the backbones on a mixed of auxiliary labels coming from different sets of datasets, then post-train the model solely on WOD-E2E's logged trajectories. The results demonstrate that pre-training with auxiliary datasets consistently improves model performance compared to training solely on WOD-E2E data. While any individual dataset

Pre-training Datasets			Metrics	
CARLA	NAVSIM	WOD-P	FDE ↓	RFS ↑
			6.33 ± 0.46	7.52 ± 0.01
✓			6.02 ± 0.14	7.75 ± 0.02
	✓		5.94 ± 0.16	7.81 ± 0.03
		✓	5.82 ± 0.10	**7.85 ± 0.04**
✓	✓		5.74 ± 0.17	7.83 ± 0.04
✓		✓	5.84 ± 0.20	7.74 ± 0.03
	✓	✓	5.90 ± 0.09	7.81 ± 0.04
✓	✓	✓	**5.73 ± 0.21**	7.84 ± 0.01

Table 3. **Impact of pre-training with diverse data.**

provides meaningful improvements, WOD-P emerges as the most effective single pre-training dataset, achieving the highest RFS score of 7.85. The combination of all three datasets yields the lowest FDE of 5.73 but does not surpass the RFS performance of WOD-P alone. This suggests that while diverse pre-training data helps with trajectory accuracy, the quality and relevance of individual datasets may be more critical for expert raters' satisfaction than simply maximizing dataset diversity.

Post-training mixture: The next experiments in Table 4 share a backbone pre-trained on all datasets. They differ by which datasets' auxiliary labels are used in the post-training step to fine-tune the backbone parallel to the training of the planning heads on WOD-E2E. Here, we increase the batch size from 64 to 128 to control the noise level of gradients of the planning heads, since only about half of the batch actually has waypoint labels of WOD-E2E and the other half has only auxiliary labels. Following the pre-training analysis, the post-training mixture results reveal a contrasting pattern. Unlike pre-training where auxiliary datasets consistently provided benefits, post-training with additional datasets generally degrades performance compared to training exclusively on WOD-E2E data. The baseline approach with no additional datasets achieves the highest RFS score of 7.84, while most auxiliary dataset combinations reduce

Additional Post-training Datasets			Metrics	
CARLA	NAVSIM	WOD-P	FDE ↓	RFS ↑
			5.73 ± 0.21	**7.84** ± 0.01
✓			6.25 ± 0.19	7.76 ± 0.12
	✓		6.17 ± 0.05	7.77 ± 0.04
		✓	5.82 ± 0.21	7.83 ± 0.03
✓	✓		5.91 ± 0.20	7.81 ± 0.07
✓		✓	**5.65** ± 0.18	7.80 ± 0.05
	✓	✓	5.88 ± 0.10	7.80 ± 0.04
✓	✓	✓	5.74 ± 0.08	**7.84** ± 0.07

Table 4. **Post-training with joint planning and perception.**

#	Submission	Val RFS ↑	Test RFS ↑
1	Post-train WOD-E2E & CARLA	7.42	7.19
2	Post-train WOD-E2E only	7.54	7.42
3	Pre- and post-train all datasets	7.86	7.54
4	Post-train validation set	**8.20**	7.49
5	Add past states to status tokens	7.90	7.59
6	Pre-train WOD-P only	7.97	**7.71**

Table 5. **Results on the WOD-E2E test server.**

both FDE and RFS performance. This preliminary result indicates the importance of curated planning-centric data in the post-training phase.

Test-set performance: To verify models' performance against the test set, each participating team was allowed to submit 6 predictions to the leaderboard per month. A summary of our submissions is shown in Table 5.

The first and second submission shared the same pre-trained backbone trained on CARLA only, differing in that the first submission is also post-trained on CARLA's logged trajectories. To enable this, we defined a learned mapping from CARLA's GPS target points to WOD-E2E's navigational commands (e.g., turn left, go straight, turn right). The CARLA trajectory post-training did not improve performance, and was excluded in our main experiments.

The third submission used all data for pre-training and post-training, and the fourth submission used in addition the validation set of WOD-E2E for training. However, we observed strong fluctuation of validation performance between each epoch, so the validation set was excluded from training in subsequent experiments for model selection purposes.

The fifth submission added past speeds and past positions as status tokens to the DiffusionLTF transformer decoder, giving a minor improvement. The final submission also included past statess, but used only WOD-P as an extra pre-training source, motivated by our results in Table 3.

4. Conclusion

Our work demonstrates that cross-dataset learning improves end-to-end driving performance, with models pre-trained on the diverse OXAV collection achieving strong results on the WOD-E2E benchmark. More importantly, our analysis reveals that while additional datasets consistently provide benefits, quality trumps quantity when selecting data sources. Individual high-quality datasets like WOD-P can outperform complex dataset combinations, and the manner of dataset integration matters. Auxiliary data proves benefits during pre-training, but can degrade performance during post-training phases. These findings suggest that strategic dataset selection and careful integration into training are as important as dataset diversity itself for developing more generalizable and capable end-to-end driving systems.

Limitations: This work relies on open-loop metrics to evaluate the performance of autonomous driving stacks. These metrics have sometimes been misleading since they do not measure closed-loop driving performance [6, 10, 16]. In particular, there is no analysis yet whether RFS correlates with closed-loop performance or not. The WOD-E2E dataset does not provide maps or bounding box labels, which is why we could not use the more reliable but label-dependent PDM-score open-loop metric [11]. For stronger conclusions, evaluation with multiple benchmkarks and metrics may be necessary [3, 11, 12].

Acknowledgments

Bernhard Jaeger and Andreas Geiger were supported by the ERC Starting Grant LEGO-3D (850533) and the DFG EXC number 2064/1 - project number 390727645. Daniel Dauner was supported by the German Federal Ministry for Economic Affairs and Climate Action within the project NXT GEN AI METHODS (19A23014S). We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Bernhard Jaeger, Daniel Dauner, and Kashyap Chitta. This research used compute resources at the Tübingen Machine Learning Cloud, DFG FKZ INST 37/1057-1 FUGG as well as the Training Center for Machine Learning (TCML).

References

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 1

[3] Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Pseudo-simulation for autonomous driving. *arXiv*, 2506.04218, 2025. 5

[4] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. 2024. 1

[5] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *Pattern Analysis and Machine Intelligence (PAMI)*, 2023. 2, 3

[6] Felipe Codevilla, Antonio M. López, Vladlen Koltun, and Alexey Dosovitskiy. On offline evaluation of vision-based driving models. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018. 5

[7] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, and Ajay Mandlekar et al. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023. 1

[8] OpenScene Contributors. Openscene: The largest up-to-date 3d occupancy prediction benchmark in autonomous driving. https://github.com/OpenDriveLab/OpenScene, 2023. 2

[9] Waymo E2E Contributors. Waymo open dataset: Vision-based end-to-end driving. https://waymo.com/open/data/e2e, 2025. 1, 2

[10] Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with misconceptions about learning-based vehicle motion planning. In *Conference on Robot Learning (CoRL)*, 2023. 5

[11] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1, 2, 5

[12] 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. 1, 2, 5

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 1

[14] Bernhard Jaeger, Kashyap Chitta, and Andreas Geiger. Hidden biases of end-to-end driving models. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2023. 1, 2, 3

[15] Napat Karnchanachari, Dimitris Geromichalos, Kok Seang Tan, Nanxiang Li, Christopher Eriksen, Shakiba Yaghoubi, Noushin Mehdipour, Gianmarco Bernasconi, Whye Kit Fong, Yiluan Guo, and Holger Caesar. Towards learning-based planning: The nuplan benchmark for real-world autonomous driving. In *IEEE International Conference on Robotics and Automation, ICRA*, 2024. 2

[16] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and José M. Álvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[17] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. *arXiv preprint arXiv:2411.15139*, 2024. 2

[18] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. 3

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 3

[20] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. 3

[21] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[22] 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 *Proc. of the European Conf. on Computer Vision (ECCV)*, 2024. 2

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 3

[24] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, and Chai et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 3

[26] Yang Zhou, Hao Shao, Letian Wang, Steven L. Waslander, Hongsheng Li, and Yu Liu. Smartrefine: A scenario-adaptive refinement framework for efficient motion prediction, 2024. 2

[27] Julian Zimmerlin. Tackling carla leaderboard 2.0 with end-to-end imitation learning. Master's thesis, University of Tübingen, 2024. 1, 2

[28] Julian Zimmerlin, Jens Beißwenger, Bernhard Jaeger, Andreas Geiger, and Kashyap Chitta. Hidden biases of end-to-end driving datasets. *arXiv.org*, 2412.09602, 2024. 2