Supplementary Material for DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving

Bencheng Liao[1,2,◇] Shaoyu Chen[2,3] Haoran Yin[3] Bo Jiang[2,◇] Cheng Wang[1,2,◇] Sixu Yan[2]
Xinbang Zhang[3] Xiangyu Li[3] Ying Zhang[3] Qian Zhang[3] Xinggang Wang[2] ✉

[1] Institute of Artificial Intelligence, Huazhong University of Science & Technology
[2] School of EIC, Huazhong University of Science & Technology
[3] Horizon Robotics

Code & Model & Demo: hustvl/DiffusionDrive

A. Further Implementation Detail

We provide additional implementation details for our method on the NAVSIM [3] and nuScenes [1] datasets.

NAVSIM Dataset. We initialize the ResNet-34 [5] backbone with ImageNet pre-trained weights, and the LiDAR range is 32m to the front, back, left, and right following Transfuser baseline [3]. We also perform auxiliary perception tasks following Transfuser baseline [3], which include 3D object detection, 2D BEV semantic segmentation. The object queries and BEV features are taken as input of the proposed diffusion decoder.

nuScenes Dataset. We follow the SparseDrive baseline [6] to perform two-stage training. The model is directly initialized with the stage-1 pre-trained weight, which is trained solely on perception tasks (3D object detection/tracking, vectorized HD map construction, and motion prediction) and provided by the official open-source implementation. We train the stage-2 model on the nuScenes dataset for 10 epochs, replacing the planning module of SparseDrive with our proposed diffusion decoder and truncated diffusion mechanism. Object queries, map queries, and PV features are taken as inputs to the diffusion decoder.

B. Further Ablation Study

Train	Infer	NC↑	DAC↑	TTC↑	Comf.↑	EP↑	PDMS↑
Anchored Dist.	Anchored. Dist.	98.2	96.2	94.7	100	82.2	88.1
	Extra. Traj.	96.3	91.7	90.4	100	76.8	81.3
Extra. Traj.	Extra. Traj.	97.3	94.0	92.6	100	79.6	84.7

Table 1. **Comparison on driving priors.** "Anchored Dist." denotes anchored Gaussian distribution. "Extra. Traj." denotes extrapolated trajectory based on current status. Row-1 marked in blue denotes the DiffusioDrive baseline of the main paper.

◇ Intern of Horizon Robotics; ✉ Corresponding author: Xinggang Wang (xgwang@hust.edu.cn).

Comparison on driving priors. In Tab. 1, we validate the superiority of prior anchors over the prior extrapolated trajectory based on the current status. Row-1 is DiffusionDrive baseline. Row-2 uses the DiffusionDrive baseline model to infer from an extrapolated trajectory instead of sampled N_{infer} trajectories. Row-3 represents DiffusionDrive trained with a single anchor, *i.e.*, the extrapolated trajectory, and infers by sampling around it. The results demonstrate the superiority of the proposed anchored Gaussian distribution over extrapolated prior, which fails to cover the potential action space and can not effectively handle challenging scenarios (*e.g.*, obstacle avoidance and turning) in real-world application (consistent with comparisons to ego-status-based planners in Tab. 1 of NAVSIM paper [3]).

Method	Anchor Source	DS↑	RC↑	IS↑
Transfuser†	-	$47.30_{\pm 5.72}$	$93.38_{\pm 1.20}$	$0.50_{\pm 0.06}$
DiffusionDrive	NAVSIM	$\mathbf{64.27}_{\pm 2.43}$	$\mathbf{94.16}_{\pm 1.46}$	$\mathbf{0.69}_{\pm 0.02}$

Table 2. **Generalization of anchor source.** We test DiffusionDrive on Carla Longest6 benchmark with clustered anchors from NAVSIM dataset. † denotes that the result is taken from Transfuser paper [2].

Generalization of anchor source. To further investigate the generalization of anchor source, we train DiffusionDrive on CARLA [4] with NAVSIM-clustered anchored Gaussian distribution (Row-2 in Tab. 2). Since the CARLA dataset is totally different from NAVSIM, the superior results validate the generalization capability of our anchored Gaussian distribution, which is designed to cover potential multi-mode driving action space instead of train/val information leakage.

C. Further Qualitative Comparison

In this section, we provide additional qualitative comparisons on challenging scenarios from the planning-oriented NAVSIM dataset `navtest` split [3].

Going straight. Fig. 1a and Fig. 1b show that the top-1 scoring trajectories of DiffusionDrive are similar to the ground truth trajectories, while the highlighted top-10 scoring trajectories can perform robust lane changes. Notably, Fig. 1c demonstrates that the diverse and highlighted top-10 trajectories can further recognize the traffic light, enabling reasonable lane changes and stopping at the stop line.

Turning left. Fig. 2 shows that the denoised diverse trajectories are dynamically adjusted based on the traffic conditions. The highlighted top-10 scoring trajectories are robust and reasonable, effectively performing lane changes.

Turning right. Fig. 3a and Fig. 3b show that the top-1 scoring trajectories of DiffusionDrive are going to perform car-following like the ground truth trajectories, while the highlighted top-10 scoring trajectories tend to overtake the leading vehicle. These results validate that DiffusionDrive can robustly generate diverse and plausible driving actions.

References

[1] 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

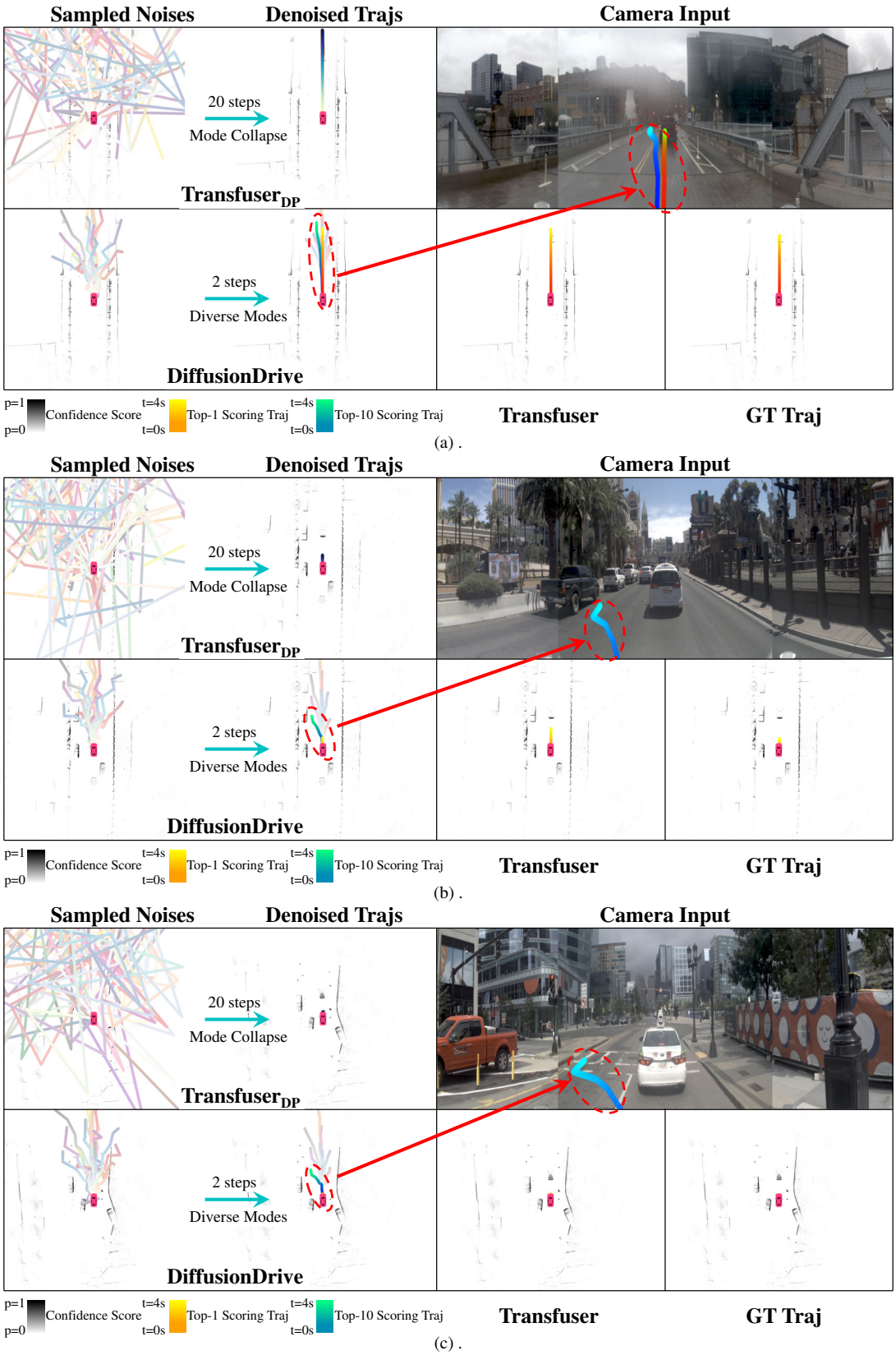
[2] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *TPAMI*, 2022. 1

[3] 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 *NeurIPS*, 2024. 1

[4] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[6] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*, 2024. 1



(a) .

(b) .

(c) .

Figure 1. **Qualitative comparison of Transfuser, Transfuser$_{DP}$ and DiffusionDrive on going straight scenarios of NAVSIM `navtest` split.**

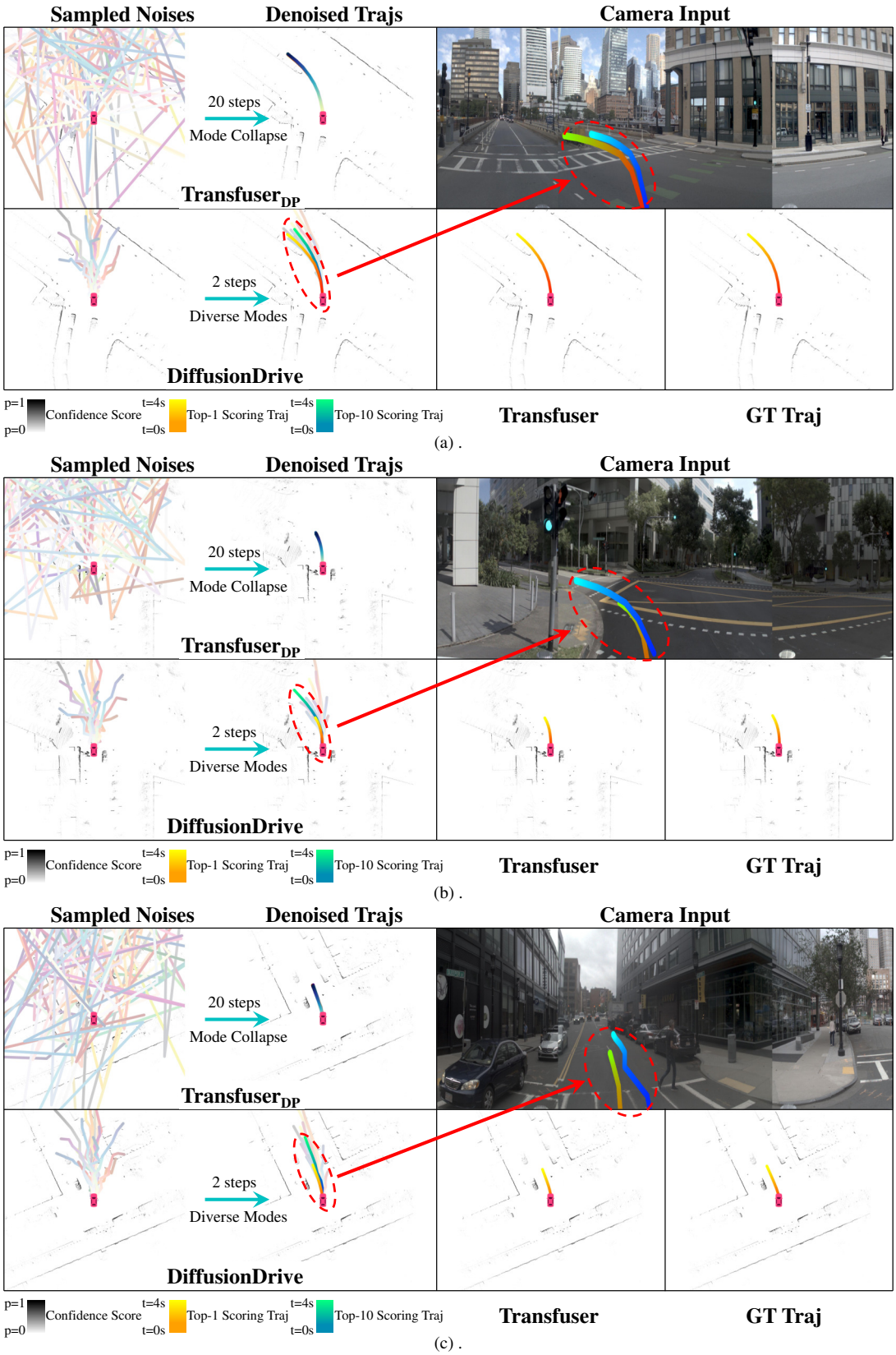


Figure 2. **Qualitative comparison of Transfuser, Transfuser$_{DP}$ and DiffusionDrive on turning left scenarios of NAVSIM `navtest` split.**

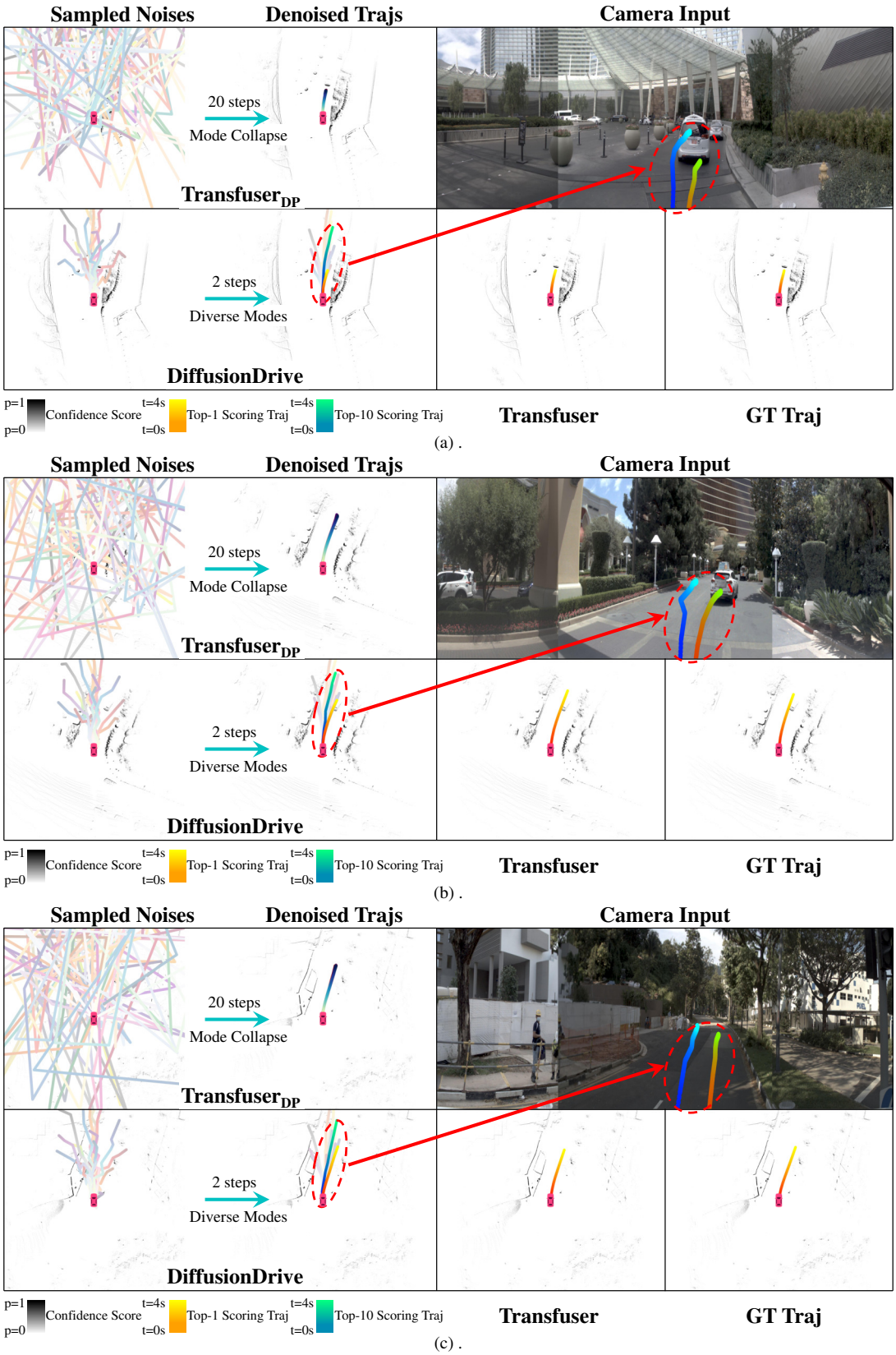


Figure 3. **Qualitative comparison of Transfuser, Transfuser$_{DP}$ and DiffusionDrive on turning right scenarios of NAVSIM `navtest` split.**