

1 Appendix

2 A Implementation Details

3 Our model is trained on the official NAVSIM [1] training set Navtrain, and evaluated on the official
4 test set Navtest. We follow the same perception setup and ResNet-34 backbone used in Transfuser for
5 a fair comparison. Specifically, we use a concatenated front-view image of size 1024×256 formed
6 by three forward-facing cameras as the visual input, fused with a 64×64 BEV LiDAR feature map.
7 In addition, the model receives a state vector consisting of the current vehicle speed, acceleration,
8 and navigation information. The size of the anchor vocabulary is set to $N = 8192$. We use fixed
9 weights $w_1 = 0.1$ and $w_2 = 1.0$ for unified policy distillation. The number of frequency bands in the
10 positional encoding is set to $L = 10$. The predefined safety threshold τ is set to 0.3. All experiments
11 are conducted on 6 NVIDIA L20 GPUs, with a batch size of 16 per GPU. We use the AdamW
12 optimizer [2] with a learning rate of $1e-4$. The model is first trained for 30 epochs using unified
13 policy distillation, followed by 10 epochs of fine-tuning with Safety DPO. We sample $K = 1024$
14 trajectories from the policy distribution for each DPO iteration and set the temperature parameter
15 to $\beta = 0.1$. To accelerate training, the PDMS for each anchor in every scene is precomputed in
16 advance. Following Transfuser [3], we adopt a multi-task learning strategy to enhance perception
17 capabilities, jointly training two auxiliary tasks: 3D object detection and BEV semantic segmentation.
18 Additionally, to further benefit policy learning, we introduce an auxiliary loss to predict the outputs
19 of each rule-based teacher. In DPO training, inspired by [4], we introduce an explicit reverse KL
20 regularization term to suppress distributional drift during training. Since the reference distribution
21 π_{Ref} targets the unified distribution p_{unified} , we directly substitute π_{Ref} with p_{unified} in practice.
22 Finally, similar to [5], we continue applying the KL loss from unified policy distillation during the
23 DPO fine-tuning stage as an auxiliary loss.

24 B Further Ablation Study

25 **Ablation on the Number of Sampled Candidate Trajectories.** To validate the choice of the
26 candidate trajectory number K in Safety DPO, we conduct an ablation study as shown in Table 1. We
27 observe that increasing K from 64 to 1024 leads to steady performance improvements. This confirms
28 that a larger candidate pool enables the selection of more representative and challenging preference
29 pairs, enhancing the effectiveness of DPO. However, performance drops when K increases to 2048,
30 likely due to a lower gradient signal from very low-probability rejected trajectories. We therefore set
31 $K = 1024$ as the default choice.

32 **Ablation on the Number of DPO Fine-Tuning Epochs.** We further investigate how the number of
33 DPO fine-tuning epochs affects policy performance, as shown in Table 2. With 5 epochs of DPO, the
34 policy already improves over the baseline. Further gains are achieved when increasing to 10 epochs,
35 reaching the highest PDMS score of 90.0. However, continuing DPO fine-tuning to 20 epochs causes
36 performance degradation, likely due to overfitting. Thus, we set the DPO fine-tuning duration to 10
37 epochs in all experiments.

38 **Ablation on the DPO Start Epoch.** We also study when to initiate DPO fine-tuning best. As
39 shown in Table 3, enabling DPO too early (e.g., at epoch 10) may destabilize training since the policy
40 has not yet converged under supervised objectives. Starting DPO at epoch 30 achieves the best result,
41 leveraging a well-initialized policy and intense pretraining supervision. Delaying DPO to epoch 40
42 slightly reduces performance, suggesting potential overfitting. We therefore adopt epoch 30 as the
43 default DPO start point.

44 C Limitations and Future Works

45 Despite the significant progress our method has made in enhancing the safety of end-to-end driving
46 policies, several limitations warrant further study. First, our approach relies on the PDMS metric as
47 the core criterion for safety evaluation. Although PDMS integrates multiple dimensions, including
48 collision avoidance, drivable area compliance, ego progress, time-to-collision, and comfort, it remains
49 a predefined weighted composite metric. As such, it cannot fully capture all potential risk factors

Table 1: **Ablation on the number of sampled candidate trajectories K .** The best result is highlighted in **bold**.

K	NC \uparrow	DAC \uparrow	EP \uparrow	TTC \uparrow	C \uparrow	PDMS \uparrow
64	97.9	97.7	83.9	93.3	100.0	89.0
256	97.9	97.8	84.0	93.6	100.0	89.2
512	98.1	98.2	83.7	94.2	100.0	89.5
1024	98.5	98.1	84.3	94.8	99.9	90.0
2048	98.0	98.0	83.8	93.8	99.9	89.3

Table 2: **Ablation on the number of DPO fine-tuning epochs.** The best result is highlighted in **bold**.

Epoch	NC \uparrow	DAC \uparrow	EP \uparrow	TTC \uparrow	C \uparrow	PDMS \uparrow
5	98.3	97.6	84.9	93.7	99.9	89.5
10	98.5	98.1	84.3	94.8	99.9	90.0
20	98.0	97.4	83.9	93.8	99.9	89.0

Table 3: **Ablation on the starting epoch for DPO fine-tuning.** The best result is highlighted in **bold**.

Epoch	NC \uparrow	DAC \uparrow	EP \uparrow	TTC \uparrow	C \uparrow	PDMS \uparrow
10	98.1	97.5	83.7	94.3	100.0	89.2
20	98.1	98.0	84.4	94.2	100.0	89.7
30	98.5	98.1	84.3	94.8	99.9	90.0
40	98.2	98.0	84.2	94.0	100.0	89.5

in complex driving scenarios. Future work may explore more expressive and flexible trajectory evaluation metrics to build a more comprehensive safety assessment mechanism. Second, our rule-based supervision depends on high-fidelity simulators to provide rule-driven evaluation scores. While effective, the preferences derived from such simulation are inherently limited by the rules’ design and the simulator’s precision. Moreover, access to such high-fidelity simulators is scarce, constraining the scale and diversity of available data. Therefore, future research should explore preference optimization methods that do not rely on ground-truth perception labels by automatically mining latent preference structures from historical trajectory data. Such efforts could facilitate the development of weakly-supervised or even fully self-supervised safety alignment strategies.

D Additional Qualitative Comparison

More Comparisons with Transfuser and Hydra-MDP. Figure 1 presents additional qualitative comparisons between our method and two baselines: Transfuser [3] and Hydra-MDP [6]. Across various challenging traffic scenes, both baselines exhibit unsafe behaviors such as collisions or off-road deviations. In contrast, our method consistently produces safer trajectories.

More Visualizations of DPO-Enhanced Policy Behaviors. Figure 2 shows more examples comparing policies before and after DPO fine-tuning. After applying Safety DPO, the policy demonstrates safer and more cautious behaviors.

References

- [1] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, et al. Navsim: Data-driven non-

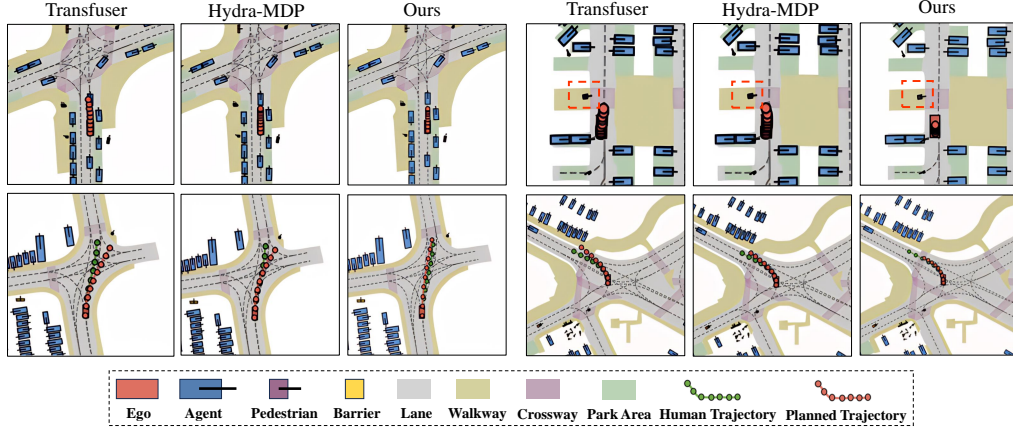


Figure 1: Additional qualitative comparisons with baselines.

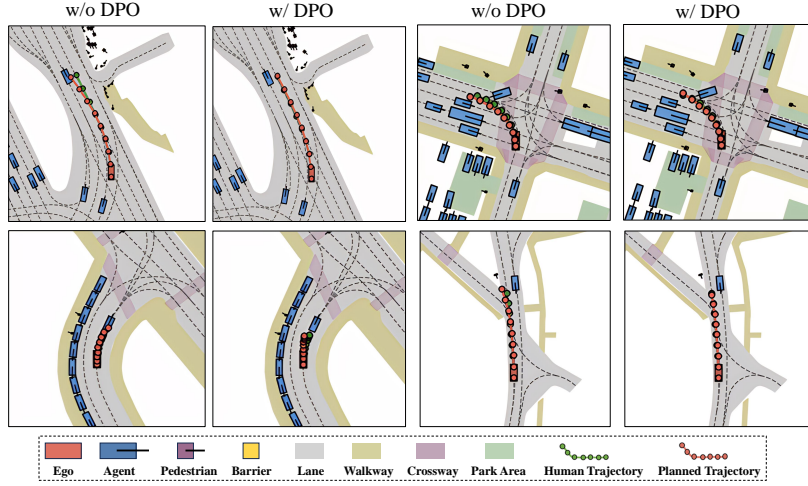


Figure 2: Additional qualitative visualizations of DPO-enhanced policy behaviors.

- 70 reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information*
71 *Processing Systems*, volume 37, pages 28706–28719, 2024.
- 72 [2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International*
73 *Conference on Learning Representations*, 2019.
- 74 [3] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger.
75 Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE*
76 *transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022.
- 77 [4] Xiong Wei, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong
78 Zhang. Iterative preference learning from human feedback: bridging theory and practice for rlhf
79 under kl-constraint. In *Proceedings of the 41st International Conference on Machine Learning*,
80 pages 54715–54754, 2024.
- 81 [5] Yiren Lu, Justin Fu, George Tucker, Xinlei Pan, Eli Bronstein, Rebecca Roelofs, Benjamin
82 Sapp, Brandyn White, Aleksandra Faust, Shimon Whiteson, et al. Imitation is not enough:
83 Robustifying imitation with reinforcement learning for challenging driving scenarios. In *2023*
84 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7553–7560.
85 IEEE, 2023.

- 86 [6] Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu,
87 Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target
88 hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024.