Supplementary Materials for SafeDrive

In these Supplementary Materials, we provide additional details that could not be included in the main paper due to space limitations. They include the following:

- Detailed Evaluation Metrics
- Additional Implementation Details
- Extensive Experimental Analysis
- Comprehensive Qualitative Results

A. Detailed Evaluation Metrics

A.1. NAVSIM Dataset

PDM Score (PDMS). PDMS evaluates driving performance using safety, progress, and comfort metrics computed in a non-reactive simulation of the predicted trajectory. Safety-critical violations such as collisions and lane departures yield a zero score through multiplicative penalties, and all remaining subscores are aggregated by a weighted average. The metrics and their weights are summarized in Table 1, and PDMS is defined as

$$\text{PDMS} = \left(\prod_{m \in \mathcal{P}} s_m \right) \left(\frac{\sum_{m \in \mathcal{W}} w_m s_m}{\sum_{m \in \mathcal{W}} w_m} \right), \quad (1)$$

where $\mathcal{P} = \{\text{NC}, \text{DAC}\}$ and $\mathcal{W} = \{\text{EP}, \text{TTC}, \text{C}\}$ denote the penalty and weighted metric sets, respectively, and s_m and w_m denote the score and weight of each metric. PDMS is computed per frame and averaged over the full sequence.

Extended PDM Score (EPDMS). PDMS assigns penalties even when the human driver also violates a rule and does not consider rule-based behaviors such as traffic-light compliance. EPDMS addresses these limitations by filtering out human-caused violations and introducing additional evaluation metrics. Newly added metrics are listed in Table 1, and all others except Comfort are inherited from PDMS. EPDMS is defined as

$$\tilde{s}_m = \begin{cases} 1 & s_m^{\text{human}} = 0 \\ s_m^{\text{pred}} & \text{otherwise} \end{cases} \quad (2)$$

$$\text{EPDMS} = \left(\prod_{m \in \mathcal{P}_{\text{ext}}} \tilde{s}_m \right) \left(\frac{\sum_{m \in \mathcal{W}_{\text{ext}}} w_m \tilde{s}_m}{\sum_{m \in \mathcal{W}_{\text{ext}}} w_m} \right), \quad (3)$$

where s_m^{pred} and s_m^{human} denote the values computed from the predicted and human trajectories, respectively. The sets $\mathcal{P}_{\text{ext}} = \{\text{NC}, \text{DAC}, \text{DDC}, \text{TLC}\}$ and $\mathcal{W}_{\text{ext}} = \{\text{EP}, \text{TTC}, \text{LK}, \text{HC}, \text{EC}\}$ represent the penalty and weighted metrics.

Table 1. Metrics used in PDMS and EPDMS. "*" indicates metrics used only in PDMS.

Metric	EPDMS only	Weight	Range
No at-fault Collisions (NC)		multiplier	{0, 1/2, 1}
Drivable Area Compliance (DAC)		multiplier	{0, 1}
Driving Direction Compliance (DDC)	✓	multiplier	{0, 1/2, 1}
Traffic Light Compliance (TLC)	✓	multiplier	{0, 1}
Ego Progress (EP)		5	[0, 1]
Time to Collision (TTC) within bound		5	{0, 1}
Comfort (C)*		2	{0, 1}
Lane Keeping (LK)	✓	2	{0, 1}
History Comfort (HC)	✓	2	{0, 1}
Extended Comfort (EC)	✓	2	{0, 1}

A.2. Bench2Drive Benchmark

Driving Score (DS). DS evaluates the overall driving performance by combining route completion with penalties for safety-related infractions. For each route, the score is obtained by multiplying the route-completion percentage by the infraction penalties defined in Table 2. DS is computed as

$$\text{DS} = \frac{1}{n_{\text{total}}} \sum_{i=1}^{n_{\text{total}}} \text{RC}^{\,i} \cdot \prod_{j=1}^{n_{\text{penalty}}^{i}} p_j^i, \quad (4)$$

where n_{total} is the number of routes, $\text{RC}^{\,i}$ is the route-completion percentage for route i, p_j^i is the penalty factor for the j-th infraction on route i, and n_{penalty}^i is the number of infractions considered for that route.

Table 2. Infraction types and penalty factors used in the DS.

Infraction	Penalty	Note
Pedestrian Collision	0.50	Each occurrence
Vehicle Collision	0.60	Each occurrence
Other Collision	0.65	Each occurrence
Running Red Light	0.70	Each occurrence
Scenario Timeout	0.70	Timeout (4 min)
Too Slow	0.70	Low speed
No Give Way	0.70	Yield failure
Off-road	–	Excluded from RC
Route Deviation	–	Deviation $>$ 30 m
Agent Blocked	–	Idle for 180 s
Route Timeout	–	Max time exceeded

Success Rate (SR). SR measures the proportion of routes that are successfully completed among all evaluation routes. A route is considered successful only if the ego vehicle reaches the goal destination within the time limit without committing any infractions. SR is computed as the ratio between the number of successful routes n_{success} and the total number of evaluation routes n_{total}, as follows:

$$\text{SR} = \frac{n_{\text{success}}}{n_{\text{total}}}. \quad (5)$$

B. Additional Implementation Details

Existing studies [3, 5, 6] typically apply BEV segmentation as an auxiliary task using BEV features extracted from the BEV backbone, covering both dynamic and static classes. Following this convention, our model is trained with BEV segmentation as an auxiliary task on BEV features, while additionally incorporating object detection for SWNet and static BEV segmentation for TwDAC. Prior works generally operate with a perception range of [0, 32] m forward and [-32, 32] m laterally and adopt a 0.25 m resolution for BEV segmentation. To enhance safety by incorporating long-range contextual cues, we extend the forward perception range to [0, 64] m and increase the BEV resolution to 0.125 m, thereby providing richer spatial information. We further stabilize motion prediction by incorporating image and LiDAR inputs from the current frame and two past frames. We also redesign the object detection decoder following the iterative refinement mechanism [4].

C. Extensive Experimental Analysis

Effect of Perception Performance to Planning. Table 3 presents the ablation results comparing different BEV backbones and planning heads. TransFuser [6] differs only in its feature encoding, while sharing the same perception range, object decoder, and related settings as BEVFormer [4]. SafeDrive is slower than DiffusionDrive [5] due to its fine-grained safety reasoning process, yet it consistently achieves higher planning performance across all configurations. Notably, DiffusionDrive shows limited planning improvements even when its perception performance increases, whereas SafeDrive yields a substantial PDMS gain when combined with the BEVFormer backbone. This indicates that perceptual outputs in DiffusionDrive do not directly influence the generated trajectories, while SafeDrive effectively leverages high-fidelity perceptual representations within its safety-critical reasoning process to produce safer trajectories.

Impact of Trajectory Refinement. Table 4 presents the ablation study on trajectory refinement. When the initial anchor trajectories are used without any refinement, the PDMS remains at 88.1. Refining ProposalNet alone raises the score to 90.6, and adding planning-branch refinement in SWNet further improves it to 90.9. Incorporating surrounding-agent motion refinement achieves the highest PDMS of 91.6. These results show that refining the motions of surrounding agents in SWNet more accurately captures instance-centric interactions between the planning query and nearby agents, which directly contributes to generating more stable and safer trajectories.

Table 3. Ablation study on BEV backbone and planning head. "TF", "BF", and "Diff" denote TransFuser, BEVFormer, and DiffusionDrive, respectively.

Backbone	Head	mAP	mIoU	NC	DAC	PDMS	Latency (ms)
TF	Diff	53.2	45.6	98.4	96.2	88.3	40
TF	Ours	51.8	44.7	99.0	97.3	89.6	55
BF	Diff	**91.4**	**56.4**	98.8	96.3	88.6	50
BF	Ours	86.6	54.7	**99.5**	**99.0**	**91.6**	67

Table 4. Ablation study on the impact of trajectory refinement

ProposalNet Refinement	SWNet Refinement Planning	SWNet Refinement Motion	NC	DAC	TTC	PDMS
			98.4	96.9	93.6	88.1
✓			99.1	98.0	96.5	90.6
✓	✓		99.1	98.3	96.7	90.9
✓	✓	✓	**99.5**	**99.0**	**97.2**	**91.6**

Table 5. Comparison of collision-avoidance methods. Conflict Filtering removes trajectories overlapping with predicted agent motions.

Method	NC	DAC	TTC	PDMS
Scene-Level NC	99.2	98.2	96.7	90.9
Conflict Filtering	99.3	98.2	96.8	91.0
Pair-wise NC	**99.5**	**98.7**	**97.3**	**91.5**

Table 6. Ablation of TwDAC Components. Temporal predicts and evaluates drivable-area compliance at each time step. Interpolated enhances compliance checking using bilinear interpolation over the predicted BEV segmentation.

Temporal	Interpolated	NC	DAC	TTC	PDMS
		99.2	98.2	96.7	90.9
✓		99.2	98.7	96.7	91.2
	✓	99.2	98.5	96.6	91.1
✓	✓	**99.2**	**99.0**	**96.7**	**91.4**

Comparison of Collision-Avoidance Methods. As shown in Table 5, we compare three methods for evaluating collision-avoidance capability. Conflict Filtering, which removes trajectories overlapping with predicted agent motions, produces results similar to Scene-Level NC (NC 99.3). Pair-wise NC enhances safety-critical decision making by explicitly modeling interactions between the planning trajectory and nearby agents, improving NC to 99.5 and TTC to 97.3. These gains are notable given the already high baseline and arise from more fine-grained assessment of collision risks.

Ablation Study for TwDAC Table 6 reports the effect of the two TwDAC components. Without either component, the model attains a DAC of 98.2. Adding the Tem-

poral evaluation improves it to 98.7, and using only the interpolation-based evaluation yields a comparable increase to 98.5. Combining both achieves the highest DAC of 99.0. These results indicate that each component enhances drivable-area compliance estimation, and that combining them enables a more fine-grained assessment of drivable-area safety.

D. Comprehensive Qualitative Results

More Visualizations of Fine-grained Reasoning. Figure 1 visualizes the fine-grained safety reasoning process of SafeDrive. For each scenario, the bottom-left panel illustrates the fine-grained safety scores—(a) PwNC and (b) TwDAC—predicted from the simulated future states of each candidate trajectory. The bottom-right panel of each scenario then shows the final trajectory selection guided by these safety signals. In contrast to ProposalNet, which often selects unsafe trajectories due to its scene-level safety estimates, FRNet leverages the fine-grained safety information from PwNC and TwDAC to choose safer trajectories. These precise and interpretable safety cues obtained through the simulation process provide essential guidance for reliable and safe trajectory selection.

Additional Comparison with SOTA Methods. Figures 2 and 3 qualitatively compare SOTA models [3, 5] across diverse scenarios. SafeDrive, in particular, produces safety-oriented trajectories by modeling fine-grained interactions with surrounding agents and the environment.

Visualization of the Bench2Drive Benchmark. Figure 4 illustrates that SafeDrive also achieves robust closed-loop driving performance across diverse and challenging scenarios in the Bench2Drive benchmark.

Failure Case. Figure 5 presents a failure where SafeDrive collides with a pedestrian stepping out of a vehicle. Although the model generates a forward trajectory at time t that does not hit the stopped vehicle, it fails to reason about the pedestrian emerging from the slightly opening door. This reveals that the model struggles to interpret subtle contextual cues associated with human behavior. Addressing such cases may require integrating VLM-based [1, 2, 7, 8] semantic reasoning with our framework to better capture these contextual signals and produce safer motion plans.

References

[1] Xuesong Chen, Linjiang Huang, Tao Ma, Rongyao Fang, Shaoshuai Shi, and Hongsheng Li. Solve: Synergy of language-vision and end-to-end networks for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025*. IEEE, 2025. 3

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 3

[3] Yingyan Li, Yuqi Wang, Yang Liu, Jiawei He, Lue Fan, and Zhaoxiang Zhang. End-to-end driving with online trajectory evaluation via bev world model. *arXiv preprint arXiv:2504.01941*, 2025. 2, 3

[4] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[5] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, et al. Diffusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12037–12047, 2025. 2, 3

[6] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7077–7087, 2021. 2

[7] Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025*. IEEE, 2025. 3

[8] 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 *Proceedings of the European Conference on Computer Vision 2024*, pages 256–274. Springer, 2024. 3

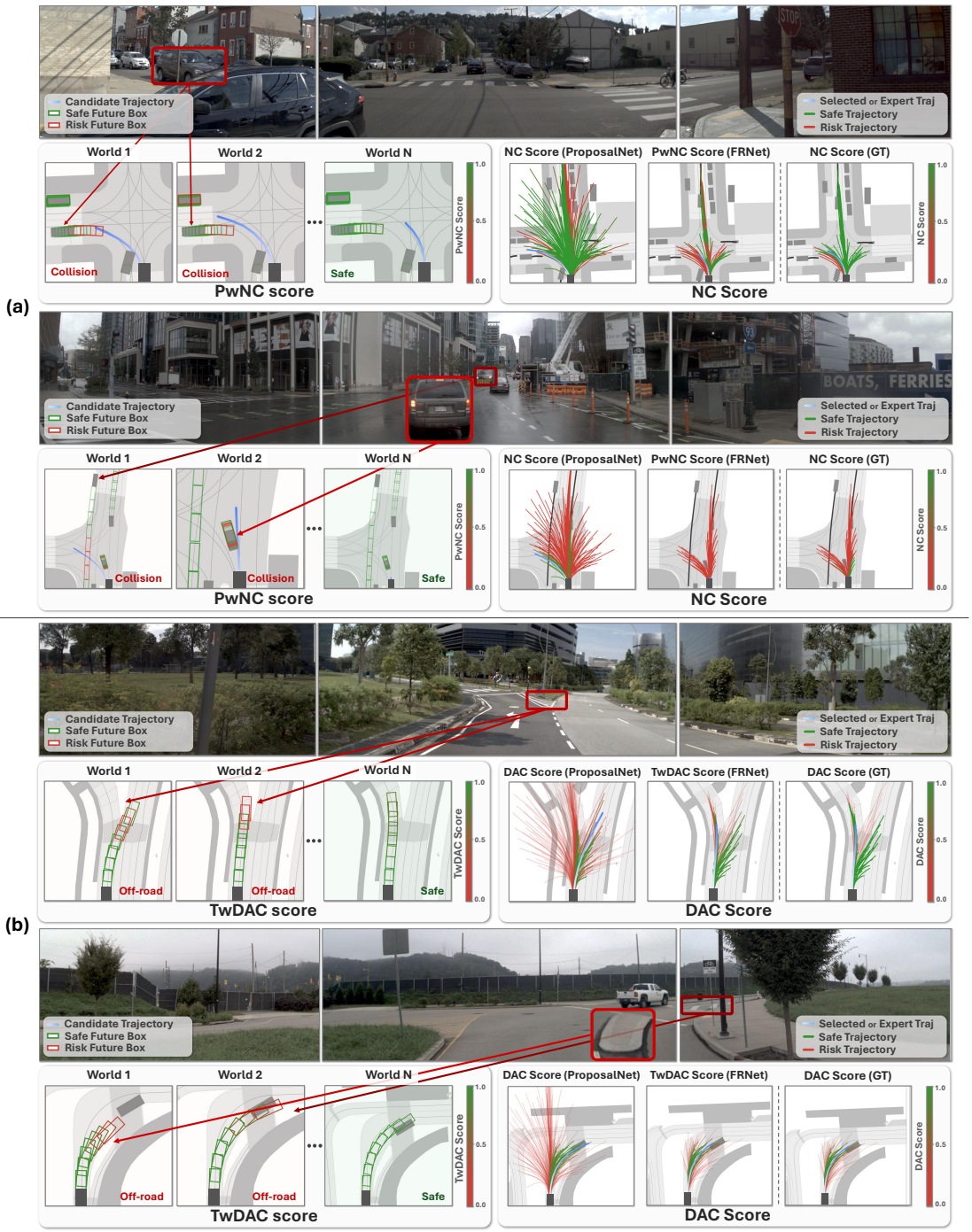

Figure 1. **Visualization of reasoning process.** The figure compares two forms of fine-grained safety reasoning, PwNC in (a) and TwDAC in (b). The bottom-left panel visualizes predicted fine-grained safety scores across Sparse Worlds using red–green shading, with (a) visualizing PwNC scores for the future boxes of surrounding agents and (b) visualizing TwDAC scores for the future ego boxes. The bottom-right panel presents the corresponding trajectory-level scores, showing NC scores for (a) and DAC scores for (b), each generated by ProposalNet, FRNet, and the ground truth.

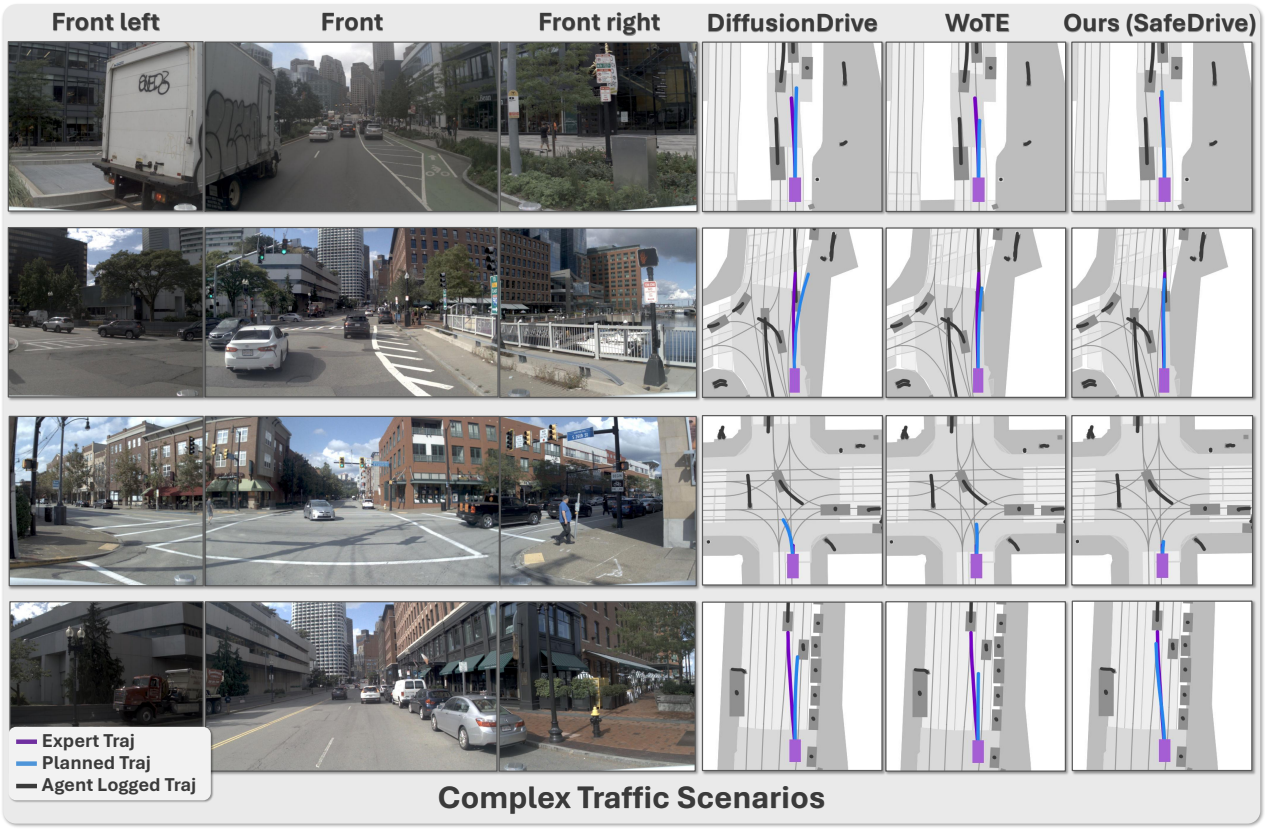

Figure 2. **Additional Comparison with SOTA methods.** Qualitative comparisons of DiffusionDrive, WoTE, and SafeDrive in complex traffic scenarios and turning scenarios.

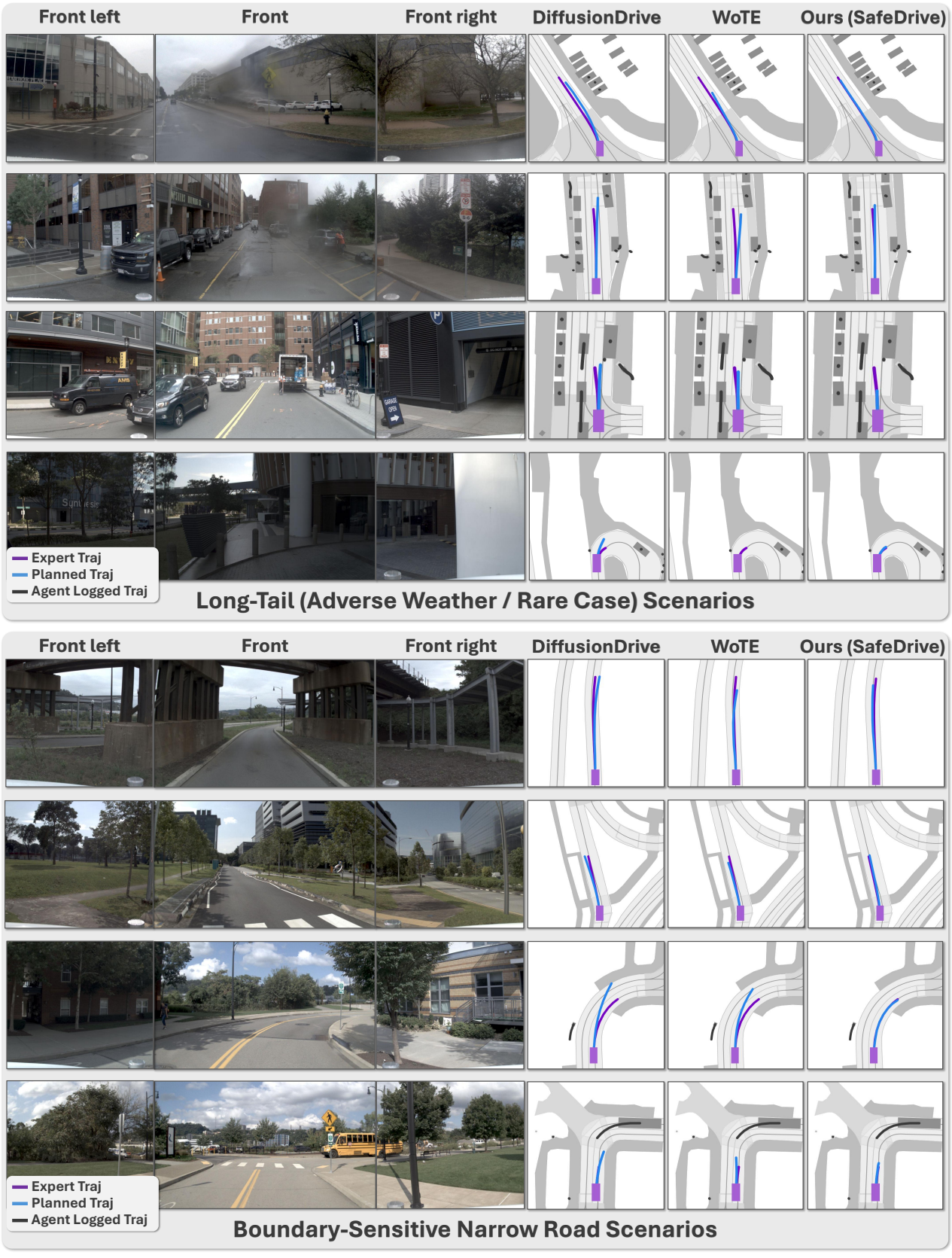

Figure 3. **Additional Comparison with SOTA methods.** Qualitative comparisons of DiffusionDrive, WoTE, and SafeDrive in long-tail scenarios and boundary-sensitive narrow road scenarios.

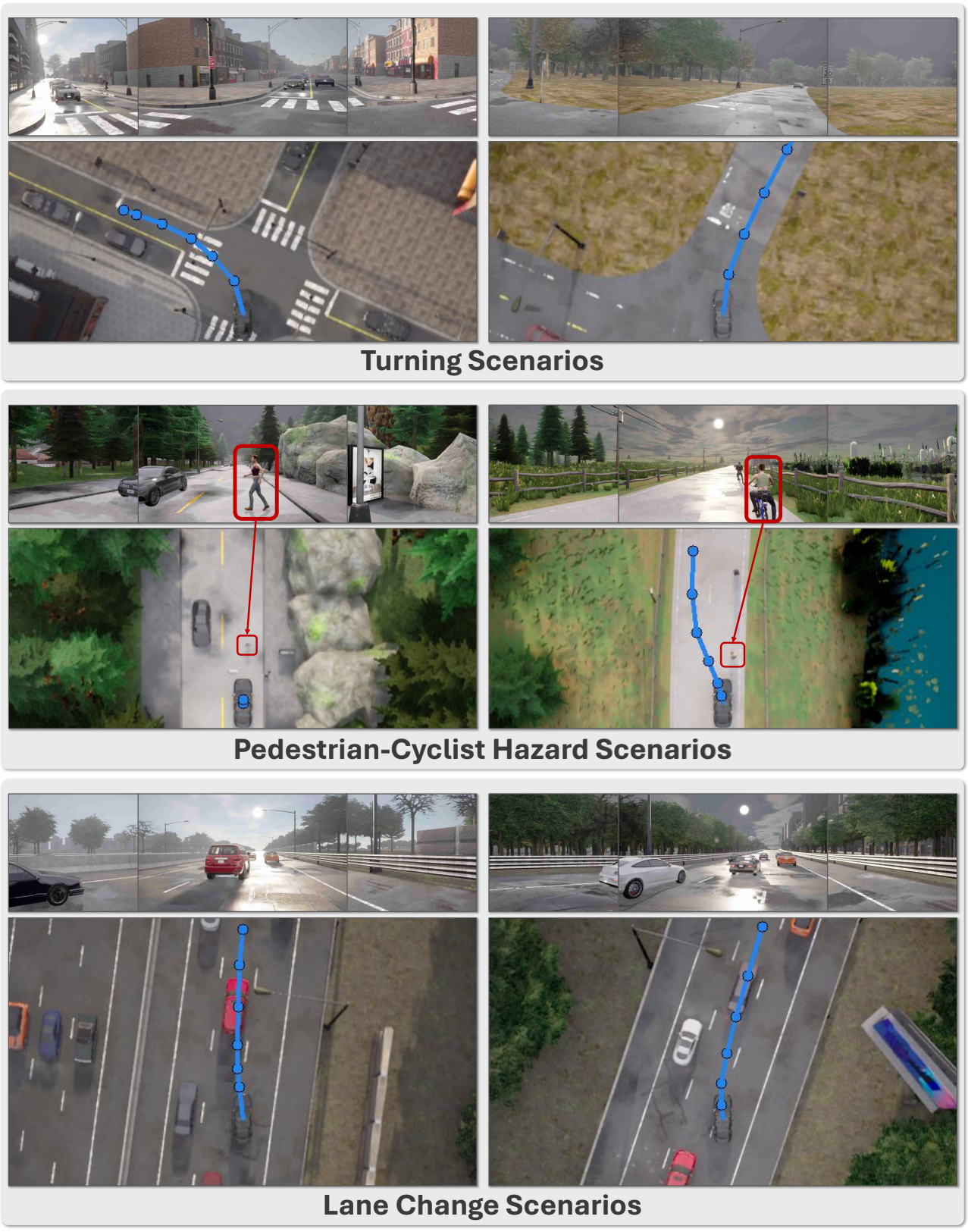

Figure 4. **Visualization of the Bench2Drive benchmark.** The predicted trajectories are visualized as blue curves in each scene. SafeDrive demonstrates robust performance across diverse scenarios in closed-loop environments, consistently generating stable and safe motions.

Figure 5. **Failure Case: Inability to Anticipate Pedestrian Exit from an Opening Door.** At time t, SafeDrive generates a trajectory that continues in its lane without colliding with the stopped vehicle, but the predicted ego path eventually collides with a pedestrian who steps out of the vehicle as the door is opening. This occurs because SafeDrive fails to incorporate the contextual cue that a slightly open door may indicate an imminent pedestrian exit. This case highlights the need for a more comprehensive understanding of contextual cues in the scene, beyond simply recognizing the current and past positions of vehicles and pedestrians.