Pseudo-Expert Regularized Offline RL for End-to-End Autonomous Driving in Photorealistic Closed-Loop Environments

Supplementary Material

A. Detailed Experimental Results for Safety-Efficiency Trade-off

In this section, we provide the full, detailed results complementing the behavior policy analysis in Section 5.7 of the main paper.

Table 5 lists the complete set of behavior policy compositions evaluated, along with their raw performance metrics on both the General Driving and Safety-Critical benchmarks. The "Mixing Ratio" column specifies the data proportions; for example, model M8 was trained on a dataset with a 5:6:1 ratio of VAD ($\sigma = 0.2$), VAD ($\sigma = 0.4$), and Random policy data, respectively.

Additional visualizations are provided to further illustrate the discussed trade-offs. Figure 5 (which is identical to Figure 3b in the main text) and Figure 6 are scatter plots that visualize the relationship between driving efficiency and safety. In these plots, data points are color-coded into four categories based on the training data composition: (1) **VAD only**: Policies using VAD with varying noise levels; (2) **Including Random**: Policies trained on VAD data mixed with random policy data; (3) **Including VADv2**: Policies trained on VAD data mixed with VADv2 data; and (4) **IL baseline**.

As discussed in Sec. 5.7, we observe a clear trend: mixing in VADv2 data or increasing data randomness (higher noise or more random policy data) generally pushes the resulting policy toward the "efficiency-focused" region of the graph (bottom-right).

As described in Sec. 5.2, SRC and JSR are defined as the product of the two axes of Figure 5 and 6, respectively. Although the two plots show a similar tendency, they highlight different aspects of performance. Because both axes of Figure 6 are based on non-collision rates, it particularly emphasizes policies that prioritize safety, even at the expense of efficiency. This structure explains why the JSR metric (the product of these two axes) can reward an overly conservative policy. A clear example is the VADv2 (w/ std. BC) model from Table 1, which achieves a high JSR (25.2%) precisely because it is extremely safety-focused. In contrast, its modest SRC (13.1%) score reflects this trade-off, as the SRC metric is designed to measure both efficiency and safety, penalizing the policy for its low route completion.

Finally, Figure 7 presents bar charts ranking all evaluated models by their unified SRC and JSR scores, offering a clear comparison of their balanced performance. These charts illustrate that the VAD only group tends to achieve higher performance than the other groups, and that nearly all offline RL models outperform the IL baseline on these metrics.

B. Additional Qualitative Results

To further illustrate the behavioral differences between the learned policies, Figure 8 provides additional qualitative examples from the safety-critical NeuroNCAP benchmark. We detail the behavior of the IL baseline, our offline RL model (VADv2*, trained on VAD($\sigma = 0.2$) + VAD($\sigma = 0.4$)), and the RL ablation (VADv2†, trained on VAD($\sigma = 0.2$) + Random). As discussed in Sec. 5.7, VADv2* is a safety-focused policy, while VADv2† is an efficiency-focused policy. The scenarios below highlight these distinct driving 'personalities'.

(a) Adversarial vehicle from left: The IL baseline and VADv2† fail to react to the approaching vehicle and collide. In contrast, VADv2* successfully identifies the hazard and avoids the collision by stopping short of the adversarial vehicle's trajectory.

(b) Stationary vehicle (center): The IL baseline proceeds straight and collides with the obstacle. VADv2† successfully avoids a collision but does so with a high-risk, sharp swerve into the oncoming lane. VADv2* performs a safer, more controlled maneuver by slowing and navigating around the vehicle.

(c) Stationary vehicle (center): In this scenario, all three models successfully avoid a collision. The IL baseline and VADv2* both slow down appropriately, while VADv2† again opts for a more aggressive swerving maneuver.

(d) Bus blocking lane: A large bus completely blocks the road, requiring a full stop. The IL baseline and VADv2† both attempt an evasive maneuver but fail to stop, resulting in a collision. VADv2* correctly identifies the situation and comes to a safe stop before the obstacle.

(e, f) Frontal head-on collision: These are among the most challenging scenarios, with the lowest success rates in the benchmark. Here, we show results for the VADv2‡ model (M14, trained on a 6-policy mix) instead of VADv2*. While all three models (IL, VADv2†, VADv2‡) ultimately fail to avoid a collision, VADv2‡ demonstrates a superior learned response by performing a minimal safety reaction and stopping just before impact.

Table 5. Dataset compositions for behavior policy analysis (Sec. 5.7). Reward parameters were fixed at $w_{\text{imitation}} = 0.1$ and $C_{\text{event}} = -10$, and $\alpha = 0.2$.

ID	Behavior Policy	Mixing Ratio	General Driving				Safety-Critical		
			CR ↓	RC ↑	Long. Jerk (m/s^3) ↓	Lat. Jerk (m/s^3) ↓	CR ↓	SRC ↑	JSR ↑
M2	VAD ($\sigma = 0.2$) + VAD ($\sigma = 0.4$)	1:1	0.511	0.465	0.636	0.158	0.281	0.335	0.352
M18	VAD ($\sigma = 0.1$) + VAD ($\sigma = 0.2$)	1:1	0.551	0.423	0.691	0.288	0.289	0.301	0.319
M14	VAD ($\sigma = 0.1$) + VAD ($\sigma = 0.2$) + VAD ($\sigma = 0.4$) + VADv2 ($\sigma = 0.1$) + VADv2 ($\sigma = 0.2$) + VADv2 ($\sigma = 0.4$)	1:1:1:1:1:1	0.38	0.681	0.747	0.291	0.586	0.282	0.257
M13	VAD ($\sigma = 0.1$) + VAD ($\sigma = 0.2$) + VAD ($\sigma = 0.4$) + VADv2 ($\sigma = 0.1$) + VADv2 ($\sigma = 0.2$) + Random	1:1:1:1:1:1	0.365	0.663	0.695	0.359	0.593	0.27	0.258
M22	VAD ($\sigma = 0.1$)	–	0.596	0.367	0.789	0.357	0.306	0.255	0.281
M7	VAD ($\sigma = 0.2$) + VAD ($\sigma = 0.4$) + Random	5:5:2	0.423	0.606	0.457	0.248	0.588	0.25	0.238
M12	VAD ($\sigma = 0.1$) + VAD ($\sigma = 0.2$) + VAD ($\sigma = 0.4$) + VADv2 ($\sigma = 0.2$)	1:1:1:1	0.438	0.689	0.58	0.312	0.658	0.236	0.192
M16	VAD ($\sigma = 0.2$) + VAD ($\sigma = 0.4$) + VADv2 ($\sigma = 0.1$) + VADv2 ($\sigma = 0.2$)	1:1:1:1	0.482	0.624	0.454	0.257	0.625	0.234	0.194
M11	VAD ($\sigma = 0.1$) + VAD ($\sigma = 0.2$) + VAD ($\sigma = 0.4$)	1:1:1	0.547	0.47	0.55	0.217	0.512	0.229	0.221
M4	VAD ($\sigma = 0.2$)	–	0.584	0.407	0.619	0.356	0.45	0.224	0.229
M15	VAD ($\sigma = 0.1$) + VAD ($\sigma = 0.2$) + VAD ($\sigma = 0.4$) + VADv2 ($\sigma = 0.1$)	1:1:1:1	0.372	0.717	0.472	0.298	0.691	0.222	0.194
M5	VAD ($\sigma = 0.4$)	–	0.387	0.709	0.482	0.363	0.694	0.217	0.188
M17	VAD ($\sigma = 0.2$) + VAD ($\sigma = 0.4$) + VADv2 ($\sigma = 0.1$)	1:1:1	0.401	0.693	0.437	0.243	0.695	0.211	0.183
M8	VAD ($\sigma = 0.2$) + VAD ($\sigma = 0.4$) + Random	5:6:1	0.54	0.467	0.389	0.2	0.565	0.203	0.2
M1	VAD ($\sigma = 0.2$) + VAD ($\sigma = 0.4$) + Random	1:1:1	0.438	0.631	0.54	0.357	0.695	0.193	0.171
M3	VAD ($\sigma = 0.2$) + Random	1:1	0.358	0.72	0.466	0.23	0.747	0.182	0.163
M21	VAD ($\sigma = 0.1$) + VADv2 ($\sigma = 0.1$)	1:1	0.423	0.699	0.496	0.285	0.741	0.181	0.149
M9	VAD ($\sigma = 0.2$) + Random	5:1	0.474	0.609	0.533	0.322	0.713	0.175	0.151
M6	Random	–	0.453	0.67	0.7	0.73	0.742	0.173	0.141
M19	VAD ($\sigma = 0.1$) + VAD ($\sigma = 0.2$) + VADv2 ($\sigma = 0.1$)	1:1:1	0.423	0.65	0.467	0.203	0.742	0.168	0.149
M20	IL	–	0.73	0.341	0.45	0.276	0.659	0.116	0.092
M10	VADv2 ($\sigma = 0.2$)	–	0.526	0.565	0.448	0.202	0.831	0.095	0.08

C. Training Stability Techniques

Offline RL is prone to instability, particularly early in training when the critic is inaccurate. We use three standard stabilization techniques: (1) ramping up the actor-loss weight, (2) EMA target networks for TD target computation, and (3) a discrete action space.

Actor Loss Weight Scheduling. Applying the full actor loss from the start can destabilize training because the critic is still immature. We therefore use an exponential ramp-up schedule. The actor-loss weight is

$$w_{\text{actor}} = w_{\text{base}} \cdot \min\left(w_{\max}, w_{\max} \cdot w_{\text{init}} \cdot \rho^t\right),$$

where t denotes the training iteration, w_{base} is a global scale, $w_{\max}$ is the cap, w_{init} is the initial value, and $\rho > 1$ is the growth rate. The scaled actor objective is

$$\mathcal{L}'_{\text{actor}} = w_{\text{actor}} \cdot \mathcal{L}_{\text{actor}}.$$

In our experiments, $w_{\text{base}} = 10$, $w_{\max} = 1$, $w_{\text{init}} = 10^{-4}$, and $\rho = 1.0004$.

EMA Target Network. Following standard deep RL practice, we maintain slowly updated target copies of the online networks. Denoting online parameters by θ and target parameters by θ', we update the latter by EMA:

$$\theta'^{(t+1)} = (1-\tau)\theta'^{(t)} + \tau\theta^{(t)},$$

where $\tau \in [0, 1)$ is the EMA coefficient. In our experiments, we use $\tau = 1 \times 10^{-4}$. The TD target in the critic loss is computed from the target actor $\pi_{\theta'}$ and target critic $Q_{\psi'}$:

$$y = r + \gamma(1-d)Q_{\psi'}(s', a'), \quad a' \sim \pi_{\theta'}(\cdot \mid s'),$$

where d is the terminal indicator.

Discrete Action Space. We also adopt a discrete action space. This makes the policy objective a categorical log-likelihood, so terms such as $\log \pi_\theta(a \mid s)$ are computed directly by a softmax classifier. In contrast, continuous-action policies typically require explicit density parameterization and more delicate log-likelihood computation. This simpler objective improves numerical stability in both the actor update and the pseudo-expert BC term.

D. Data Collection Details

The offline dataset is generated by running behavior policies in the NeuroNCAP simulator on scenes from the nuScenes training split. Each scene is approximately 20 seconds long. To increase coverage, we start rollouts at 3-second intervals from the beginning of each scene, yielding approximately six rollouts per scene. We refer to one complete pass over all training scenes as a *sweep*. Each rollout terminates either when the scene ends or when a terminal event occurs (collision, off-road, or off-route). In the latter case, the terminal

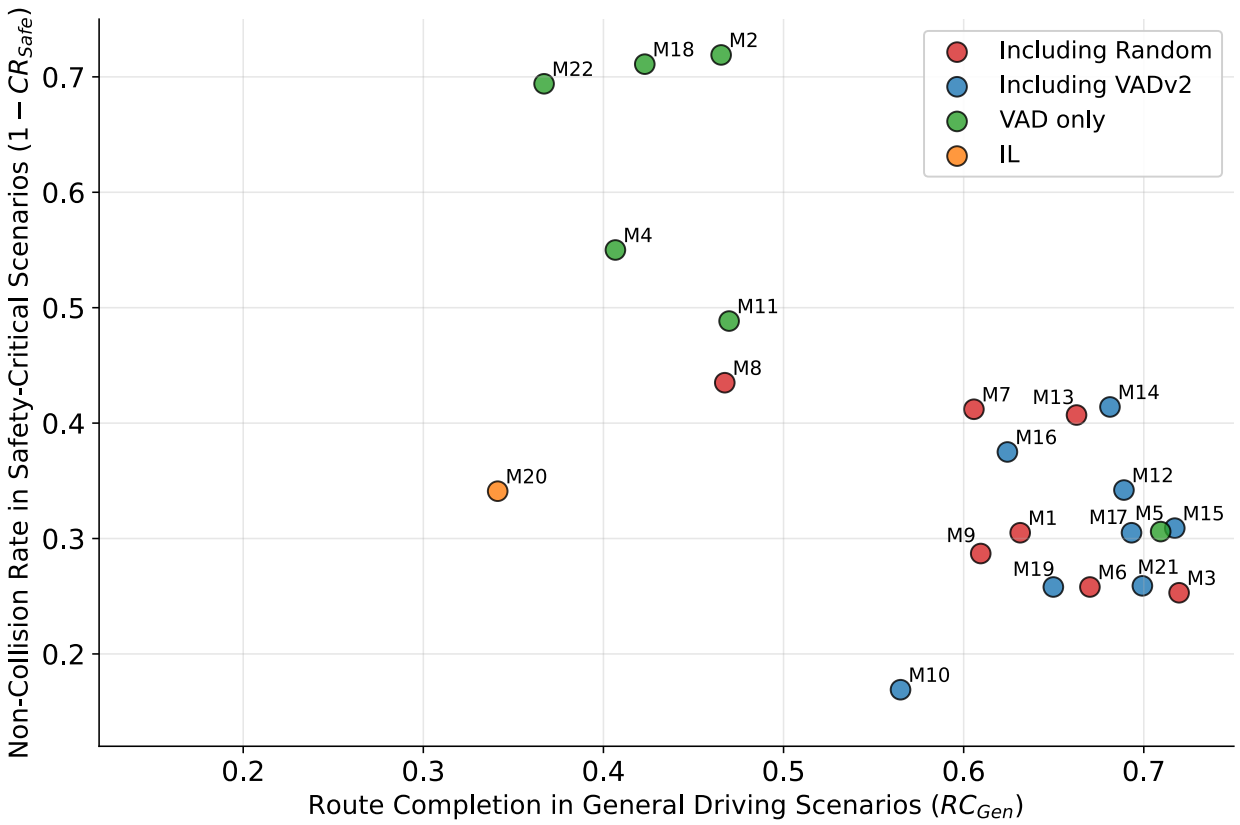


Figure 5. Scatter plot showing the trade-off between route completion in general driving scenarios and non-collision rate in safety-critical scenarios.

flag is set and used in the subsequent TD target computation.

All datasets used in our experiments consist of 12 sweeps. For mixed-policy datasets, the sweeps are allocated in proportion to the specified mixing ratio. For example, M2 in Table 5 uses VAD ($\sigma = 0.2$) + VAD ($\sigma = 0.4$) with a 1:1 ratio, corresponding to 6 sweeps collected with VAD ($\sigma = 0.2$) and 6 sweeps collected with VAD ($\sigma = 0.4$).

The average rollout length varies across behavior policies because the frequency of early termination differs. For example, VAD ($\sigma = 0.2$) yields an average of approximately 11.3 frames per rollout, whereas the Random policy yields approximately 7.6 frames. The shorter average rollout length of the Random policy reflects its higher likelihood of triggering terminal events due to uniformly random action selection.

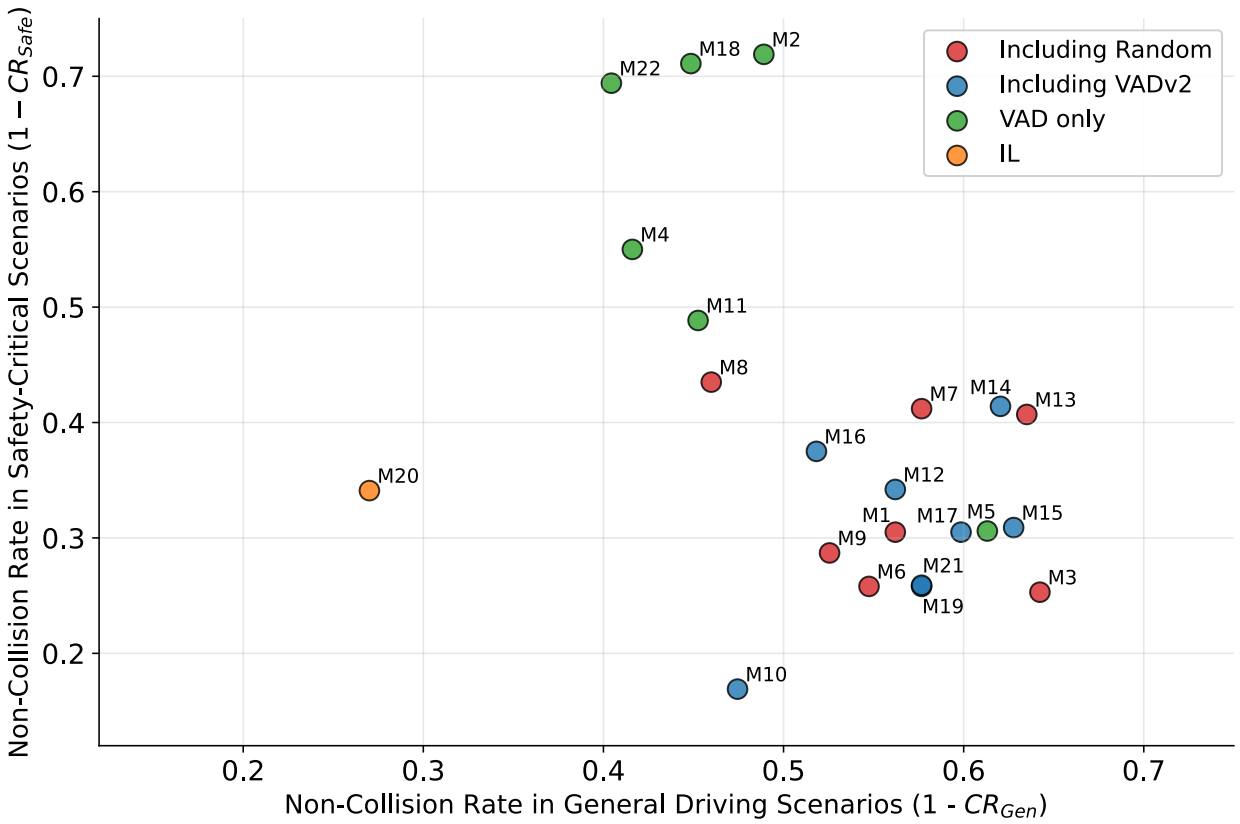
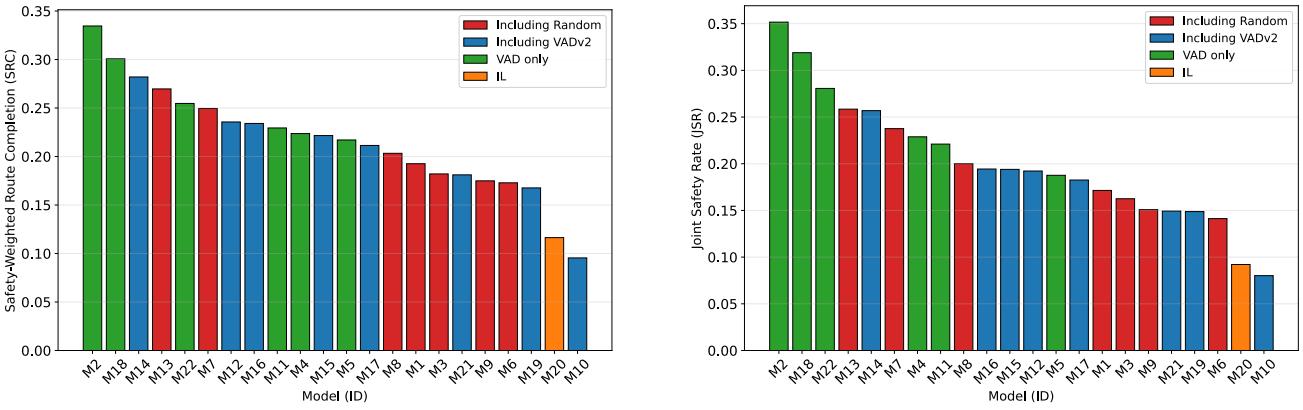


Figure 6. Scatter plot showing the trade-off between non-collision rate in general driving scenarios and non-collision rate in safety-critical scenarios.

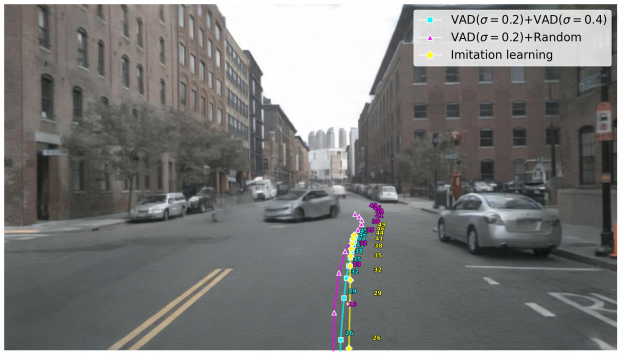


(a) (b)

Figure 7. Comparison of (a) Safety-Weighted Route Completion (SRC) scores and (b) Joint Safety Rate (JSR) across all evaluated models. The models are sorted by each score in descending order. The Model IDs (e.g., M2, M18, M10) correspond to the specific behavior policy compositions detailed in Table 5 in Appendix A.

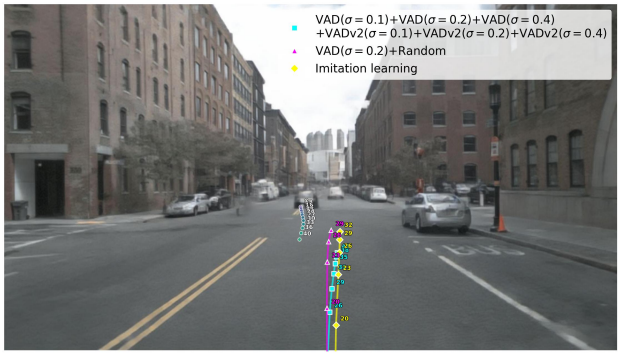
(a)

(b)

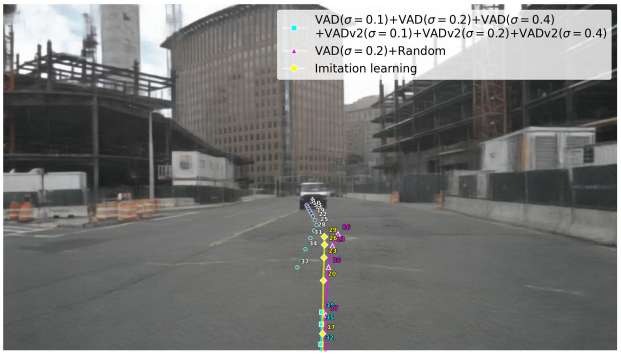


(c)

(d)



(e)



(f)

Figure 8. **Additional qualitative results in safety-critical NeuroNCAP scenarios.** We compare trajectories from the IL baseline, the VADv2† model (trained with Random data), and our VADv2* model. **(a)** An adversarial vehicle approaches from the left. **(b, c, d)** A stationary vehicle or bus obstructs the lane. **(e, f)** An adversarial vehicle approaches head-on. In (e, f), VADv2* is replaced with the VADv2‡ model (M14, a 6-policy mix).