Inverse Reinforcement Learning on GPUDrive

Anonymous Authors

Abstract

We explore the use of inverse reinforcement learning (IRL) to develop robust driving policies in GPUDrive. Using demonstrations from either human experts or PPO-trained agents, we investigate GAIL-style approaches with PPO as inner-loop optimizers, and discriminators trained on egocentric observations or observation-action pairs. Our experiments span 75 worlds with varying numbers of controlled agents and we investigated the difficulty of scaling IRL from single-agent to multi-agent environments. We evaluate policies using task-relevant metrics such as off-road counts, collisions, goal-reaching rates, and hand-crafted episodic returns. These early results raise intriguing questions about reward generalization, scalability, and the design of efficient algorithms for multi-agent autonomy.

Keywords: Inverse Reinforcement Learning, Autonomous Driving, Zero-sum Games, Scaling

1 Extended Abstract

Imitation learning (IL) is the problem of reproducing the behavior of an expert policy $\pi_{\mathbf{E}}$ from demonstrations induced from the policy. The simplest approach is *behavioral cloning* [5], which reduces IL to a supervised learning problem: fitting a policy by maximizing the likelihood of expert actions given expert states. However, because the learner is only trained on expert data, errors compound when the learner visits states outside the expert's distribution, accumulating compounding errors [6]. This issue is unavoidable in general [11].

Inverse reinforcement learning (IRL) works like [4, 13] take a different approach: rather than directly cloning actions, they infer a cost function that rationalizes expert behavior. However, IRL is *ill-posed* without additional structure—many reward functions (e.g., potential-based shapings or near-constant costs) can render the expert optimal. A useful view is the two-player zero-sum game between a policy and a reward model [2, 11]:

$$\min_{r\in R}\max_{\pi\in\Pi}\ \mathbb{E}_{\pi}[r(s,a)] - \mathbb{E}_{\pi_{\mathbf{E}}}[r(s,a)] - \lambda H(\pi). \tag{1}$$

This frames learning as matching trajectory-level statistics while regularizing the policy causal entropy.

Generative adversarial imitation learning (GAIL) [2] is one practical solver for eq. (1). We optimize the reward model (discriminator) by gradient descent on the cross-entropy classification loss distinguishing expert and generated samples, and we optimize the policy through any RL policy updates. Concretely, GAIL seeks a saddle point (π, D) of

$$\mathbb{E}_{\pi}\left[\log D(s,a)\right] + \mathbb{E}_{\pi_E}\left[\log(1 - D(s,a))\right] - \lambda H(\pi), \tag{2}$$

by alternating updates to increase eq. (2) with respect to D and decrease it with respect to π.

GPUDrive [3] is a GPU-accelerated, large-scale, multi-agent driving simulator built upon the Waymo open motion dataset [1]. While much prior IRL work has focused on MuJoCo environments [2, 11], we aim to tackle a more real-world level challenging problem: multi-agent driving in complex, diverse scenes. What makes GPUDrive particularly well-suited for IRL is twofold: *multi-agent interactions*, where small deviations in one agent's policy can cause large shifts in the environment due to other agents, rendering more difficulties in IRL training [12]; and *wide scene diversity*, which can be easily used to test the generalization of learned reward models.

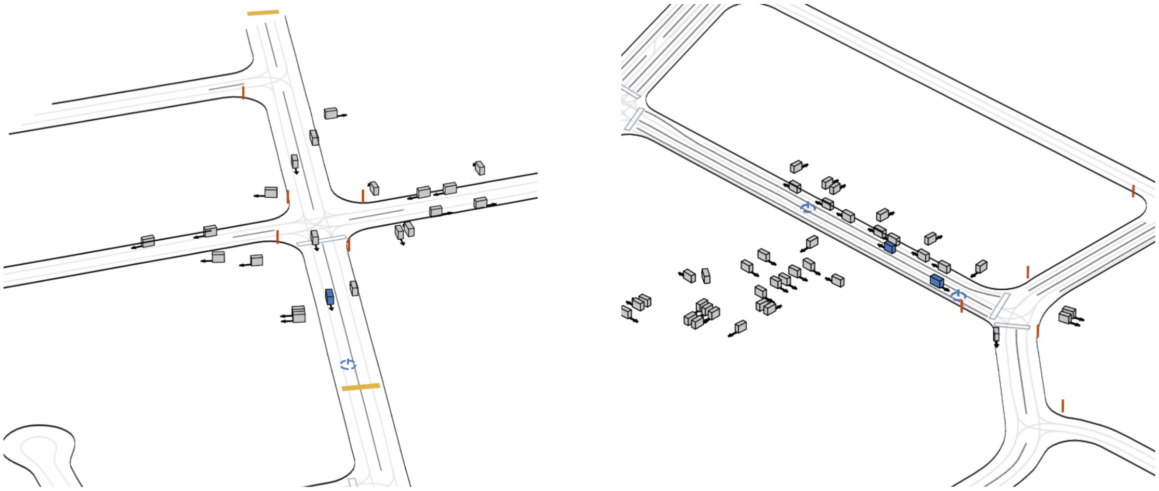

Figure 1: Visualizations of GPUDrive worlds with one controlled agent (left) and three controlled agents (right).

1.1 Experiments

We are in the early stages of implementing IRL algorithms on GPUDrive. Our goal is to learn robust policies from expert demonstrations, using GAIL with PPO as the inner-loop policy optimizer [2, 8]. We conduct experiments across 75 worlds with varying numbers of controlled agents, using discriminators trained on either egocentric observations or state-action pairs. In all our current experiments, we are learning from one human demonstration per agent in each world.

Reward function and evaluation metrics. The discriminator provides a learned reward signal that guides policy training, but we evaluate on metrics on which the policy is not directly trained on. For evaluation, we measure off-road count, collision count, and goal-reaching count. These metrics are not visible to the training algorithm but serve as independent proxies for task performance. Furthermore, we calculate the metric 'episodic reward' with $+1$ for goal-reaching and -1 for crashing. Note that since we terminate the agent after goal-reaching, the cumulative episodic reward is hard capped at 1.

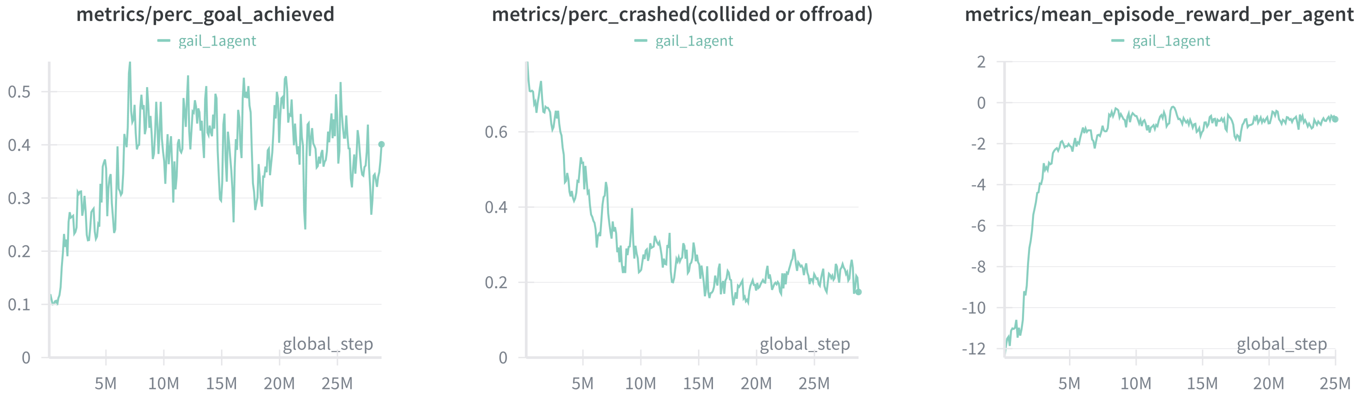

Figure 2: Over training, performance improves across all metrics without access to the hand-crafted reward. Results shown for 1 seed; human expert achieves 98.67% goal and 1.33% collisions (avg. episodic reward 0.97).

Single-agent learning from human demonstrations. In the single-agent setting, we find that vanilla GAIL—without stabilizing tricks or GAN-style regularization—can produce reasonably strong policies with minimal hyperparameter tuning, as shown in fig. 2.

Multi-agent learning from human demonstration Scaling from single-agent to multi-agent settings remains a significant challenge, even when only two agents are controlled within a world. In our experiments, we attempt to control up to 2, 16, and 64 agents, and observe a clear degradation in performance as the number of agents increases. Since one agent's observation inevitably contains other agents in the scene, the reward assignment problem makes it difficult to scale up multi-agent control.

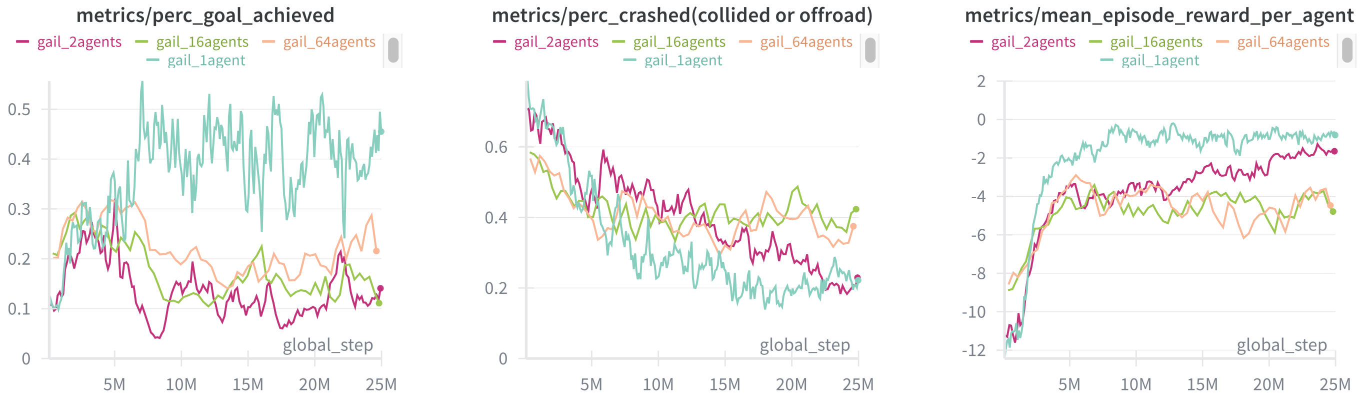

Figure 3: As the maximum number of controlled agents increases, performance degrades across metrics. Results shown for 1 seed.

Behavior Cloning baseline Behavioral cloning is not applicable here because the dataset lacks action labels; we study imitation from observation [9] instead in all the experiments above.

1.2 Future directions and vision

1. Can GAIL learn reward models that support re-training? We plan to test whether the learned reward model can guide a freshly initialized policy to expert-level performance by re-training on the same environment. This directly addresses whether the reward model captures a generalizable signal rather than just overfitting the initial policy trajectory.

2. Can learned reward models generalize to new environments? By training on one set of worlds and evaluating on previously unseen worlds, we can test whether GAIL produces reward models that generalize, as hypothesized for IRL [4, 10].

3. How do inner-loop game dynamics affect the above? Classic IRL algorithms like MaxEnt IRL require full retraining from scratch in the inner loop, naturally supporting policy re-training with the final reward model. GAIL, in contrast, does not possess this property, and prior work [7] has explored resetting the policy to simulate a best-response. More generally, we aim to investigate whether game-theoretic ideas (e.g., optimism, magnetic mirror descent) can enable re-training without the computational expense of full inner-loop RL.

References

[1] Scott Ettinger et al. *Large Scale Interactive Motion Forecasting for Autonomous Driving : The Waymo Open Motion Dataset*. 2021. arXiv: 2104.10133 [cs.CV]. URL: https://arxiv.org/abs/2104.10133.

[2] Jonathan Ho and Stefano Ermon. *Generative Adversarial Imitation Learning*. 2016. arXiv: 1606.03476 [cs.LG]. URL: https://arxiv.org/abs/1606.03476.

[3] Saman Kazemkhani et al. *GPUDrive: Data-driven, multi-agent driving simulation at 1 million FPS*. 2025. arXiv: 2408.01584 [cs.AI]. URL: https://arxiv.org/abs/2408.01584.

[4] Andrew Y Ng, Stuart Russell, et al. "Algorithms for inverse reinforcement learning." In: *Icml*. Vol. 1. 2. 2000, p. 2.

[5] Dean A Pomerleau. "Alvinn: An autonomous land vehicle in a neural network". In: *Advances in neural information processing systems* 1 (1988).

[6] Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. *A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning*. 2011. arXiv: 1011.0686 [cs.LG]. URL: https://arxiv.org/abs/1011.0686.

[7] Silvia Sapora et al. *EvIL: Evolution Strategies for Generalisable Imitation Learning*. 2024. arXiv: 2406.11905 [cs.NE]. URL: https://arxiv.org/abs/2406.11905.

[8] John Schulman et al. "Proximal Policy Optimization Algorithms". In: *CoRR* abs/1707.06347 (2017). arXiv: 1707.06347. URL: http://arxiv.org/abs/1707.06347.

[9] Wen Sun et al. *Provably Efficient Imitation Learning from Observation Alone*. 2019. arXiv: 1905.10948 [cs.LG]. URL: https://arxiv.org/abs/1905.10948.

[10] Gokul Swamy et al. *All Roads Lead to Likelihood: The Value of Reinforcement Learning in Fine-Tuning*. 2025. arXiv: 2503.01067 [cs.LG]. URL: https://arxiv.org/abs/2503.01067.

[11] Gokul Swamy et al. *Of Moments and Matching: A Game-Theoretic Framework for Closing the Imitation Gap*. 2021. arXiv: 2103.03236 [cs.LG]. URL: https://arxiv.org/abs/2103.03236.

[12] Jingwu Tang et al. *Multi-Agent Imitation Learning: Value is Easy, Regret is Hard*. 2024. arXiv: 2406.04219 [cs.LG]. URL: https://arxiv.org/abs/2406.04219.

[13] Brian D Ziebart et al. "Maximum entropy inverse reinforcement learning." In: *Aaai*. Vol. 8. Chicago, IL, USA. 2008, pp. 1433–1438.