# Discriminative Barrier Functions for Safe Adversarial Imitation Learning from Observation

Anubhav Vishwakarma*, Bhaumik Mehta, Caleb Hsu, Byron Boots, Karen Leung, and Tyler Han
University of Washington

***Abstract*—Inverse Reinforcement Learning (IRL) algorithms are powerful tools for learning from and generalizing expert demonstrations, but they often rely on unconstrained exploration, rendering them unsafe for real-world deployment. Meanwhile, Control Barrier Functions (CBFs) can guarantee the safety of control systems, but the analytical design of CBFs can be time-consuming and esoteric. In this work, we address these limitations jointly by constraining reward function candidacy during IRL to the space of CBFs, yielding a formulation that exhibits safe online control with continuous experiential improvement. Crucially, this framework enables the data-driven recovery of barrier functions directly from unlabeled expert observations. We demonstrate that the recovered barrier function is robust to unsafe states entirely absent from the expert data. Furthermore, we benchmark our method against standard IRL baselines in a simulated navigation environment, demonstrating improved safety performance. Finally, we investigate the trade-offs of planning-based versus policy-based IRL methods across both simulation and a real world obstacle avoidance task.**

## I. Introduction

Inverse Reinforcement Learning (IRL) serves as a principled approach for extracting the intrinsic objectives behind expert behavior, enabling the synthesis of robust, generalizable imitation policies [25, 2, 44, 3]. However, most IRL methods rely on action-labeled demonstrations, which are often unavailable or expensive to collect in real-world robotics [37]. In contrast, humans are capable of learning to imitate behaviors directly from visual observation without access to the demonstrator's precise actions, improving through experience; in robot learning, this setting can be referred to as Inverse Reinforcement Learning from Observation (IRLfO), in which a robot recovers a reward function and an imitative policy from state-only expert trajectories [37]. IRLfO methods typically learn online, refining the policy through exploration in the environment [37, 15].

However, exploration poses safety risks that limit the real-world deployment of these algorithms [11]. Humans, by contrast, exhibit the ability to avoid *unsafe states* when learning from expert observations. In cognitive science, the suppression of actions at unsafe states is associated with inhibitory control [26, 39, 20]. Yet existing IRLfO methods focus solely on behavior fidelity and provide no explicit mechanism for avoiding unsafe states during the exploration and learning process [36, 37, 15]. To address this limitation, we study the following problem within the IRLfO setting: *Given expert observational trajectories, how can a robot learn an imitative policy that also performs inhibitory control when encountering unsafe states during learning?*

To instantiate this notion of inhibitory control, we look to control theory, in particular, Control Barrier Functions (CBFs) [7]. CBF theory consists of a barrier function and a dynamics model, and, under certain conditions, it can render a set control-invariant (i.e., there exists a control that ensures the system remains inside the (safe) set indefinitely). Owing to these guarantees, CBFs have been widely employed as explicit safety layers in applications such as adaptive cruise control [6], aerial robotics [16], and legged locomotion [21]. However, manually designing task-specific barrier functions to form CBFs for safety-critical systems remains challenging, analogous to reward engineering in reinforcement learning (RL) [5, 12, 43].

In this work, we investigate the hypothesis that constraining reward function candidacy with the forward invariance property required by CBFs may lead to the recovery of both the barrier function and a policy that learns to inhibit its control near unsafe observations through experience. In particular, we employ a discrete form of CBFs to relax the continuity and differentiability requirements of the barrier function during training. *In this work, we propose DBFs (**Discriminative** Barrier Functions) for IRLfO, which recovers a learned barrier function and an imitative policy jointly from state-only expert observations, enabling safe online imitation learning without action supervision and hand-designed constraints.* We summarize our key contributions as follows:

**Learning Interpretable Barrier Functions:** We demonstrate that restricting the cost function class within the IRLfO framework to Discrete-time CBFs [4] yields highly interpretable and generalizable barrier functions; this constrained objective to synthesize cost function through IRLfO is called DBF. These functions successfully capture expert intent while inherently discriminating against unsafe observations, such as obstacles completely absent from the expert demonstrations.

**Safe Online Learning from Observation:** We show that incorporating DBF regularization into the reward function search significantly reduces collisions in online navigation settings. This mitigates the risks of environment interaction during IRLfO, paving the way toward safe online deployment in the real world.

**Empirical Validation of the DBF Formulation:** We test and compare DBF across both policy-based and planning-based IRLfO frameworks, demonstrating its efficacy in obstacle avoidance tasks through both simulation and real-

*Corresponding Author: av72@uw.edu

world robotic experiments unseen during training time, thereby showcasing the generalization capability of DBF.

## II. Related Work

### A. Learning Control Barrier Functions

CBFs can be challenging to design by hand for general, high-dimensional systems, yet their safety desiderata are intuitive: forward invariance within the safe set and recovery enforcement when violated [7]. These properties make CBFs an attractive structural object for robot learning, where they have already garnered substantial research attention [7, 14]. A common approach to learning CBFs from data exploits discriminative boundaries between safe and unsafe regions [34, 29]. For instance, Srinivasan et al. [34] optimize for a separating hyperplane in a kernel-induced feature space. Along a related line, Yang et al. [41, 42] employ Adversarial Inverse Reinforcement Learning (AIRL) as a pre-training step to obtain a policy, which is subsequently used to recover a CBF. More recently, Nakamura et al. [24] learn a value-function-based CBF from action-labeled demonstrations, similarly requiring a pre-specified safe set and offline supervision [32]. In contrast, a parallel line of work learns barrier-like certificates directly from demonstration data without explicit safe/unsafe labels [8, 29], or imposes conservatism toward out-of-distribution states [18, 23]. However, these methods uniformly require action-labeled demonstrations, operate strictly offline, and rely on a pre-specified unsafe set — limiting their applicability to settings where safety structure must be inferred from unlabeled observations alone. Our work addresses this gap by learning a discrete barrier function directly within the adversarial imitation loop, requiring neither action labels, safety labels, nor offline pretraining.

### B. Safe Reinforcement and Imitation Learning

Towards high-dimensional systems that learn from interaction, safe RL methods aim to adhere to safety-oriented constraints while simultaneously maximizing reward [13]. These approaches naturally depend on a pre-specified constraint set and do not address settings where safety structure must be derived from data. SafeDICE [19] addresses offline safe imitation learning by correcting stationary distributions using labeled non-preferred demonstrations, avoiding constraint-violating behavior without explicit reward learning. However, it operates purely offline, requires action-labeled demonstrations and explicit non-preferred labels, and lacks any mechanism for online adaptation or OOD recovery — assumptions and limitations our framework removes entirely. Closer to our method, CBF-RL [40] incorporates a hand-specified CBF as a constraint term in the policy loss during RL training, internalizing safety without requiring a runtime safety filter at inference. However, the CBF itself must be provided a priori and is not learned from data. While our DBF cost plays an analogous role — discouraging actions that push toward unsafe regions — it is learned entirely from unlabeled observational demonstrations rather than derived from a hand-specified model, and operates within an adversarial imitation loop rather than a pure RL setting.

## III. Discriminative Barrier Functions

This section describes our methodology for learning a barrier function through the AIL framework.

### A. Preliminaries:

Consider a Partially Observable Markov Decision Process (POMDP). In an unknown world state $s_w \in S_w$, the agent makes an observation $o \sim p(o|s_w)$. From a history of observations $\mathbf{o}_{0:t}$, the agent perceives its *state* $s_t \sim p(s|\mathbf{o}_{0:t})$ and performs *action* $a \in \mathcal{A}$. Through a predictive model $f : S \times \mathcal{A} \to S$, the agent self-predicts forward in time for a model-based method. Partial observations of $S$ are desirably used for demonstrations to perform IRL and AIL, as full observation history would quickly become intractable [38]. The expert $\pi_E$ may be acting to optimize some reward based on current and next state $r_w(s, s')$ in the POMDP, which may also potentially encode safety constraints. While in this work we do not assume its availability, known constraints can be employed seamlessly as we discuss in Section III-E.

### B. Inverse Reinforcement Learning from Observation

While there are many IRL formulations from which to begin, Ho and Ermon [17] show how apprenticeship learning with regularization unifies various perspectives on IRL. In short, this objective seeks cost functions $c(\cdot)$ under which the performance of the expert is maximized while other *optimal* policy (i.e. learner, $\pi$) candidates are minimized:

$$\begin{aligned}\text{IRLfO}_\psi(\pi_E) = &\arg\max_{c \in \mathbb{R}^{S\times S}} -\psi(c) \\ &+ \underbrace{\min_{\pi\in\Pi} -\lambda \mathrm{H}(\pi) + \mathbb{E}_\pi[c(s,s')]}_{\text{Entropy-regularized Policy Optimization}} \\ &- \mathbb{E}_{\pi_E}[c(s,s')], \end{aligned} \tag{1}$$

where the observation-only cost $c(s, s')$ [37] is introduced, reflecting the absence of expert action data in our problem setting.

It is through the cost regularizer, $\psi : \mathbb{R}^{S\times S} \to \mathbb{R}$, that Ho and Ermon instantiate other IRL algorithms. Likewise, to introduce our specific form of regularization for our DBF formulation, a brief introduction to CBFs is required.

### C. Discrete-Time Control Barrier Functions

Let safe states be the set $\mathcal{S}$ and unsafe states $\mathcal{U}$. Towards barrier functions, we also introduce the *specification function* $h(\cdot)$ such that, $\mathcal{S} \coloneqq \{s \in \mathcal{S} \mid h(s) \geq 0\}$. Finally, $h$ can be certified as a Discrete-Time CBF if

$$\sup_{a\in\mathcal{A}} h(f(s,a)) - h(s) \geq -\alpha(h(s)) \ . \tag{2}$$

where $\alpha$ is a class $\mathcal{K}$ function. Under this condition, CBFs guarantee that there is an action which maintains forward

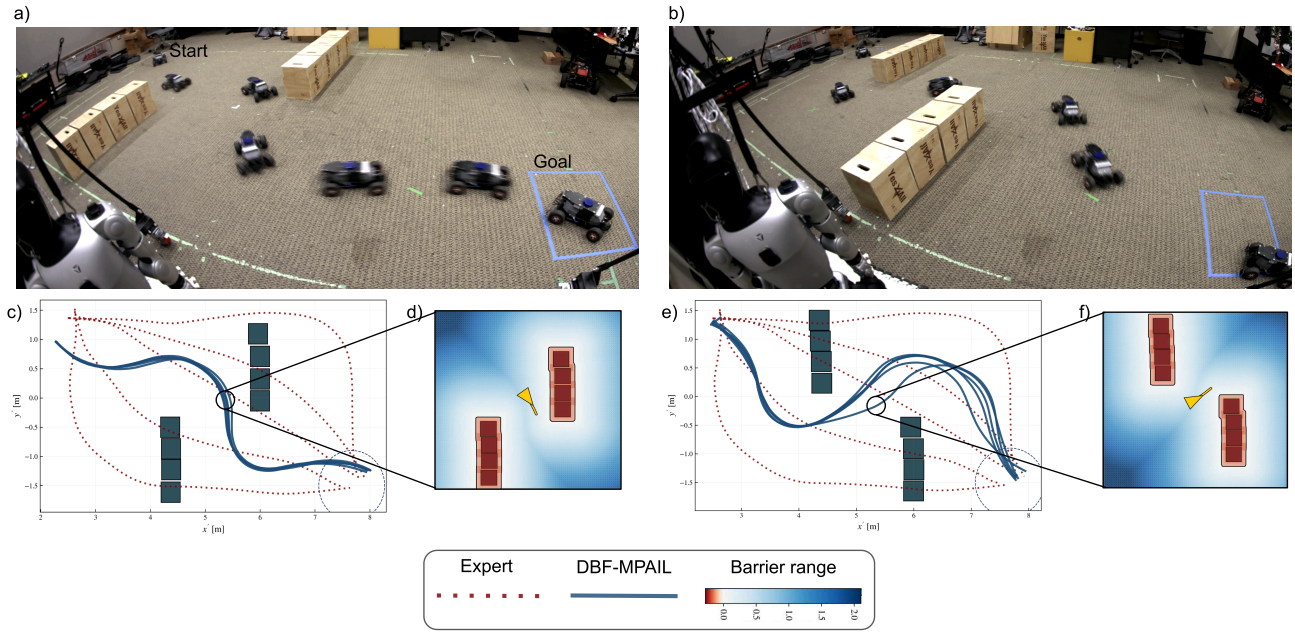



**Fig. 1:** Hardware deployment across two obstacle configurations unseen during training (a,b), demonstrating generalization to unseen unsafe observations. Robots are shown mid-inference navigating toward the goal. (c,e) Closed-loop trajectories for DBF based planning method with expert reference (dotted). (d,f) bird's-eye view map of the robot representing the barrier function range for the map region. See Figure 6 and Table III for performance on additional unseen scenarios.

invariance and stability. For instance: if $h(s_t) < 0$ (i.e. $s_t$ is unsafe), then there is an action for which the agent can increase $h$ for $s_{t+1}$ thereby returning to safety; if $h(s_t) \to 0^+$ (i.e. agent approaches unsafe from safe states), the allowable rate at which the agent may continue approaching $\mathcal{U}$ goes to 0 thereby stably avoiding unsafe states.

### *D. Discriminative Barrier Functions via Cost Regularization*

Now equipped with the IRL optimization objective and CBF safety constraints, we can proceed with the Discriminative Barrier Function (DBF) formulation. First, we observe a simple relationship between AIL and CBFs, which will allow us to view CBFs as a constraint on the AIL objective.

In AIL, an ideal discriminator $D^*$ achieves a boundary which perfectly separates expert states $s \in \mathcal{D}_E$ from learner states $s \in \mathcal{D}_\pi$. Meanwhile for CBFs, the barrier function $h$ can be viewed as separating safe states $s \in \mathcal{S}$ from unsafe states $s \in \mathcal{U}$ as illustrated in Figure 2. These binary classification objectives appear adjacent: *can a barrier function act as a discriminator?* Rather than only learning to discriminate expert and learner data, we wish to also discriminate safe from unsafe data through enforcing control-theoretic structure in the discriminator. Namely, we wish to classify a transition $(s, s')$ based on the condition in Equation (2), through the *dynamic constraint Lagrangian* $q(s, s')$:

$$q_h(s, s') \coloneqq h(s') - h(s) + \alpha(h(s)) \tag{3}$$

which should be more positive for safe transitions that satisfy the constraint and more negative for unsafe transitions that violate the constraint.

Returning our attention to the IRL objective in Equation (1), consider restricting the cost function space to $\mathbb{R}^S$ than in standard AIL dependent on $\mathbb{R}^{S\times S}$.

$$\begin{aligned} \mathcal{C}_{\text{CBF}} \coloneqq \{c \mid\ & c(s, s') = q_h(s, s'),\ h \in \mathbb{R}^S \\ \text{s.t.}\quad & h(s) > 0,\ \forall s \in \mathcal{S} \\ & h(s) < 0,\ \forall s \in \mathcal{U}\}\,. \end{aligned} \tag{4}$$

That is, we wish to search through cost function candidates that are *also Control Barrier Functions* (thereby also imposing control-theoretic structure on the cost function). By combining this restriction on $c$ with the convex regularization from Ho and Ermon [17], as derived in Appendix B, we obtain the *DBF objective*,

$$\begin{aligned} \arg\max_{h\in\mathbb{R}^S} \quad & \underbrace{\mathbb{E}_{\tau_{\text{safe}}}[D_h(s, s')] - \mathbb{E}_{\tau_{\text{unsafe}}}[D_h(s, s')]}_{\mathcal{L}_{\text{Wasserstein-GAN}}} \\ & \underbrace{- \mathbb{E}_{\hat{\mathcal{T}}}\left[\left(\left\|\nabla_{(s,s')} D_h(s, s')\right\|_2 - 1\right)^2\right]}_{\mathcal{L}_{\text{Gradient Penalty}}} \\ \text{s.t.} \quad & \begin{cases} h(s) > 0, & \forall s \in \mathcal{S}_{\text{safe}} \\ h(s) < 0, & \forall s \in \mathcal{S}_{\text{unsafe}} \end{cases} \end{aligned} \tag{5}$$

where $D_h(s, s') = q_h(s, s')$, $\hat{\mathcal{T}} \sim \eta\mathcal{T}_{safe} + (1-\eta)\mathcal{T}_{unsafe}$ and $\sigma(x) \coloneqq (1 + e^{-x})^{-1}$ is the sigmoid function. Note that for the barrier function, to achieve stable training of GAN's we use a Wasserstein GAN formulation, which encourage 1-Lipschitz continuity via a gradient penalty.

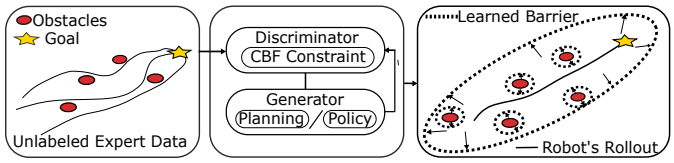



**Fig. 2:** Overview of Learning barrier function from unlabeled expert data. Arrow in learned barrier shows repulsion and expansion of barrier where unsafe and safe respectively.

### *E. Learning DBFs via Adversarial Imitation Learning*

What remains is how this objective should be computed. Since we have not assumed that safety constraints are known, we make the simplifying assumptions that (i) learner data is unsafe, $\mathcal{T}_{\text{unsafe}} \sim \pi$ and (ii) expert data is safe, $\mathcal{T}_{\text{safe}} \sim \pi_E$. Meaning, we can then employ Adversarial Imitation Learning to compute the expectations in Equation (5). Altogether, the AIL objective with DBFs can be summarized:

$$\begin{aligned}\min_{\pi\in\Pi}\max_{h\in\mathbb{R}^S} \quad & \lambda_{\text{wgan}}\Big[\mathbb{E}_{\pi_E}[D_h(s,s')] - \mathbb{E}_{\pi}[D_h(s,s')]\Big] \\ & - \lambda_{\text{GP}}\,\mathbb{E}_{\hat{\pi}}\Big[(\|\nabla_{(s,s')}D_h(s,s')\|_2 - 1)^2\Big] - \mathrm{H}(\pi) \\ & + \lambda_{\text{sign}}\Big[\underbrace{\mathbb{E}_{s\sim\pi_E}\big[(\delta + h(s))_+\big]}_{\textbf{hinge loss}} \\ & \qquad + \underbrace{\mathbb{E}_{s\sim\pi}\big[(\delta - h(s))_+\big]}\Big] \end{aligned} \tag{6}$$

This formulation optimizes the DBF directly from data without requiring explicit, pre-defined safety labels, leveraging the adversarial framework to dynamically separate safe and unsafe states.

In the setting where safety constraints *are known*, it is straightforward to incorporate this additional information, for example, by also populating $\mathcal{T}_{\text{safe}}$ conditionally from the learner or by decomposing $h(s) = \min\{\bar{h}(s), h_\theta(s)\}$ where $\bar{h}$ and $h_\theta$ are prior and learned barrier functions, respectively. In addition, all demonstrations need not be safe as assumed by $\mathcal{T}_{\text{safe}} \sim \pi_E$. It is straightforward to include demonstrations that are unsafe directly in $\mathcal{T}_{\text{unsafe}}$. While we do not explore these extensions here due to their assumption of additional safety knowledge, they remain as promising future work in more applied settings.

As in Robey et al. [29], our intuition behind states being drawn from closed and non-empty safety sets requires notions of $\epsilon$-nets. Put simply, each demonstrated state subsumes an $\epsilon$ ball of states around it, implying that the expert operates with some minimum distance to the unsafe boundary. This is equivalent to constraining the value of the barrier function at safe and unsafe states while limiting the Lipschitz constant of the function, which (along with gradient penalties) is coincidentally common practice for stabilizing AIL algorithms [28].

See appendix A for experimental results.

## IV. Conclusion

In this work, we introduce Discriminative Barrier Functions (DBFs), a framework for incorporating Control Barrier Function (CBF) theory directly into the AIL framework. By constraining the learned discriminator through barrier-function objectives, DBFs enable AIL methods to recover notions of safety and inhibitory control directly from unlabeled demonstrations without requiring explicit unsafe labels or hand-engineered safety costs. Across both policy-based and planning-based AIL algorithms, we demonstrate that DBFs improve the generalization of learned safe sets beyond the expert data distribution without sacrificing imitation performance. DBFs recover coherent safety boundaries, exhibit robustness to previously unseen obstacles and environmental changes, and induce safer exploration during online learning in both simulation and the real world. Overall, our results suggest that incorporating control-theoretic structure into adversarial imitation learning provides a promising direction toward safer and more robust online learning from observation.

## V. Limitations and Future Work

Although DBFs effectively learn to discriminate safe from unsafe states during learning and inference, the safety signal remains confined to the discriminator and does not directly shape the policy's value function. Future work will consider approximate Hamilton-Jacobi reachability methods [9, 22] to modify the generalized advantage estimator, synthesizing a safety-aware value function for improved performance. Additionally, DBFs may become overly conservative when expert data covers only a narrow subset of valid strategies, incorrectly suppressing safe but underrepresented behaviors. Future work may investigate uncertainty-aware DBFs, active data collection, or mechanisms for reasoning about demonstration coverage and ambiguity. The $\mathcal{K}$-class function is currently specified by hand; prior work suggests these functions may themselves be learnable [30], potentially enabling more adaptive barrier representations across tasks with complex or heterogeneous safety constraints.

## Acknowledgments

## References

[1] J. Achiam A. Ray and D. Amodei. Benchmarking safe exploration in deep reinforcement learning. URL https://cdn.openai.com/safexp-short.pdf.

[2] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 1, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015430. URL https://doi.org/10.1145/1015330.1015430.

[3] Pieter Abbeel, Adam Coates, and Andrew Ng. Autonomous helicopter aerobatics through apprenticeship learning. *I. J. Robotic Res.*, 29:1608–1639, 11 2010. doi: 10.1177/0278364910371999.

[4] Ayush Agrawal and Koushil Sreenath. Discrete control barrier functions for safety-critical control of discrete systems with application to bipedal robot navigation. In *Robotics: Science and Systems*, volume 13, pages 1–10. Cambridge, MA, USA, 2017.

[5] Prithvi Akella, Apurva Badithela, Richard M. Murray, and Aaron D. Ames. Lipschitz continuity of signal temporal logic robustness measures: Synthesizing control barrier functions from one expert demonstration, 2023. URL https://arxiv.org/abs/2304.03849.

[6] Aaron D. Ames, Jessy W. Grizzle, and Paulo Tabuada. Control barrier function based quadratic programs with application to adaptive cruise control. In *53rd IEEE Conference on Decision and Control*, pages 6271–6278, 2014. doi: 10.1109/CDC.2014.7040372.

[7] Aaron D. Ames, Samuel Coogan, Magnus Egerstedt, Gennaro Notomista, Koushil Sreenath, and Paulo Tabuada. Control Barrier Functions: Theory and Applications. In *2019 18th European Control Conference (ECC)*, pages 3420–3431, June 2019. doi: 10.23919/ECC.2019.8796030. URL https://ieeexplore.ieee.org/abstract/document/8796030.

[8] Rowan McAllister Koushil Sreenath Adrien Gaidon Fernando Castañeda, Haruki Nishimura. In-distribution barrier functions: Self-supervised policy filters that avoid out-of-distribution states, January 2023. URL https://arxiv.org/abs/2301.12012. arXiv:2301.12012 [cs.RO].

[9] Jaime F. Fisac, Neil F. Lugovoy, Vicenç Rubies-Royo, Shromona Ghosh, and Claire J. Tomlin. Bridging hamilton-jacobi safety analysis and reinforcement learning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8550–8556, 2019. doi: 10.1109/ICRA.2019.8794107.

[10] Justin Fu, Katie Luo, and Sergey Levine. Learning Robust Rewards with Adverserial Inverse Reinforcement Learning. February 2018. URL https://openreview.net/forum?id=rkHywl-A-.

[11] Javier GarcÃa and Fernando FernÃ¡ndez. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015. URL http://jmlr.org/papers/v16/garcia15a.html.

[12] Kunal Garg, James Usevitch, Joseph Breeden, Mitchell Black, Devansh Agrawal, Hardik Parwana, and Dimitra Panagou. Advances in the theory of control barrier functions: Addressing practical challenges in safe control synthesis for autonomous and robotic systems, 2023. URL https://arxiv.org/abs/2312.16719.

[13] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A Review of Safe Reinforcement Learning: Methods, Theories, and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):11216–11235, December 2024. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3457538. URL https://ieeexplore.ieee.org/document/10675394/.

[14] Maeva Guerrier, Hassan Fouad, and Giovanni Beltrame. Learning Control Barrier Functions and their application in Reinforcement Learning: A Survey, April 2024. URL http://arxiv.org/abs/2404.16879. arXiv:2404.16879 [cs].

[15] Tyler Han, Yanda Bao, Bhaumik Mehta, Gabriel Guo, Anubhav Vishwakarma, Emily Kang, Sanghun Jung, Rosario Scalise, Jason Zhou, Bryan Xu, and Byron Boots. Model Predictive Adversarial Imitation Learning for Planning from Observation, July 2025. URL http://arxiv.org/abs/2507.21533. arXiv:2507.21533 [cs].

[16] Marvin Harms, Mihir Kulkarni, Nikhil Khedekar, Martin Jacquet, and Kostas Alexis. Neural control barrier functions for safe navigation, 2024. URL https://arxiv.org/abs/2407.19907.

[17] Jonathan Ho and Stefano Ermon. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/cc7e2b878868cbae992d1fb743995d8f-Abstract.html.

[18] Hussein Sibai Ihab Tabbara. Learning neural control barrier functions from offline data with conservatism, September 2025. URL https://arxiv.org/abs/2505.00908v1. arXiv:2505.00908 [cs.LG].

[19] Youngsoo Jang, Geon-Hyeong Kim, Jongmin Lee, Sungryull Sohn, Byoungjip Kim, Honglak Lee, and Moontae Lee. Safedice: Offline safe imitation learning with non-preferred demonstrations. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 74921–74951. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed2fb79f2664c3d9ba878be7e575b2af-Paper-Conference.pdf.

[20] Armin Lederer, Erfaun Noorani, John S. Baras, and Sandra Hirche. Risk-sensitive inhibitory control for safe reinforcement learning. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 1040–1045, 2023. doi: 10.1109/CDC49753.2023.10383524.

[21] Qiayuan Liao, Zhongyu Li, Akshay Thirugnanam, Jun Zeng, and Koushil Sreenath. Walking in narrow spaces: Safety-critical locomotion control for quadrupedal robots with duality-based optimization, 2023. URL https://arxiv.org/abs/2212.14199.

[22] I.M. Mitchell, A.M. Bayen, and C.J. Tomlin. A time-dependent hamilton-jacobi formulation of reachable sets for continuous dynamic games. *IEEE Transactions on Automatic Control*, 50(7):947–957, 2005. doi: 10.1109/TAC.2005.851439.

[23] Aditya Singh Shishir Kolathaya Ravi Prakash Mumuksh Tayal, Manan Tayal. V-ocbf: Learning safety filters from offline data via value-guided offline control barrier functions, April 2026. URL https://arxiv.org/abs/2512.10822. arXiv:2512.10822 [cs.AI].

[24] Kensuke Nakamura, Arun L. Bishop, Steven Man, Aaron M. Johnson, Zachary Manchester, and Andrea

Bajcsy. How to train your latent control barrier function: Smooth safety filtering under hard-to-model constraints. In *Learning for Dynamics and Control*, 2026.

[25] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

[26] Joel T. Nigg. On Inhibition/Disinhibition in Developmental Psychopathology: Views from Cognitive and Personality Psychology and a Working Inhibition Taxonomy. *Psychological Bulletin*, 126(2):220–246, 2000.

[27] NVIDIA, :, Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, Lukasz Wawrzyniak, Milad Rakhsha, Alain Denzler, Eric Heiden, Ales Borovicka, Ossama Ahmed, Iretiayo Akinola, Abrar Anwar, Mark T. Carlson, Ji Yuan Feng, Animesh Garg, Renato Gasoto, Lionel Gulich, Yijie Guo, M. Gussert, Alex Hansen, Mihir Kulkarni, Chenran Li, Wei Liu, Viktor Makoviychuk, Grzegorz Malczyk, Hammad Mazhar, Masoud Moghani, Adithyavairavan Murali, Michael Noseworthy, Alexander Poddubny, Nathan Ratliff, Welf Rehberg, Clemens Schwarke, Ritvik Singh, James Latham Smith, Bingjie Tang, Ruchik Thaker, Matthew Trepte, Karl Van Wyk, Fangzhou Yu, Alex Millane, Vikram Ramasamy, Remo Steiner, Sangeeta Subramanian, Clemens Volk, CY Chen, Neel Jawale, Ashwin Varghese Kuruttukulam, Michael A. Lin, Ajay Mandlekar, Karsten Patzwaldt, John Welsh, Huihua Zhao, Fatima Anes, Jean-Francois Lafleche, Nicolas Moënne-Loccoz, Soowan Park, Rob Stepinski, Dirk Van Gelder, Chris Amevor, Jan Carius, Jumyung Chang, Anka He Chen, Pablo de Heras Ciechomski, Gilles Daviet, Mohammad Mohajerani, Julia von Muralt, Viktor Reutskyy, Michael Sauter, Simon Schirm, Eric L. Shi, Pierre Terdiman, Kenny Vilella, Tobias Widmer, Gordon Yeoman, Tiffany Chen, Sergey Grizan, Cathy Li, Lotus Li, Connor Smith, Rafael Wiltz, Kostas Alexis, Yan Chang, David Chu, Linxi "Jim" Fan, Farbod Farshidian, Ankur Handa, Spencer Huang, Marco Hutter, Yashraj Narang, Soha Pouya, Shiwei Sheng, Yuke Zhu, Miles Macklin, Adam Moravanszky, Philipp Reist, Yunrong Guo, David Hoeller, and Gavriel State. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning, 2025. URL https://arxiv.org/abs/2511.04831.

[28] Manu Orsini, Anton Raichuk, Leonard Hussenot, Damien Vincent, Robert Dadashi, Sertan Girgin, Matthieu Geist, Olivier Bachem, Olivier Pietquin, and Marcin Andrychowicz. What matters for adversarial imitation learning? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14656–14668. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/7b647a7d88f4d6319bf0d600d168dbeb-Paper.pdf.

[29] Alexander Robey, Haimin Hu, Lars Lindemann, Hanwen Zhang, Dimos V. Dimarogonas, Stephen Tu, and Nikolai Matni. Learning Control Barrier Functions from Expert Demonstrations. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3717–3724, December 2020. doi: 10.1109/CDC42340.2020.9303785. URL https://ieeexplore.ieee.org/document/9303785/. ISSN: 2576-2370.

[30] Vivek Sharma, Negar Mehr, and Naira Hovakimyan. Learning differentiable and safe multi-robot control for generalization to novel environments using control barrier functions, 2024.

[31] Oswin So, Zachary Serlin, Makai Mann, Jake Gonzales, Kwesi Rutledge, Nicholas Roy, and Chuchu Fan. How to train your neural control barrier function: Learning safety filters for complex input-constrained systems, 2023. URL https://arxiv.org/abs/2310.15478.

[32] Oswin So, Zachary Serlin, Makai Mann, Jake Gonzales, Kwesi Rutledge, Nicholas Roy, and Chuchu Fan. How to train your neural control barrier function: Learning safety filters for complex input-constrained systems. In *2024 IEEE International Conference on Robotics and Automation (ICRA) (Under Review)*, 2024.

[33] Siddhartha S. Srinivasa, Patrick Lancaster, Johan Michalove, Matt Schmittle, Colin Summers, Matthew Rockett, Joshua R. Smith, Sanjiban Chouhury, Christoforos Mavrogiannis, and Fereshteh Sadeghi. MuSHR: A low-cost, open-source robotic racecar for education and research. *CoRR*, abs/1908.08031, 2019.

[34] Mohit Srinivasan, Amogh Dabholkar, Samuel Coogan, and Patricio A. Vela. Synthesis of Control Barrier Functions Using a Supervised Machine Learning Approach. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7139–7145, October 2020. doi: 10.1109/IROS45743.2020.9341190. URL https://ieeexplore.ieee.org/document/9341190/. ISSN: 2153-0866.

[35] Jiankai Sun, Lantao Yu, Pinqian Dong, Bo Lu, and Bolei Zhou. Adversarial Inverse Reinforcement Learning With Self-Attention Dynamics Model. *IEEE Robotics and Automation Letters*, 6(2):1880–1886, April 2021. ISSN 2377-3766. doi: 10.1109/LRA.2021.3061397. URL https://ieeexplore.ieee.org/document/9361118.

[36] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation, 2018. URL https://arxiv.org/abs/1805.01954.

[37] Faraz Torabi, Garrett Warnell, and Peter Stone. Generative Adversarial Imitation from Observation, June 2019. URL http://arxiv.org/abs/1807.06158. arXiv:1807.06158 [cs].

[38] Samuel Triest, Mateo Guaman Castro, Parv Maheshwari, Matthew Sivaprakasam, Wenshan Wang, and Sebastian Scherer. Learning risk-aware costmaps via inverse reinforcement learning for off-road navigation, 2023. URL https://arxiv.org/abs/2302.00134.

[39] Núñez D. E. Gaspar P. A. Brass M. Ulloa, J. L. Imitative inhibitory control is associated with psychotic experiences in a sample from the general population. *Frontiers in psychiatry, 15, 1470030*, 126, 2024.

[40] Lizhi Yang, Blake Werner, Massimiliano de Sa, and Aaron D. Ames. Cbf-rl: Safety filtering reinforcement learning in training with control barrier functions, 2026. URL https://arxiv.org/abs/2510.14959.

[41] Yue Yang, Letian Chen, and Matthew Gombolay. Safe Inverse Reinforcement Learning via Control Barrier Function, March 2023. URL http://arxiv.org/abs/2212.02753. arXiv:2212.02753 [cs].

[42] Yue Yang, Letian Chen, Zulfiqar Zaidi, Sanne van Waveren, Arjun Krishna, and Matthew Gombolay. Enhancing Safety in Learning from Demonstration Algorithms via Control Barrier Function Shielding. In *2024 19th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 820–829, March 2024. URL https://ieeexplore.ieee.org/document/10660943/. ISSN: 2167-2121.

[43] Songyuan Zhang, Oswin So, Kunal Garg, and Chuchu Fan. Gcbf+: A neural graph control barrier function framework for distributed safe multiagent control. *IEEE Transactions on Robotics*, 41:1533–1552, 2025. ISSN 1941-0468. doi: 10.1109/tro.2025.3530348. URL http://dx.doi.org/10.1109/TRO.2025.3530348.

[44] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1433–1438. AAAI Press, 2008. ISBN 9781577353683.

## APPENDIX A
## EXPERIMENTAL SETUP AND ADDITIONAL RESULTS

| Algorithm | Reward | Learner |
|---|---|---|
| GAIL | $\log D$ | PPO (Policy) |
| AIRL | $\sigma^{-1} \circ D$ | PPO (Policy) |
| MPAIL | $\sigma^{-1} \circ D$ | MPPI (Planner) |

**TABLE I:** AIL Algorithm Summaries

We evaluate the DBF objective with different AIL Algorithm 1, including GAIL [17], AIRL [10], and MPAIL [15], summarized in Table I. These algorithms are primarily differentiated by their reward structure and whether they are policy-based or planning-based. GAIL and AIRL are model-free methods, while MPAIL is a model-based AIL method that achieves strong out-of-distribution recovery through approximated policy value-function bootstrapping at the terminal state [15]. In experiments we aim to answer the following questions:

**Q1** Can DBFs learn barrier functions that generalize to novel environments from unlabeled observations unseen during training?

**Q2** How do DBF-augmented AIL algorithms compare against existing AIL baselines in terms of safety?

**Algorithm 1** Discriminative Barrier Function based Adversarial Inverse Reinforcement Learning

Obtain state-only expert observation trajectories $\tau_E$
Initialize policy $\pi_\phi$ and discriminator $D_{h_\theta}$.
Initialize experience replay buffer $\mathcal{B} \leftarrow \{\}$.
**for** step $t$ in $\{1, \ldots, N\}$ **do**
  Collect trajectories $\tau_i = (s_0, \ldots, s_T)$ by executing $\pi_\phi$ and $\tau_i \rightarrow \mathcal{B}$.
  Sample interpolations $\hat{\tau} \sim \eta\tau_E + (1-\eta)\tau_u$; where $\tau_u \sim \mathcal{B}$.
  Update discriminator parameters $\theta$ by minimizing $\mathcal{L}(\theta)$:

$$\begin{aligned}\mathcal{L}(\theta) = \lambda_{\text{wgan}} &\Big[\mathbb{E}_{\tau_u}[D_{h_\theta}(s,s')] - \mathbb{E}_{\tau_E}[D_{h_\theta}(s,s')]\Big] \\ &+ \lambda_{\text{GP}}\mathbb{E}_{\hat{\tau}}\Big[(\|\nabla_{(s,s')} D_{h_\theta}(s,s')\|_2 - 1)^2\Big] \\ &- \mathbb{E}_{s\sim\tau_E}\big[(\delta + h_\theta(s))_+\big] - \mathbb{E}_{s\sim\tau_u}\big[(\delta - h_\theta(s))_+\big]\end{aligned}$$

  $r_\theta(s,s') \leftarrow \text{AIL}(D_{h_\theta}(s,s'))$ (see Table I)
  Update policy parameters $\phi$ using $r_\theta(s,s')$ with any policy optimization method.
**end for**

**Q3** Can DBFs improve safety in real-world and safety-critical environments?

*Note:* In all experiments, the expected value of the discriminator in Equation (6) is calculated off-policy.

### A. Navigation In Simulated Environment:

We first qualitatively analyze the learned boundary around expert trajectory in a simple no-obstacle simulation setting, where unsafe states can be defined as those not present in the expert data. We use MUSHR [33], a 10-DOF ego vehicle, as a testbed in the Isaac Lab [27] simulation environment for navigation to a goal point. Figure 3 compares the barrier function learned by DBF-based methods, $h(s)$, with the single-state (position-only) reward function, $r(s)$, used by baseline AIL methods for learning the discriminative boundary around the expert trajectory. This comparison is necessary because, in standard AIL, the reward function may implicitly encode safety constraints. We retain the circling motion near the goal in the expert trajectory for qualitative comparison, since this behavior is more difficult to imitate from state-only observations, as corroborated by [28]. The zero contour curve serves as the discriminative boundary — states inside are classified as safe ($h(s), r(s) > 0$) and states outside as unsafe ($h(s), r(s) < 0$). AIL algorithms equipped with DBFs generalize substantially better around the expert demonstrations, while also correctly modeling the challenging circling behavior at the goal. In contrast, the baselines collapse and overestimate the discriminative boundary near goal position.

### B. Q1: Barrier Function Robustness to Unsafe States not Present in Expert Data

**Experiment Setup:** For this study, we use a maze environment enclosing the expert trajectories (Figure 7), with hori-

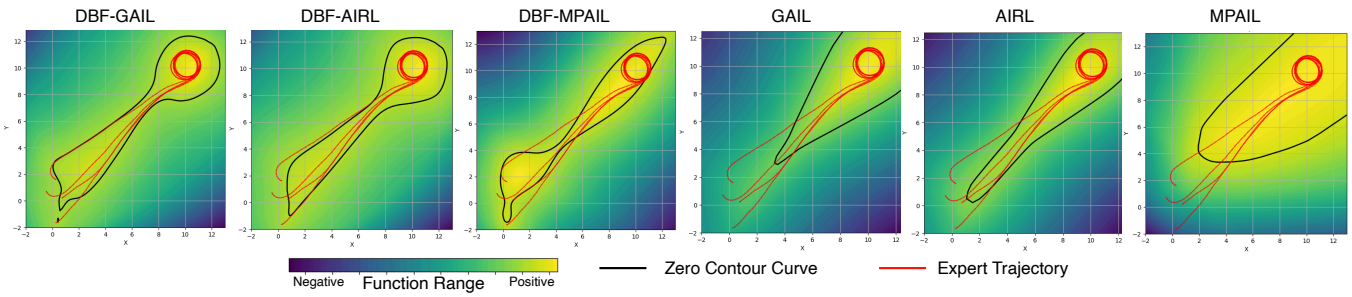



**Fig. 3:** Qualitative comparison of the learned barrier function $h(s)$ for DBF-based methods and the reward function $r(s)$ for non-DBF variants of GAIL, AIRL, and MPAIL. For DBF methods, the learned barrier function $h(s)$ is plotted; for non-DBF baselines, the learned reward function $r(s)$ is plotted. The black curve denotes the zero contour, and the red curve denotes the expert trajectory. Here $s \in \mathbb{R}^2$ represents the agent's position.

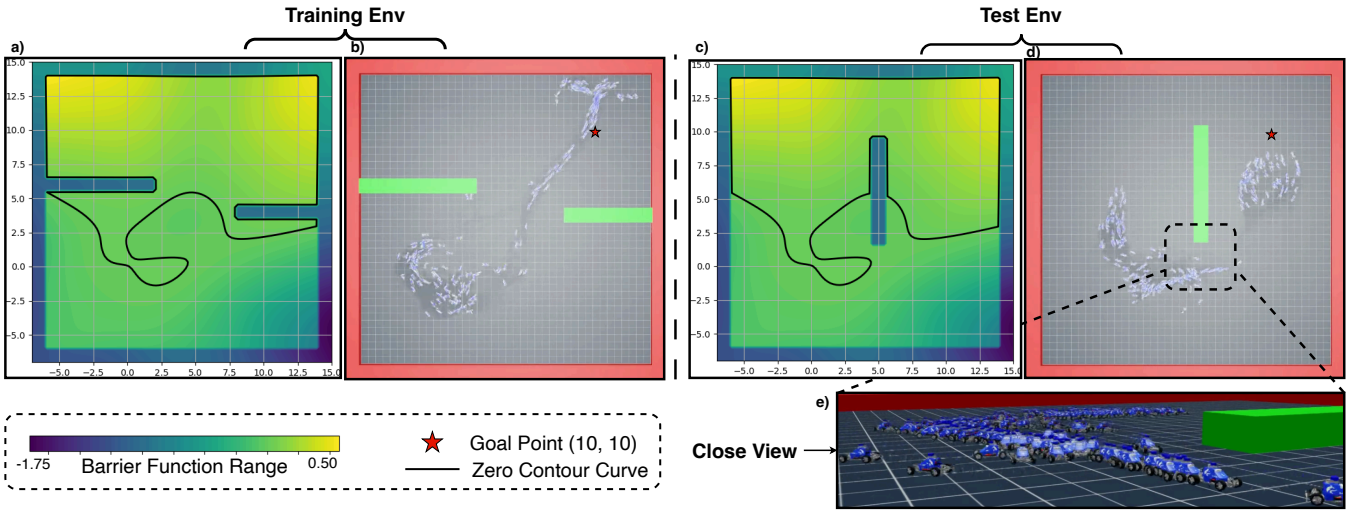



**Fig. 4: Qualitative analysis of barrier function generalization synthesized via DBF-MPAIL.** **(a)** Learned continuous barrier function $h(s)$ from expert state-only observations (Figure 7). **(b)** Top view of the training environment with the marked goal point. **(c)** The same barrier function adapted to a novel test environment. **(d)** Top view of the test environment showing agent trajectories. **(e)** Close-up view of robots inhibiting their control inputs to avoid obstacles in the unseen environment.

zontal wall positions randomized during training (Figure 4b) to promote generalization, the discriminator is evaluated with expectations from an off-policy replay buffer, exposing it to a diverse set of states beyond the current policy's distribution. A vertical wall is placed at the center of the environment during test time to investigate generalization to unseen obstacle configurations (Figure 4d).

**Qualitative Analysis:** We investigate robustness to unsafe states of the DBF-based method unseen in expert data after having established that DBF-based methods learn expressive discriminative boundaries (Figure 3), especially DBF-MPAIL, among having the highest variation in expressivity within the safe set, i.e., zero contour. We next analyze DBF-MPAIL in the obstacle-avoidance task using unlabeled expert observational trajectories to train the method. During training, the expert data is classified as safe; however, when learning to imitate in the simulation environment, the robot may encounter obstacles unseen in the expert data that hinder its navigation to the goal due to walls. As shown in Figure 4 parts a and c, DBF-MPAIL preserves meaningful notions of safety even under these distribution shifts both while training and during evaluation, respectively. Although portions of the expert trajectories intersect obstacle regions, the learned barrier function adapts by rejecting unsafe transitions while maintaining connectivity to the demonstrated safe set. The continuity imposed by the barrier objective (Equation (6)) forces the learned safe set to remain topologically consistent around the obstacle.

**Quantitative comparison:** To quantify above analysis for all the DBF based AIL methods, we evaluate the performance of them in the test environment (Figure 4d) by comparing the total number of collisions out of the total number of robot spawns, as well as the average environmental reward, to quantify the task objective of reaching the goal while avoiding

**TABLE II:** Obstacle navigation evaluation (mean $\pm$ std across $N = 5$ training seeds: 1, 2, 3, 22, and 312).

| Algorithm | Collision (%) $\downarrow$ | Avg. Env Reward $\uparrow$ |
|---|---|---|
| DBF-AIRL | $\mathbf{0.0 \pm 0.0}$ | $\mathbf{0.264 \pm 0.078}$ |
| DBF-GAIL | $\mathbf{0.0 \pm 0.0}$ | $\mathbf{0.300 \pm 0.014}$ |
| DBF-MPAIL | $\mathbf{0.0 \pm 0.0}$ | $\mathbf{0.338 \pm 0.022}$ |
| AIRL | $0.0 \pm 0.0^*$ | $0.246 \pm 0.041$ |
| GAIL | $20.5 \pm 25.4$ | $-0.030 \pm 0.174$ |
| MPAIL | $12.3 \pm 1.8$ | $0.455 \pm 0.029$ |

*AIRL achieves zero collisions due to its reward structure, not explicit safety constraints A-F.

obstacles. The results are shown in Table II.[1] It is evident from the performance table that DBF objective strongly influences the performance of planning-based method, since the reward function (i.e., the CBF constraint) is directly used as a stage cost of MPPI planner, providing strong guidance toward safe actions. In addition, value bootstrapping at terminal states encodes safety context, providing guidance analogous to the role of the learned policy at terminal states leading to control inhibition, as shown in Figure 4e. Moreover, we observe improved performance in policy-based AIL methods: GAIL with the DBF formulation leads to zero collisions than its counterpart, indicating that the learned policy derived from this discriminative reward function does promote inhibitory control. These comparisons demonstrate that the DBF objective learns a generalizable barrier function, as shown qualitatively in Figure 4 and demonstrated performance quantitatively in Table II.

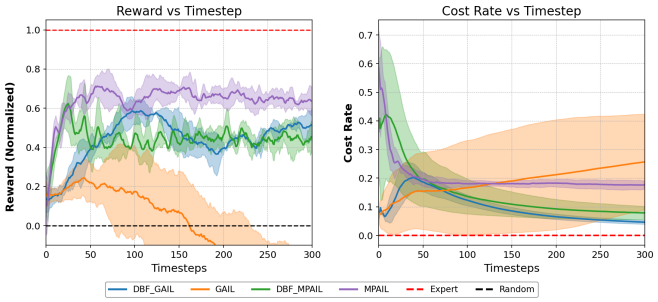
### *C. Q2: Safe Online Imitation Learning Performance Comparison and Safety Analysis*

For online performance evaluation we compare both imitation performance via normalized reward and safety via cost rate [1]: a metric that quantifies constraint-violation regret throughout learning in the same environment as Q1.

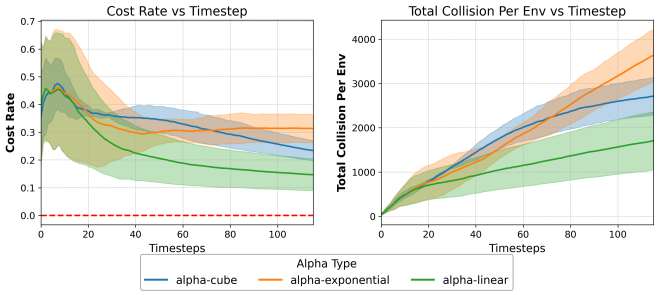
As shown in Figure 5a, DBF-based methods consistently reduce cost rate as learning progresses, while GAIL collapses entirely, failing to imitate the expert and accumulating high constraint violations due to failure to discriminate against unseen obstacle configurations absent from the expert demonstrations. MPAIL, as a planning-based method, shows greater robustness to out-of-distribution obstacle states represented as unseen obstacle occupancy than policy-based GAIL, but achieves higher cost rate than DBF-MPAIL. This is because DBF-based methods constrain the reward function search to satisfy the CBF dynamic constraint (Equation (5)), encouraging both policy and planning-based methods to learn safe imitative inhibitory behavior with minimal regret.

Notably, DBF-GAIL achieves the lowest cost rate at convergence, while DBF-MPAIL trades marginally higher cost rate for significantly better reward than DBF-GAIL. Although MPAIL achieves the highest reward, it does so at the cost of higher constraint violations compared to DBF-based methods, highlighting the safety-performance trade-off that DBFs are designed to address. AIRL and DBF-AIRL achieve comparable performance shown in Figure 8

[1]In all experiments involving baselines, the reward structure was $r(s, s')$, whereas in Section A-A, $r(s)$ was used to qualitatively distinguish each method.




**(a)** Normalized reward and cost rate.




**(b)** Effect of $\mathcal{K}$-class functions.

**Fig. 5:** Online learning evaluation across training seeds. (a) Normalized reward and cost rate comparison between DBF and non-DBF variants. (b) Trade-offs evaluated on DBF-MPAIL.

**Effect of different $\mathcal{K}$-class functions:** AIL addresses the occupancy matching problem relative to the expert distribution, with the learned reward serving as a dynamic constraint for the Control Barrier Function (CBF). Since this reward directly affects safe exploration, we quantify its impact using the cost rate and total collisions per environment. In our experiments, discriminator logit outputs remained in the range $[-0.5, 0.0]$ across all considered $\mathcal{K}$-class functions (linear, cubic, and exponential) in training environment Figure 4b. Within this range, the slopes of these functions follow the order linear $>$ cubic $>$ exponential. As the reward is learned, the cost-function logits for generated trajectories shift from large negative values toward zero. This progression is reflected in the safety metrics, including cost rate and total collisions per environment, as shown in Figure 5b. The figure indicates that cost rate decreases fastest for linear and slowest for exponential, consistent with their slope ordering. While total collisions per environment continue to accumulate throughout learning, linear accumulates them at the lowest rate, further corroborating its superior safety performance. Together, these results demonstrate that DBF-based AIL methods learn to discriminate against unseen unsafe states, enabling safe online imitation learning whose inhibitory behavior is further modulated by the choice of $\mathcal{K}$-class function.

### D. Q3: Real World Validation Through Real-Sim-Real Experiment

**Experiment Setup:** Our hardware experiment evaluates GAIL, MPAIL, DBF-GAIL, and DBF-MPAIL through a Real-to-Sim-to-Real pipeline: 1) a small number of state-only (position) observations are collected from the real world, 2) the method is trained using interactions from simulation with randomized obstacle positions, and 3) the method is deployed zero-shot in the real world for evaluation.

**Findings:** We find that DBF-MPAIL consistently trace path during inference not directly present in the expert demonstrations, as shown in Figure 1, to avoid obstacles while imitating the behavior needed to reach the goal. Unlike DBF-MPAIL, which plans over the barrier landscape, policy-based methods take positional state and a BEV map directly as input to a black-box policy. In addition, Figure 1 and Figure 6 illustrate a key advantage of planning-based DBF-AIL: success in imitating goal-reaching behavior while avoiding obstacles unseen in the expert data and at unseen positions in the training environment. By granting access to the agent's optimization landscape, DBF-MPAIL significantly improves the interpretability of agents trained on unlabeled and suboptimal data compared with black-box policies. We find that although the success rate of DBF-GAIL in reaching the goal is low due to the sim-to-real gap and the lack of structure in planning, it achieves zero collisions across all obstacle configurations, unlike the non-DBF baselines, as quantified in Table III.

At deployment time, baseline model-free AIL policies (GAIL and MPAIL without DBF) fail catastrophically on hardware, either drifting from the trajectory or executing severe collisions, which aligns with findings on real-world AIL brittleness [35]. As shown quantitatively in Table III, DBF-based methods drastically eliminate these collisions. However, a clear algorithmic trade-off emerges between policy-based and planning-based execution under learned safety regularizers. While DBF-GAIL achieves zero collisions, it does so at the cost of task completion, yielding only a $5\%$ success rate. Because model-free policy optimization lacks predictive look-ahead, an aggressive learned safety regularizer easily traps the policy in conservative local minima, effectively achieving safety via task stagnation. In contrast, by pairing the learned DBF regularizer with a predictive MPC planner, DBF-MPAIL successfully balances safety and task progress, navigating around physical obstacles to achieve a $65\%$ $(13/20)$ success rate with zero collisions.

### E. Data Collection in Maze Environment

To benchmark the AIL algorithms, we designed a maze environment and measured cumulative regret over the course of learning. Expert trajectories were collected by training PPO with a reward consisting of distance-to-goal and a collision penalty across diverse wall configurations. Using varied configurations produces a wider expert state-occupancy distribution, which is important for benchmarking: it allows both the policy and the planner to explore different trajectories when evaluated under randomized maze layouts. We save just state-only observations for training AIL algorithms.

**TABLE III:** Hardware deployment results (5 trials per variant across four obstacle configurations (Figure 1, Figure 6)). DBF variants achieve zero collisions across all trials. $d$ (m) denotes the mean minimum distance to obstacles throughout each trial.

| Variant | Succ. ↑ | Coll. ↓ | Min $d$ (m) ↑ |
|---|---|---|---|
| *Without DBF* | | | |
| GAIL | 0/20 (0%) | 10/20 (50%) | $0.035 \pm 0.274$ |
| MPAIL | 0/20 (0%) | 6/20 (30%) | $0.337 \pm 0.534$ |
| *With DBF* | | | |
| DBF-GAIL | 1/20 (5%) | **0/20 (0%)** | $0.500 \pm 0.437$ |
| DBF-MPAIL | **13/20 (65%)** | **0/20 (0%)** | $0.188 \pm 0.138$ |

### F. Comparison of AIRL and DBF-AIRL

Both AIRL and DBF-AIRL achieve comparable performance during training (Fig. 8) and evaluation (Table II). Although DBF-AIRL incurs marginally higher regret during training, we hypothesize that safety context for black-box policies can be more robustly embedded via a safety value function based on the Hamilton-Jacobi fixed-point approximation [24, 31], which we leave as future work. Furthermore, DBF-MPAIL and MPAIL share the same disentangled reward structure as AIRL—important for sim-to-real transfer—yet DBF-MPAIL clearly outperforms MPAIL in hardware evaluation (Table III), supporting the claim that DBF yields a more safety-aware discriminative reward.

## APPENDIX B
## PROOF OF THEOREMS AND IMPLEMENTATION DETAILS

*a) DBF as Cost-Regularized Observation-Only IRL.:* We now show how the DBF objective in Equation (5) can be interpreted as a restricted form of the entropy-regularized observation-only adversarial imitation learning formulation in Equation (1).

To start, suppose that

$$\mathcal{C}_{\text{CBF}} := \{c : c(s, s') = q_h(s, s'),\ h \in \mathcal{H}\} \subset \mathbb{R}^{S \times S} \tag{7}$$

is a nonempty, closed, convex class of cost functions parametrized by $h$.

Now recall that Proposition 3.2 in [17] states that for any proper, closed, convex regularizer $\psi$,

$$\text{RL} \circ \text{IRL}_\psi(\pi_E) = \arg\min_\pi [-H(\pi) + \psi^*(\rho_\pi - \rho_E)] \tag{8}$$

where $\text{IRL}_\psi(\pi_E) = \arg\max_{c \in \mathbb{R}^{\mathbb{S} \times \mathbb{A}}} -\psi(c) + \min_{\pi \in \Pi} -\lambda \mathbb{H}(\pi) + \mathbb{E}_\pi[c(s,a)] - \mathbb{E}_{\pi_E}[c(s,a)]$ and $\rho_\pi$, $\rho_E$ are the state-action occupancy measures of the policy and expert respectively. Following [37], we instead consider the state-transition occupancy measure

$$\rho_\pi(s, s') = \sum_a P(s' \mid s, a)\, \pi(a \mid s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \pi). \tag{9}$$

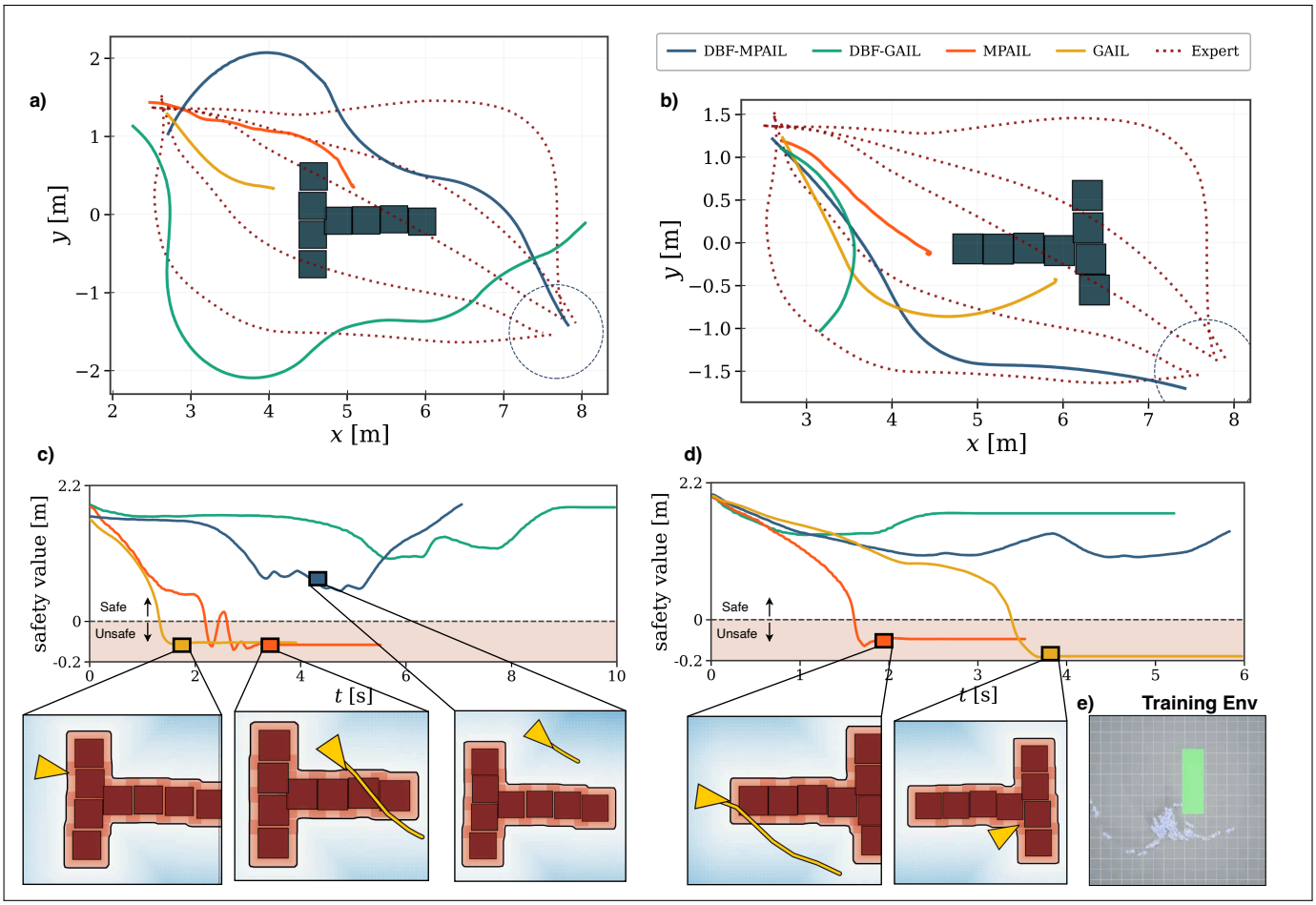



**Fig. 6:** Hardware deployment across two obstacle configurations. (a,b) Closed-loop trajectories for all variants with expert reference (dotted). (c,d) Safety value represented by minimum of distance to all obstacle. Insets show robot state at near-collision timesteps. (e) Shows the training environment with randomized obstacle position for this and Fig. 1.

While [37] derives the observation-only analogue of Equation 8 in the absence of entropy regularization, in practice we optimize the corresponding entropy-regularized objective

$$\text{RL} \circ \text{IRLfO}_{\psi}(\pi_E) = \arg\min_{\pi}[-H(\pi) + \psi^*(\rho_\pi - \rho_E)] \quad (10)$$

where IRLfO is defined in Equation 1 and $\rho_\pi$ and $\rho_E$ are now state-transition occupancy measures. This can be viewed as the natural entropy-regularized extension of the GAIfO objective.

Now consider the indicator cost function regularizer

$$\psi_{\text{CBF}}(c) = \begin{cases} 0, & \text{if} \quad c \in \mathcal{C}_{\text{CBF}} \\ \infty, & \text{otherwise} \end{cases} \quad (11)$$

Note that this regularizer is proper, closed, and convex by assumption on $\mathcal{C}_{\text{CBF}}$. Now recall that

$$\psi^*_{\text{CBF}}(u) = \sup_{c \in \mathbb{R}^{S \times S}} [\langle u, c \rangle - \psi_{\text{CBF}}(c)] \quad (12)$$

(13)

But since $\psi_{\text{CBF}}(c) = 0$ for all $c \in \mathcal{C}_{\text{CBF}}$ and is $\infty$ everywhere else, we obtain that

$$\psi^*_{\text{CBF}}(u) = \sup_{c \in \mathcal{C}_{\text{CBF}}} \langle u, c \rangle \quad (14)$$

Interpreting the discriminator as a parametrization of the cost function, we let $D_h := q_h$. Substituting $u = \rho_\pi - \rho_E$ and reparameterizing by $h \in \mathcal{H}$ yields

$$\psi^*_{\text{CBF}}(\rho_\pi - \rho_E) = \sup_{h \in \mathcal{H}} [\mathbb{E}_{\rho_\pi}[D_h(s, s')] - \mathbb{E}_{\rho_E}[D_h(s, s')]] \quad (15)$$

From our earlier observation that $\psi_{\text{CBF}}$ is proper, closed and convex we may apply Equation 10 and then obtain that

$$\text{RL} \circ \text{IRLfO}_{\psi}(\pi_E) = \arg\min_{\pi} \max_{h \in \mathcal{H}} [-H(\pi) + \mathbb{E}_{\rho_\pi}[D_h(s, s')] - \mathbb{E}_{\rho_E}[D_h(s, s')]] \quad (16)$$

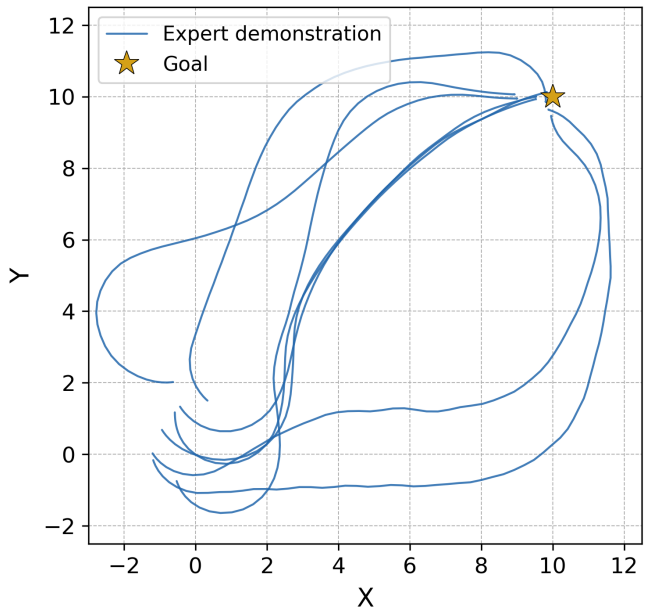



**Fig. 7:** Expert trajectories collected in the maze environment by training PPO with varied wall configurations and reward heuristics.

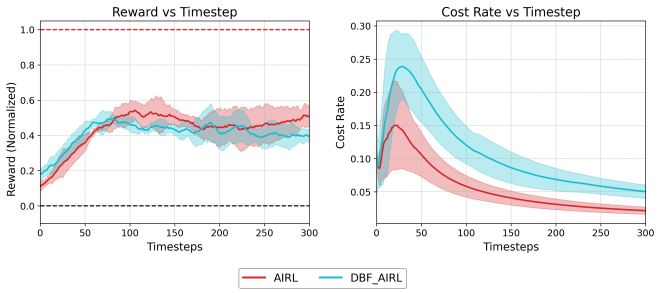



**Fig. 8:** Performance comparison between AIRL and DBF-AIRL during online learning. AIRL achieves both higher reward and lower cost rate than DBF-AIRL, suggesting that AIRL's inherently conservative reward structure ($\sigma^{-1} \circ D$) already provides implicit safety regularization in this setting. Additionally, DBF-AIRL consistently decreases cost rate as learning progresses whilst maintaining comparable task performance.

Furthermore, if we restrict the class of cost functions we consider to be 1-Lipschitz $\mathcal{C}_{\text{CBF}} \subset \{c : \|c\|_L \leq 1\}$ then the objective becomes an integral probability metric over the restricted CBF class. If additionally $\mathcal{C}_{\text{CBF}}$ equals the full 1-Lipschitz ball then by Kantorovich-Rubinstein duality we get that

$$\sup_{c \in \mathcal{C}_{\text{CBF}}} [\mathbb{E}_{\rho_\pi}[c(s,s')] - \mathbb{E}_{\rho_E}[c(s,s')]] = W_1(\rho_\pi, \rho_E)$$

where $W_1$ is the Wasserstein distance between the two distributions. Thus the discriminator objective corresponds to the Kantorovich dual formulation underlying WGANs.

**Remark:** In practice, we use the penalty term

$$\lambda_{\text{GP}} \mathbb{E}_{\hat{\pi}} \left[ (\|\nabla_{(s,s')} D_h(s,s')\|_2 - 1)^2 \right] \tag{17}$$

to encourage approximate 1-Lipschitz continuity. Also, the exact parametrized neural network class used to represent $\mathcal{C}_{\text{CBF}}$ is not guaranteed to be convex or closed. However, the idealized function class can be approximated by:

$$\mathcal{H}_{\text{CBF},\epsilon} := \{h \in \mathbb{R}^S : h(s) \geq \epsilon \; \forall s \in \mathcal{S}, \; h(s) \leq -\epsilon \; \forall s \in \mathcal{U}\} \tag{18}$$

$$\alpha(r) = \kappa r + \beta \tag{19}$$

$$\mathcal{Q}_{\text{CBF}} := \{c : c(s,s') = q_h(s,s'), \; h \in \mathcal{H}_{\text{CBF},\epsilon}\} \cap \{c : \|c\|_L \leq 1\} \tag{20}$$

Then $\mathcal{H}_{\text{CBF},\epsilon}$ is convex and closed as the intersection of half-spaces. It is also non-empty as long as $\mathcal{S}$ and $\mathcal{U}$ are disjoint and $\epsilon$ is small enough. Expanding the definition of $q_h$ yields $q_h(s,s') = h(s') + (\kappa - 1)h(s) + \beta$ which is an affine map. We already know the Lipschitz ball is closed and convex and the parametrized cost function set is also closed and convex since $h \mapsto q_h$ is affine. Thus, $\mathcal{Q}_{\text{CBF}}$ is also closed and convex. Assuming there exists at least one feasible $h$ such that the induced $q_h$ is 1-Lipschitz, a mild assumption given the continuous setting of many robotics tasks, the class is nonempty so the assumptions on $\mathcal{C}_{\text{CBF}}$ are satisfied.