WAIT OR TAKE OVER: FORECASTING OOD PERSISTENCE AND RECOVERY FOR EFFICIENT HUMAN-IN-THE-LOOP IMITATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Robot-gated interactive imitation learning allows an agent to request human intervention and learn from the resulting corrective demonstrations. Deciding when to request human takeover is central to using limited human effort effectively. Existing intervention mechanisms commonly rely on pointwise novelty, uncertainty, or disagreement, leaving an ambiguity: a novice outside human support may recover autonomously or continue along a persistent deviation. Intervening in the former case wastes limited human effort. We formulate robot-gated interactive imitation learning as a sequential intervention decision problem and develop Recovery-aware Gating for Interactive Imitation Learning (ReGIL). Drawing on Hamilton–Jacobi reachability analysis, we learn a support-persistence value that estimates whether the novice can remain within human support under its own policy. A short-horizon dynamics model predicts a return to support, while the persistence value assesses behavior after that return. A sequential recovery assessment combines these predictions with observed changes in persistence to determine when human intervention is warranted. This allows the novice to continue when autonomous recovery remains plausible and requests human takeover when predicted recovery is insufficient and observed persistence declines. Experiments in MetaDrive and MetaWorld demonstrate the effectiveness of the framework in learning autonomous policies from limited human intervention. The corresponding source code is available at `https://anonymous.4open.science/r/ReGIL-7864`.

1 INTRODUCTION

Human-in-the-loop (HIL) imitation learning enables agents to acquire task skills through repeated interaction with human experts (Leike et al., 2018; Everitt et al., 2017; Lee et al., 2021; Wang et al., 2021; Ouyang et al., 2022). A central motivation is that demonstrations collected in advance may provide limited guidance for situations encountered by a novice acting under its own policy, where small errors can compound over time (Ross et al., 2011). Human intervention allows the expert to demonstrate how to proceed from these situations, supplying corrections that reflect the novice's evolving behavior (Kelly et al., 2019; Peng et al., 2023). These corrections become part of the training data for subsequent autonomous execution. Deciding when to obtain them is therefore central to learning effectively from limited human intervention.

Human-in-the-loop imitation learning methods can therefore be categorized as human-gated or robot-gated according to whether the human expert or the agent decides when to initiate human intervention. Human-gated interactive imitation learning assigns this decision to the human expert, who monitors the novice and chooses when to take control (Kelly et al., 2019; Spencer et al., 2020; Li et al., 2022b; Peng et al., 2023; Ling et al., 2025). This allows the expert to direct the collection of corrective demonstrations, but requires continued attention even while the novice acts autonomously. The monitoring burden becomes difficult to sustain over long interactions, motivating robot-gated interactive imitation learning, in which the agent itself requests human assistance.

Robot-gated IIL reduces the need for continuous human monitoring by using an intervention mechanism to select when assistance is requested. The gate determines when the novice continues au-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

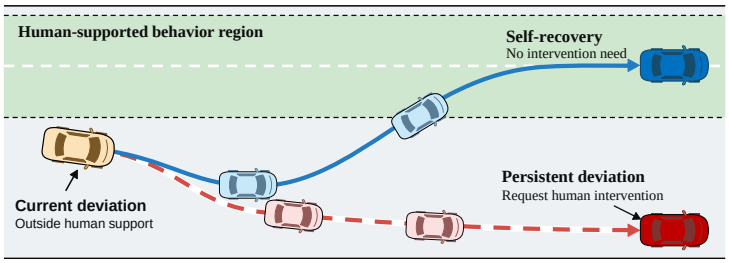

Figure 1: Motivation for recovery-aware intervention. A novice outside human support may recover autonomously (blue) or continue along a persistent deviation requiring human intervention (red). Robot-gated IIL should therefore consider the novice's future recovery prospects when deciding whether to request human takeover.

tonomously and when a human expert takes over to provide corrective demonstrations (Menda et al., 2019). Its decisions shape both the amount of human intervention data collected and the situations represented in that data. An effective gate must therefore balance the demand for human assistance with the learning benefit of the requested interventions.

However, existing robot-gated mechanisms commonly make this decision from a pointwise signal. Novelty, uncertainty, risk, or policy disagreement is evaluated at the current state-action pair and compared with a threshold (Biré et al., 2025; Menda et al., 2019; Hoque et al., 2021a; Zhang & Cho, 2016). This view creates a fundamental ambiguity:

> *A novice may temporarily leave the human-supported region and recover under its own closed-loop behavior, or it may enter a deviation that persists and progressively deteriorates.*

These two situations can produce similar instantaneous scores while requiring different intervention decisions. The central question is consequently not only whether the novice is currently outside human support, but whether it is likely to recover while continuing to act autonomously.

We address this ambiguity by formulating intervention as a sequential support-recovery decision. The relevant evidence includes whether the novice can return to human-supported behavior, whether it can remain supported after returning, and whether the observed interaction provides sufficient evidence to request human intervention.

Based on this view, we develop Recovery-aware Gating for Interactive Imitation Learning (ReGIL), a recovery-aware intervention framework for robot-gated IIL. We draw on Hamilton–Jacobi (HJ) reachability analysis to characterize whether trajectories remain within a prescribed region (Fisac et al., 2019; Ganai et al., 2023). Applying this perspective to human support, we learn a support-persistence value function that estimates whether the novice can remain within region supported by human data under its own policy. We then use a short-horizon model of novice dynamics to predict future support recovery. During interaction, a sequential recovery assessment combines the predicted recovery outcome with the evolution of support persistence as the novice continues to act. Human takeover is requested only when the current behavior is outside human support and the combined assessment indicates that autonomous recovery is no longer reliable. Our contributions are summarized as follows:

- We formulate robot-gated interactive imitation learning as a sequential intervention decision problem, extending pointwise gating to account for the novice's future closed-loop behavior.
- We introduce a recovery-aware intervention framework that integrates HJ-based support-persistence estimation and novice dynamics prediction within a sequential recovery assessment to distinguish deviations the novice can resolve autonomously from those requiring human intervention.
- Experiments in MetaDrive and MetaWorld demonstrate the effectiveness of our framework in learning from limited human intervention.

2 Related Work

2.1 Robot-Gated Interactive Imitation Learning

A long line of work studies how to incorporate human demonstrations into the imitation learning loop. Human-gated IIL grants the expert authority to take control during execution and steer the agent toward safer or more informative states. Human-Gated DAgger (HG-DAgger) (Kelly et al., 2019), a variant of DAgger (Ross et al., 2011), is the canonical instantiation: a human supervisor decides when to intervene, and the resulting demonstrations are aggregated into the training set to mitigate the compounding-error problem of behavior cloning (Ross & Bagnell, 2010). To improve data efficiency, some methods additionally exploit agent-generated trajectories as supplementary supervision: Expert Intervention Learning (EIL) (Spencer et al., 2020) and Intervention Weighted Regression (IWR) (Mandlekar et al., 2020) re-weight agent trajectories in a supervised manner, while HACO (Li et al., 2022b) and PVP (Peng et al., 2023) adopt off-policy reinforcement learning to jointly imitate human interventions and agent rollouts. A more recent line of work, FaithDaIL (Ling et al., 2025), further formulates the problem as dynamic regret minimization with an unbiased off-policy objective for faithful expert imitation. A common assumption across this branch, however, is that the *human* decides when to intervene. As a result, the gating burden scales linearly with the deployment horizon and becomes the dominant bottleneck for large-scale or long-horizon tasks, motivating the *robot-gated* alternative in which the agent itself requests help.

Robot-gated IIL approaches depend on an intervention mechanism, where the central challenge is to reduce the workload of human experts while still collecting enough informative data to train an effective imitation policy. Existing methods can be broadly organized by the *gating signal* they use to decide when to request intervention. A first line of work treats gating as a *discrepancy-classification* problem. SafeDAgger (Zhang & Cho, 2016) trains an auxiliary binary classifier that predicts whether the novice action would deviate from the expert beyond a safety threshold, and hands over control whenever it does. LazyDAgger (Hoque et al., 2021b) extends this idea with an explicit hysteresis between the novice-to-expert and expert-to-novice transitions, reducing the number of context switches and the cognitive load on the supervisor. A second line of work replaces the explicit classifier with *epistemic disagreement* signals derived from the novice itself. EnsembleDAgger (Menda et al., 2019) maintains an ensemble of behavior-cloning policies and uses the variance across their predicted actions as a Bayesian proxy for novice uncertainty. ThriftyDAgger (Hoque et al., 2021a) further combines such disagreement with a learned *risk* network—a Q-function estimating the probability of task failure under the novice—and calibrates both thresholds against an explicit intervention *budget* α_h. A third line of work casts gating as an *out-of-distribution detection* problem on the state space. RND-DAgger (Biré et al., 2025) adapts Random Network Distillation (Burda et al., 2019) to the IIL setting: a predictor network is trained on expert-visited states, and its prediction error on the current observation serves as a density-free novelty score that flags states substantially outside the expert distribution. Despite their different proxy signals, these pointwise gates base intervention on the current state–action pair and can mistake a transient departure for a persistent deviation, wasting human effort on cases where the novice would recover autonomously. One recent method, AIM (Cai et al., 2025), partially extends this pointwise view through temporal propagation with a Bellman backup, but recursively propagates a binary human-versus-novice classification rather than explicitly modeling future recovery or support persistence. Our method instead uses HJ reachability to estimate support persistence and combines it with predicted novice dynamics and sequential evidence to determine whether human takeover is warranted.

3 Hamilton-Jacobi Reachability Analysis

Hamilton–Jacobi (HJ) reachability analysis provides a principled framework for characterizing feasible sets from which system trajectories can remain within a prescribed region over time under admissible controls. Classical HJ methods compute these sets through the solution of Hamilton–Jacobi partial differential equations, but suffer from the curse of dimensionality due to grid-based computation. A number of methods have been proposed to improve scalability, including decomposition-based methods, efficient initialization and warm-starting, sampling-based reachability analysis, and neural approximations of HJ solutions (Chen et al., 2018; Herbert et al., 2019; Lew et al., 2022; Bansal & Tomlin, 2021). Recent works bridge HJ reachability and reinforcement learning through discounted dynamic-programming and Bellman-style formulations, enabling reachability values to

be learned from transition samples (Fisac et al., 2019; Hsu et al., 2021; Akametalu et al., 2023; Kokolakis et al., 2023). Building on this connection, RCRL incorporates HJ reachability values as state-wise feasibility constraints while optimizing task rewards (Yu et al., 2022), and RESPO further develops a probabilistic reachability formulation for stochastic reinforcement learning (Ganai et al., 2023). In this work, we use the complement of the reachability value to quantify whether novice behavior remains feasible within human support over time, yielding the support-persistence signal for recovery assessment and human intervention.

4 Problem Formulation

We consider human-in-the-loop imitation learning in a Markov Decision Process (MDP) , denoted by $\mathcal{M}$, and $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma, d_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, p is the transition probability function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\gamma \in (0, 1)$ is the discount factor, and d_0 is the distribution of initial state s_0. A policy $\pi(a|s)$ takes action a in $\mathcal{A}$ given state s in $\mathcal{S}$. Conventional Reinforcement Learning (RL) aims to find a policy π that maximizes the expected return: $\mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$, where $\mathbb{E}_\pi$ denotes the expectation under the distribution induced by $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim p(\cdot|s_t, a_t)$.

We denote the learning agent by a novice policy $\pi_N(a|s)$ and the human expert by $\pi_H(a|s)$. At each time step t, the novice first proposes an action $a_t^N \sim \pi_N(\cdot|s_t)$. An intervention gate determines whether the novice retains control or requests human assistance. If intervention is triggered, the expert provides $a_t^H \sim \pi_H(\cdot|s_t)$; otherwise, a_t^N is executed. Accordingly, the action applied to the environment is $a_t = g_t a_t^H + (1 - g_t)a_t^N$, where $g_t \in \{0, 1\}$ denotes the intervention decision. The novice policy is updated by behavioral cloning on the accumulated human intervention dataset $\mathcal{D}_H = \{(s_t, a_t^H) : g_t = 1\}$, i.e., $\pi_N = \arg\max_\pi \mathbb{E}_{(s,a^H)\sim\mathcal{D}_H}\left[\log \pi(a^H \mid s)\right]$.

Crucially, whether intervention is necessary need not be determined solely by the current state-action pair. Intervention is inherently a temporal decision: an action that appears unfamiliar at the current step may lead the novice back to well-supported behavior, while a seemingly familiar state-action pair may initiate a trajectory that progressively departs from prior human experience. Therefore, the relevant question for gating is not only *where the novice is now*, but also *where its own closed-loop behavior is likely to lead*. We accordingly formulate intervention as a decision over the novice trajectory initialized from the current state-action pair:

$$g_t = G\big(\tau_t^N(s_t, a_t^N)\big), \qquad G : \mathcal{T} \to \{0, 1\}, \tag{1}$$

where $\tau_t^N(s_t, a_t^N) = \big((s_t, a_t^N), (s_{t+1}, a_{t+1}^N), \dots\big)$ denotes the closed-loop trajectory generated by the novice policy π_N and environment dynamics P starting from (s_t, a_t^N), and $\mathcal{T}$ denotes the corresponding trajectory space. In stochastic environments, τ_t^N is a random trajectory induced by π_N and P, and the gate therefore reasons over its conditional future behavior rather than a single deterministic rollout.

Since $\tau_t^N(s_t, a_t^N) = \big((s_t, a_t^N), \tau_{t+1:}^N\big)$, the proposed formulation jointly captures the current state-action pair and the novice's future closed-loop evolution. Conventional pointwise gates $\widetilde{G} : \mathcal{S}\times\mathcal{A} \to \{0, 1\}$ can be viewed as a special case obtained by ignoring $\tau_{t+1:}^N$, making them a restricted subclass of Eq. 1.

5 Method

Our key insight is that human intervention should not be determined solely by whether the novice is currently outside the support of prior human demonstrations. A transient deviation may still be recoverable under the novice's own closed-loop behavior, whereas intervention is more critical when the novice is unlikely to return to, or subsequently remain within, a human-supported region. Motivated by this distinction, we formulate gating as a temporal support-recovery problem rather than a pointwise OOD detection problem. ReGIL consists of two components. We first learn a *support-persistence value* that estimates whether the novice will continue to remain within human-supported behavior from a given state-action pair. We then combine this long-horizon estimate with a learned dynamics model to predict short-horizon recovery from out-of-support states, yielding a dynamics-aware intervention gate that intervenes only when autonomous recovery is unlikely.

5.1 From Pointwise Support Estimation to Support Persistence

Existing intervention mechanisms commonly rely on pointwise novelty or uncertainty estimates of the current state-action pair. However, an out-of-support observation does not necessarily require human intervention: the novice may only deviate transiently and subsequently recover. To reason beyond instantaneous support membership, we introduce a policy-conditioned value that characterizes whether the novice can *persist* within the human-supported region under its future closed-loop behavior.

Human support. Let $\mathcal{H} \subseteq \mathcal{S} \times \mathcal{A}$ denote a support region estimated from the human expert dataset $\mathcal{D}_H$. This region represents behavior supported by the available demonstrations. We write the support indicator as

$$m_{\mathcal{H}}(s,a) := \mathbb{I}[(s,a) \in \mathcal{H}]\,. \tag{2}$$

The support indicator provides the interface between the estimated human support and the persistence value.

Support persistence value. The support indicator describes whether the current novice state-action pair is covered by human demonstrations, but it does not reveal whether the novice will remain in that covered region. For a fixed novice policy π_N and support region $\mathcal{H}$, we define the *support-persistence value* as

$$V_{\mathcal{H}}^{\pi_N}(s,a) := \Pr_{\pi_N,P}\big((s_k,a_k) \in \mathcal{H} \text{ for all } k \geq 0 \,|\, s_0 = s,\ a_0 = a\big)\,. \tag{3}$$

After the initial action a, subsequent actions follow π_N. A low value indicates that the current novice behavior is already outside the demonstrated support or is likely to leave it under its own future actions, making the state a candidate for further recovery assessment and possible human intervention. Following the one-step decomposition used in probabilistic reachability formulations (Ganai et al., 2023), the trajectory event can be decomposed as

$$\begin{aligned} V_{\mathcal{H}}^{\pi_N}(s,a) &= m_{\mathcal{H}}(s,a)\,\mathbb{E}_{s',a'}[V_{\mathcal{H}}^{\pi_N}(s',a')] \\ &= \min\{m_{\mathcal{H}}(s,a),\ \mathbb{E}_{s',a'}[V_{\mathcal{H}}^{\pi_N}(s',a')]\}\,, \end{aligned} \tag{4}$$

where $s' \sim P(\cdot \mid s,a)$ and $a' \sim \pi_N(\cdot \mid s')$. The first line follows by conditioning on the next state-action pair, and the second uses $m_{\mathcal{H}}(s,a) \in \{0,1\}$: an unsupported pair has zero persistence, whereas a supported pair inherits the expected persistence of its novice-controlled continuation.

The same recursion has a reachability interpretation. Its complement, $\Phi_{\mathcal{H}}^{\pi_N} = 1 - V_{\mathcal{H}}^{\pi_N}$, is the probability of ever leaving $\mathcal{H}$ and satisfies the corresponding max recursion, $\Phi_{\mathcal{H}}^{\pi_N}(s,a) = \max\{1 - m_{\mathcal{H}}(s,a), \mathbb{E}_{s',a'}[\Phi_{\mathcal{H}}^{\pi_N}(s',a')]\}$.

Learning from replay. We use a network V_ψ with outputs in $[0,1]$ and a target network $V_{\bar{\psi}}$ to approximate the persistence value. Let $\mathcal{D}_R$ contain transitions collected during both novice and human control. For $(s,a,s') \sim \mathcal{D}_R$, we sample $a' \sim \pi_N(\cdot \mid s')$ and form the target and regression loss

$$\begin{aligned} y(s,a,s') &= \min\big\{m_{\mathcal{H}}(s,a),\ V_{\bar{\psi}}(s',a')\big\}\,, \\ \mathcal{L}_V(\psi) &= \mathbb{E}_{\substack{(s,a,s')\sim\mathcal{D}_R \\ a'\sim\pi_N(\cdot|s')}}\Big[\big(V_\psi(s,a) - y(s,a,s')\big)^2\Big]\,. \end{aligned} \tag{5}$$

The replay action supplies the observed transition, while the next action is drawn from the novice to evaluate its continuation. The definition above holds the policy and support region fixed; during online learning, both change as new human demonstrations are collected, so the network must track an evolving target. In experiments, we use a discounted, terminal-aware target, detailed in the Appendix.

5.2 Dynamics-Aware Recovery Gating

The support-persistence value characterizes whether the novice can remain within human support. To decide whether to wait or request human assistance after a deviation, we combine short-horizon dynamics prediction with a sequential recovery assessment. This assessment considers both the predicted return to support and the observed evolution of persistence estimates to determine when to request human intervention.

Prospective novice dynamics. The persistence value evaluates behavior after a trajectory has returned to human support, but it does not determine whether such a return will occur from the current deviation. We therefore learn a one-step predictive model $\hat{f}_\omega$ from replay transitions and use it to estimate the short-term consequences of continuing with the novice. Given the current observation s_t and proposed novice action a_t^N, the predicted trajectory is generated by

$$\begin{aligned} \hat{s}_{t+1} &= \hat{f}_\omega(s_t, a_t^N), \\ \hat{a}_{t+k} &= \pi_N(\hat{s}_{t+k}), \qquad k = 1, \dots, H, \\ \hat{s}_{t+k+1} &= \hat{f}_\omega(\hat{s}_{t+k}, \hat{a}_{t+k}), \quad k = 1, \dots, H-1. \end{aligned} \tag{6}$$

The model is trained with a transition-prediction objective,

$$\mathcal{L}_{\mathrm{dyn}}(\omega) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}_R}[\ell(\hat{f}_\omega(s,a), s')], \tag{7}$$

where ℓ is the state-prediction loss. This rollout is used to estimate whether the novice can resolve the current deviation within a short horizon; it is not treated as an additional source of human supervision.

Sequential recovery assessment. We assess whether to allow further autonomous recovery or end the current period of novice control. A predicted return to a region with high support persistence favors continued autonomy. When predicted recovery is insufficient, we request assistance once the observed persistence estimates exhibit sustained decline. To evaluate the recovery prospect, we locate the first predicted return to human support:

$$j_t := \min\left\{k \in \{1, \dots, H\} : m_{\mathcal{H}}(\hat{s}_{t+k}, \hat{a}_{t+k}) = 1\right\}, \tag{8}$$

with $j_t = \varnothing$ if no such step exists. The predicted support-recovery value is

$$\rho_t := \begin{cases} V_\psi(\hat{s}_{t+j_t}, \hat{a}_{t+j_t}), & j_t \neq \varnothing, \\ 0, & j_t = \varnothing. \end{cases} \tag{9}$$

The score ρ_t evaluates persistence at the predicted point of re-entry, combining short-horizon recovery with the longer-term support estimate.

We update this prediction as the novice continues and assess it together with the observed change in persistence, allowing the gate to wait while recovery remains plausible and to request assistance when the evidence deteriorates. Let $v_i := V_{\psi_i}(s_i, a_i^N)$ be the persistence estimate recorded at step i, using the value network available at that step, and write $v_{t-K:t} := (v_{t-K}, \dots, v_t)$ for the recent sequence of estimates. For a recovery threshold η, a history length K, and a small tolerance $\epsilon \geq 0$, we define the sequential recovery assessment function $\mathcal{R}$ as

$$\mathcal{R}(\rho_t, v_{t-K:t}) := \mathbb{I}\left[(\rho_t < \eta) \ \wedge \ \left(\min_{i=t-K+1,\dots,t} \big(v_{i-1} - v_i\big) > \epsilon\right)\right]. \tag{10}$$

Here, when the current state-action pair is outside human support, $\mathcal{R}(\rho_t, v_{t-K:t}) = 1$ indicates that human intervention is warranted: the prediction offers insufficient support for recovery, and every observed one-step decrease in the recent window exceeds ϵ. The history follows consecutive novice-controlled transitions across support boundaries. It is cleared at episode reset and restarted from the release state after human control; the function returns zero until K consecutive differences are available. A zero output allows the novice to continue, with the assessment updated at the next step.

Recovery-aware intervention decision. While the novice controls the system, the support indicator identifies deviations for which the sequential assessment may request assistance. A new takeover is requested when the current state-action pair is outside human support and the assessment calls for human intervention:

$$g_t = \mathbb{I}\big[\big(m_{\mathcal{H}}(s_t, a_t^N) = 0\big) \ \wedge \ (\mathcal{R}(\rho_t, v_{t-K:t}) = 1)\big]. \tag{11}$$

The gate permits continued autonomous execution and reassessment until an out-of-support state-action pair satisfies the sequential intervention criterion. Once human control begins, it is maintained until the mean squared difference between the novice and human actions falls below the calibrated release threshold.

Algorithm 1 ReGIL: Recovery-Aware Gating for Interactive Imitation Learning

Require: Human expert π_H, rollout horizon H, assessment parameters K, ϵ
Collect warm-up demonstrations in $\mathcal{D}_H$ and transitions in replay $\mathcal{D}_R$
Initialize π_N, support estimator $m_{\mathcal{H}}$, dynamics $\hat{f}_\omega$, and value V_ψ from warm-up data
Calibrate support and recovery thresholds and the fixed release threshold τ_a
for trajectory $i = 1, 2, \ldots$ **do**
  for timestep $t = 1, 2, \ldots$ **do**
    Sample novice action $a_t^N \sim \pi_N(s_t)$
    Record the current persistence estimate and evaluate g_t by Eq. (11), using Eqs. (9)–(10)
    if $g_t = 1$ **then**
      Sample human action $a_t^H \sim \pi_H(s_t)$
      repeat
        Observe $s_{t+1} \sim p(\cdot \mid s_t, a_t^H)$.
        Add (s_t, a_t^H, s_{t+1}) to $\mathcal{D}_R$ and $\mathcal{D}_H$
        Train π_N by BC and update $m_{\mathcal{H}}$ from $\mathcal{D}_H$
        $t \leftarrow t + 1$;
        Sample human action $a_t^H \sim \pi_H(s_t)$
        Sample novice action $a_t^N \sim \pi_N(s_t)$
      until $\|a_t^N - a_t^H\|^2 < \tau_a$
    else
      Execute a_t^N; observe $s_{t+1} \sim p(\cdot \mid s_t, a_t^N)$
      Add (s_t, a_t^N, s_{t+1}) to $\mathcal{D}_R$
    end if
    At each timestep, train $\hat{f}_\omega$ by Eq. 7 and V_ψ by Eq. 5 on $\mathcal{D}_R$
    Recalibrate support and recovery thresholds
  end for
end for
return π_N

6 Experiments

6.1 Experimental Setup

Environments. We evaluate autonomous driving in MetaDrive (Li et al., 2022a) and robotic manipulation on MetaWorld ML1 `drawer-open-v3` (McLean et al., 2025). MetaDrive uses 50 training and 50 separate test scenarios. MetaWorld uses 40 training instances and 10 development instances; the current comparison evaluates the latter with five reset seeds each. In both environments, interventions are simulated: MetaDrive uses a neural expert trained in its training environment with the human-gated FaithDaIL method (Ling et al., 2025), while MetaWorld uses its built-in task-solving scripted expert.

Baselines. We compare with EnsembleDAgger (Menda et al., 2019), ThriftyDAgger (Hoque et al., 2021a), RND-DAgger (Biré et al., 2025), and AIM (Cai et al., 2025). All methods use behavioral cloning on expert data with a common policy architecture and update schedule, while retaining method-specific intervention rules.

Evaluation metrics. *Human Data* counts expert-controlled transitions, including initialization. We report autonomous success rate in both environments and episode return in MetaDrive. *Completion* denotes the fraction of the route traversed in MetaDrive (higher is better) and normalized completion time in MetaWorld (lower is better). The latter measures steps to five consecutive successes divided by the 500-step horizon, assigning 1 to unfinished episodes. Budget checkpoints, final data usage, and aggregation rules are specified in the tables.

Implementation Details. MetaDrive runs for 8,000 interaction steps, including 200 initial expert transitions; MetaWorld runs for 1,500 online steps after 50 expert transitions. Each method is trained using four different random seeds. Policies use two 512-unit hidden layers and Adam with learning rate 10^{-4}, updating every 25 new expert transitions. Our support indicator thresholds state-based RND error (Burda et al., 2019) at 1.75 times its mean on expert observations. The state-based persistence estimator and residual dynamics model learn from mixed expert–novice replay. Recovery assessment uses $H = 10$, $K = 2$, and $\epsilon = 10^{-6}$. More details are provided in the Appendix.

Table 1: MetaDrive results using an human-data budget of 1,000 transitions and a total training budget of 8,000 steps. For runs that do not exhaust the human-data budget, we report the performance of the final policy.

Method	**Under Human-Data Budget**				**Final Results**	
	Human Data ↓	**Success ↑**	**Return ↑**	**Completion ↑**	**Human Data ↓**	**Success ↑**
EnsembleDAgger	1000.0 ±0.0	0.672 ±0.234	313.9 ±44.8	0.837 ±0.115	2253.6 ±262.9	0.868 ±0.059
ThriftyDAgger	1000.0 ±0.0	0.305 ±0.264	225.8 ±72.9	0.611 ±0.188	4591.0 ±453.9	0.855 ±0.072
RND-DAgger	1000.0 ±0.0	0.535 ±0.184	289.5 ±40.2	0.778 ±0.113	3348.0 ±432.8	0.890 ±0.012
AIM	1000.0 ±0.0	0.745 ±0.075	**336.1 ±6.9**	0.892 ±0.015	2316.3 ±256.0	0.855 ±0.050
ReGIL (Ours)	**945.5 ±75.9**	**0.800 ±0.063**	335.2 ±14.8	**0.898 ±0.031**	**1006.8 ±141.7**	0.830 ±0.066

Table 2: MetaWorld drawer-opening results over 50 development episodes, using an human-data budget of 400 transitions and a total training budget of 1,500 steps. For runs that do not exhaust the human-data budget, we report the performance of the final policy.

Method	**Under Human-Data Budget**			**Final Results**	
	Human Data ↓	**Success ↑**	**Completion ↓**	**Human Data ↓**	**Success ↑**
EnsembleDAgger	400.0 ±0.0	0.700 ±0.476	0.4320 ±0.3924	590.0 ±96.3	1.000 ±0.000
ThriftyDAgger	400.0 ±0.0	0.475 ±0.377	0.6131 ±0.3088	769.8 ±303.7	1.000 ±0.000
RND-DAgger	397.0 ±6.0	0.900 ±0.200	0.2567 ±0.1822	503.5 ±109.3	1.000 ±0.000
AIM	400.0 ±0.0	0.825 ±0.287	0.3521 ±0.2320	723.8 ±161.2	1.000 ±0.000
ReGIL (Ours)	**393.8 ±12.5**	**1.000 ±0.000**	**0.2048 ±0.0671**	**425.8 ±46.1**	1.000 ±0.000

6.2 RESULTS

Table 1 compares autonomous driving policies learned through robot-gated human intervention. Under the human-data budget, ReGIL achieves the highest mean success and route completion, with return close to the best baseline, while using the least human intervention data on average. This comparison addresses the central goal of robot-gated IIL: deciding when to request human takeover so that the collected corrections help the novice learn to act independently. The results support accounting for autonomous recovery when selecting these interventions, allowing effective driving behavior to be learned from a smaller amount of human-controlled experience.

At the end of training, the baselines attain higher success while collecting substantially more human intervention data. Our method maintains the lowest human-data usage, with its advantage in success most apparent under the intervention-data budget. This distinction matters when human experts have limited availability to take control and provide corrective actions. Evaluating both budgeted and final policies reveals how much human-controlled experience is required to obtain the resulting autonomous behavior.

Table 2 shows a similar pattern on the MetaWorld drawer-opening development set. Under the human-data budget, our method achieves the highest success and the shortest normalized completion time among the reported runs. The learned novice therefore completes the manipulation task more reliably and efficiently without human assistance during evaluation. Our method also uses the least human intervention data by the end of training. The budgeted comparison distinguishes how effectively the requested human corrections support learning before that endpoint. Across driving and manipulation, the results favor considering the novice's recovery prospects when deciding whether human takeover is needed.

6.3 ABLATION STUDY

Table 3 examines how prospective dynamics, the HJ value formulation, and sequential evidence contribute to deciding when the novice should retain control or request human takeover. We con-

Table 3: MetaDrive ablations after 8,000 training steps. Values are mean ± sample standard deviation.

Variant	Human Data Usage ↓	Success ↑	Return ↑	Route ↑
ReGIL (Ours)	1006.8 ± 141.7	0.830 ± 0.066	342.0 ± 17.3	0.914 ± 0.039
w/o Dynamic Model	963.8 ± 162.2	0.696 ± 0.181	319.1 ± 30.5	0.847 ± 0.084
w/o HJ Formulation	1211.6 ± 245.6	0.856 ± 0.067	348.5 ± 11.0	0.928 ± 0.030
w/o Sequential Evidence	1569.8 ± 534.3	0.800 ± 0.157	334.5 ± 32.1	0.896 ± 0.080

sider human intervention data together with autonomous driving performance to assess whether each variant reduces reliance on human control while preserving effective policy learning.

The *w/o Dynamic Model* variant replaces the predicted support-recovery value with the current persistence estimate. It collects slightly less human intervention data than the full method but yields lower success, return, and route completion. A low current estimate indicates that the novice may struggle to remain within human support, but does not resolve whether its subsequent actions can bring it back to behavior covered by human demonstrations. Dynamics prediction provides this recovery prospect before the gate requests takeover. The weaker driving performance of the current-value variant is consistent with the importance of considering where continued novice control may lead when selecting human interventions.

The *w/o HJ Formulation* variant replaces the binary support signal and HJ-style min backup with a continuous normalized RND cost and a standard Bellman backup, retaining dynamics prediction and sequential intervention assessment. It achieves somewhat higher final driving performance while collecting more human intervention data. The HJ formulation assesses whether the novice can remain within human-supported behavior after a predicted return, providing a criterion for allowing continued autonomy. The alternative bases this decision on accumulated future novelty. Their comparison shows a trade-off between the amount of human-controlled correction collected during training and the performance of the resulting autonomous policy, with the HJ formulation requiring less intervention data.

The *w/o Sequential Evidence* variant retains recovery prediction but requests human takeover whenever predicted recovery is insufficient, without requiring sustained decline in observed persistence. It collects substantially more human intervention data without improving mean driving performance. In the full method, an unfavorable recovery prediction is assessed together with how the novice's behavior evolves during autonomous execution. This gives the novice an opportunity to continue before human control is requested, while sustained decline supplies additional evidence that assistance is needed. The observed data usage and policy performance are consistent with this sequential assessment avoiding premature human interventions while retaining useful corrective demonstrations.

7 Conclusion

Robot-gated interactive imitation learning requires deciding when a novice's departure from human-supported behavior warrants human takeover and when continued autonomy may suffice. We formulated this choice as a sequential intervention decision and developed ReGIL, a recovery-aware gate that combines HJ-based support-persistence estimation, short-horizon novice dynamics prediction, and observed changes in persistence. The resulting assessment requests human intervention when predicted recovery is insufficient and the novice's behavior continues to deteriorate, while allowing transient deviations time to resolve.

In MetaDrive and MetaWorld, our method learned effective autonomous policies with less human intervention data than the compared gates, with its clearest performance advantage under fixed human-data budgets. The MetaDrive ablations further show that prospective dynamics supports policy performance, while the HJ formulation and sequential evidence help limit human-controlled data collection. Together, these results support evaluating the novice's recovery prospects over time when deciding which deviations require human correction.

8 AI USE STATEMENT

We used generative AI tools only to aid and polish writing and to assist with retrieval and discovery of related work. We did not use generative AI for developing the methodology, conducting experiments, or generating scientific conclusions. All AI-assisted content was reviewed by the authors, and all references were independently verified. The authors take full responsibility for the final content of the paper.

REFERENCES

Anayo K Akametalu, Shromona Ghosh, Jaime F Fisac, Vicenc Rubies-Royo, and Claire J Tomlin. A minimum discounted reward hamilton–jacobi formulation for computing reachable sets. *IEEE Transactions on Automatic Control*, 69(2):1097–1103, 2023.

Somil Bansal and Claire J Tomlin. Deepreach: A deep learning approach to high-dimensional reachability. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1817–1824. IEEE, 2021.

Emilien Biré, Anthony Kobanda, Ludovic Denoyer, and Rémy Portelas. Efficient active imitation learning with random network distillation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GFgn2LprFR.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.

Haoyuan Cai, Zhenghao Peng, and Bolei Zhou. Robot-gated interactive imitation learning with adaptive intervention mechanism. *ArXiv*, abs/2506.09176, 2025. URL https://api.semanticscholar.org/CorpusID:279305793.

Mo Chen, Sylvia L Herbert, Mahesh S Vashishtha, Somil Bansal, and Claire J Tomlin. Decomposition of reachable sets and tubes for a class of nonlinear systems. *IEEE Transactions on Automatic Control*, 63(11):3675–3688, 2018.

Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. Reinforcement learning with a corrupted reward channel. *arXiv preprint arXiv:1705.08417*, 2017.

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)*, pp. 8550–8556. IEEE, 2019.

Milan Ganai, Zheng Gong, Chenning Yu, Sylvia Herbert, and Sicun Gao. Iterative reachability estimation for safe reinforcement learning. *Advances in Neural Information Processing Systems*, 36:69764–69797, 2023.

Sylvia L Herbert, Somil Bansal, Shromona Ghosh, and Claire J Tomlin. Reachability-based safety guarantees using efficient initializations. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pp. 4810–4816. IEEE, 2019.

Ryan Hoque, Ashwin Balakrishna, Ellen Novoseller, Albert Wilcox, Daniel S Brown, and Ken Goldberg. Thriftydagger: Budget-aware novelty and risk gating for interactive imitation learning. *arXiv preprint arXiv:2109.08273*, 2021a.

Ryan Hoque, Ashwin Balakrishna, Carl Putterman, Michael Luo, Daniel S Brown, Daniel Seita, Brijen Thananjeyan, Ellen Novoseller, and Ken Goldberg. Lazydagger: Reducing context switching in interactive imitation learning. In *2021 IEEE 17th international conference on automation science and engineering (case)*, pp. 502–509. IEEE, 2021b.

Kai-Chieh Hsu, Vicenç Rubies-Royo, Claire J Tomlin, and Jaime F Fisac. Safety and liveness guarantees through reach-avoid reinforcement learning. *arXiv preprint arXiv:2112.12288*, 2021.

Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Hg-dagger: Interactive imitation learning with human experts. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8077–8083. IEEE, 2019.

Nick-Marios Kokolakis, Kyriakos G Vamvoudakis, and Wassim Haddad. Reachability analysis-based safety-critical control using online fixed-time reinforcement learning. In *Learning for Dynamics and Control Conference*, pp. 1257–1270. PMLR, 2023.

Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*, 2018.

Thomas Lew, Lucas Janson, Riccardo Bonalli, and Marco Pavone. A simple and efficient sampling-based algorithm for general reachability analysis. In *Learning for Dynamics and Control Conference*, pp. 1086–1099. PMLR, 2022.

Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3461–3475, 2022a.

Quanyi Li, Zhenghao Peng, and Bolei Zhou. Efficient learning of safe driving policy via human-ai copilot optimization. *arXiv preprint arXiv:2202.10341*, 2022b.

Bo Ling, Zhengyu Gan, Wanyuan Wang, Guanyu Gao, Weiwei Wu, and Yan Lyu. Faithful dynamic imitation learning from human intervention with dynamic regret minimization. *Advances in Neural Information Processing Systems*, 38:132251–132270, 2025.

Ajay Mandlekar, Danfei Xu, Roberto Martín-Martín, Yuke Zhu, Li Fei-Fei, and Silvio Savarese. Human-in-the-loop imitation learning using remote teleoperation. *arXiv preprint arXiv:2012.06733*, 2020.

Reginald McLean, Evangelos Chatzaroulas, Luc McCutcheon, Frank Röder, Tianhe Yu, Zhanpeng He, K.R. Zentner, Ryan Julian, J K Terry, Isaac Woungang, Nariman Farsad, and Pablo Samuel Castro. Meta-world+: An improved, standardized, RL benchmark. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=1de3azE606.

Kunal Menda, Katherine Driggs-Campbell, and Mykel J Kochenderfer. Ensembledagger: A bayesian approach to safe imitation learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5041–5048. IEEE, 2019.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Zhenghao Mark Peng, Wenjie Mo, Chenda Duan, Quanyi Li, and Bolei Zhou. Learning from active human involvement through proxy value propagation. *Advances in neural information processing systems*, 36:77969–77992, 2023.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668. JMLR Workshop and Conference Proceedings, 2010.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Jonathan Spencer, Sanjiban Choudhury, Matthew Barnes, Matthew Schmittle, Mung Chiang, Peter Ramadge, and Siddhartha Srinivasa. Learning from interventions: Human-robot interaction as both explicit and implicit feedback. In *16th robotics: science and systems, RSS 2020*. MIT Press Journals, 2020.

Zizhao Wang, Xuesu Xiao, Garrett Warnell, and Peter Stone. Apple: Adaptive planner parameter learning from evaluative feedback. *IEEE Robotics and Automation Letters*, 6(4):7744–7749, 2021.

Dongjie Yu, Haitong Ma, Shengbo Li, and Jianyu Chen. Reachability constrained reinforcement learning. In *International conference on machine learning*, pp. 25636–25655. PMLR, 2022.

Jiakai Zhang and Kyunghyun Cho. Query-efficient imitation learning for end-to-end autonomous driving. *arXiv preprint arXiv:1605.06450*, 2016.