WHEN INTERACTION CHANGES CONTROLLABILITY: PHYSICS-INFORMED GOAL-CONDITIONED REINFORCEMENT LEARNING FOR MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning to reach arbitrary goals requires agents to infer a rich notion of reachability across state-goal pairs, which becomes increasingly difficult over long decision horizons and in systems with complex dynamics. Physics-informed Goal-Conditioned Reinforcement Learning (Pi-GCRL) addresses this problem primarily via Eikonal-based losses, achieving strong results in navigation while remaining less effective in robotic manipulation. In this paper, we show that this gap stems from the state-dependent controllability of manipulation, where object coordinates are not locally controllable throughout the task. We first establish that full-state Eikonal regularization becomes inconsistent whenever goal-relevant coordinates are locally uncontrollable. This insight leads to two complementary approaches: a loss that regularizes value gradients only along feasible directions, and a hierarchical decomposition that restricts physics-informed learning to a controllable low-level space. We evaluate both approaches on offline manipulation benchmarks and a physical robot, showing substantial improvements over naive full-state Eikonal regularization and demonstrating that accounting for local controllability improves the effectiveness of Pi-GCRL in robotic manipulation.

1 INTRODUCTION

Reinforcement Learning (RL) offers a powerful framework for learning sequential decision-making policies from experience (Sutton & Barto, 2018), but realizing its potential remains challenging, in part due to the complexity of reward design that often requires substantial fine-tuning, simulation access and trial-and-error (Ibrahim et al., 2024). Sparse rewards reduce this burden, but often make the learning problem considerably more difficult (Giammarino et al., 2024).

This tension has motivated increasing interest in Goal-Conditioned RL (GCRL) (Kaelbling, 1993; Schaul et al., 2015), which shifts the focus from designing task-specific rewards to learning reusable goal-reaching behaviors from sparse supervision. Rather than learning a policy for a single task, GCRL conditions policies and value functions on desired goals, allowing a single agent to reason about how to reach many possible outcomes from different initial states. This formulation also enriches the available supervision, as visited states can serve as alternative goals, effectively transforming sparse feedback into a richer set of learning signals (Andrychowicz et al., 2017). Under this paradigm, the learning challenge shifts toward inferring the state-goal reachability structure induced by the environment and the system dynamics, including which goals are reachable from a given state, through which feasible transitions, and at what cost.

Physics-informed GCRL (Pi-GCRL) directly targets this reachability problem by regularizing value functions with optimal-control-inspired partial differential equations (PDEs) (Settai et al., 2026; Giammarino et al., 2026; Giammarino & Qureshi, 2026). Rather than learning an arbitrary function over state-goal pairs, Pi-GCRL encourages the value landscape to behave like a physically meaningful cost-to-go. Viewed this way, Pi-GCRL provides an inductive bias that aligns value learning with the underlying control dynamics without requiring any explicit reward-shaping procedure.

So far, Pi-GCRL has shown promising results in navigation and object-free goal-reaching tasks, primarily through Eikonal regularization (Giammarino et al., 2026; Giammarino & Qureshi, 2026). Eikonal-based losses provide a convenient way to inject reachability structure into value learning,

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.

owing to the simplicity and tractability of the Eikonal PDE. However, these benefits do not appear to transfer as readily to goal-conditioned manipulation. In this paper, we study this limitation and trace it back to the entity-centric nature of robotic manipulation (Haramati et al., 2026), where the state space includes coordinates describing both the robot and the objects it must control. Unlike robot coordinates, object coordinates are not locally controllable throughout the task, since their controllability changes with interaction modes, for instance when the system transitions from free-space motion to grasping or object transport. This state-dependent controllability fundamentally alters the local geometry of the goal-reaching problem. In particular, we show that full-state Eikonal regularization becomes inconsistent with the system dynamics when goal-relevant coordinates are locally uncontrollable. Motivated by this analysis, we develop two complementary formulations that explicitly account for local controllability and extend Pi-GCRL to goal-conditioned manipulation.

In summary, our contributions are:

- **Analysis of Eikonal-based Pi-GCRL under state-dependent controllability.** We characterize why full-state Eikonal regularization becomes inconsistent when goal-relevant coordinates are locally uncontrollable. Drawing on hybrid systems and optimal control, we show that the Eikonal regularizer can penalize value variation along directions that are not locally feasible, providing a principled explanation for its degradation in goal-reaching manipulation tasks.
- **Physics-informed value learning under local controllability.** Guided by this analysis, we derive two complementary design principles: (i) regularize value gradients only along locally feasible transitions, and (ii) apply physics-informed constraints within controllable representations. We realize these principles using a transition-based Hamilton-Jacobi-Bellman (HJB) residual and a hierarchical decomposition that separates high-level object-centric reasoning from controllable low-level goal reaching.
- **Empirical validation in goal-conditioned manipulation.** We evaluate both formulations across offline goal-conditioned manipulation benchmarks, where they substantially improve over naive full-state Eikonal regularization. We further validate our approaches on a physical robot, demonstrating that the proposed algorithms extend beyond simulation.

2 Related work

GCRL GCRL conditions policies on desired goals, enabling agents to reuse goal-reaching experience across tasks (Kaelbling, 1993; Schaul et al., 2015). This perspective has motivated a broad family of methods for learning from sparse supervision, including hindsight relabeling (Andrychowicz et al., 2017), contrastive objectives (Eysenbach et al., 2022), state-occupancy matching (Ma et al., 2022a), and hierarchical approaches that decompose long-horizon problems into shorter goal-reaching subtasks (Giammarino & Qureshi, 2026; Park et al., 2024; Haramati et al., 2026; Park et al., 2026; Ahn et al., 2026). In the offline setting, where policies must be learned entirely from fixed datasets, these challenges have further motivated specialized algorithms for offline goal reaching (Park et al., 2024; Chebotar et al., 2021; Yang et al., 2022; 2023; Mezghani et al., 2023; Sikchi et al., 2023). A complementary line of work interprets optimal goal-conditioned values through the geometry of shortest feasible paths, inducing an asymmetric distance-like structure over the state space (Sontag, 1995; Liu et al., 2023). Quasimetric RL (QRL) (Wang et al., 2023) formalizes this perspective by constraining goal-conditioned value functions through quasimetric parameterizations that reflect the dynamics-induced geometry of goal reaching (Liu et al., 2023; Pitis et al.; Durugkar et al., 2021). More recently, Myers et al. (2026) further expanded this quasimetric perspective by introducing additional structural conditions, whereas Kang et al. (2026) relaxed quasimetric constraints proposing a different asymmetric parameterization for the goal-conditioned value function.

Physics-informed RL Recent work has introduced physics-informed inductive biases into value learning by regularizing value functions with PDE-based constraints inspired by continuous-time optimal control. Several studies have proposed HJB-style constraints for online or offline RL (Settai et al., 2026; Lien et al., 2024), demonstrating the potential of PDE-based value regularization. However, these approaches do not focus on GCRL or address the specific challenges posed by robotic manipulation. Within GCRL, Giammarino et al. (2026) introduce Eikonal regularization for goal-conditioned value learning (Noack & Clark, 2017), using the Eikonal equation as a computationally convenient way to impose reachability structure on the value function. Building on this idea, Giammarino & Qureshi (2026) combine Eikonal constraints with quasimetric value parameterizations,

yielding a trajectory-free Eikonal-constrained formulation of QRL (Wang et al., 2023). While these methods achieve strong results in navigation and object-free goal-reaching, their direct application to robotic manipulation remains limited. Our work addresses this gap by formally characterizing when Eikonal regularization provides a consistent inductive bias for goal-conditioned value learning and why this consistency can break under state-dependent controllability. This perspective also distinguishes our work from recent HJB-inspired approaches to GCRL (Viswanath et al., 2026). In our work, rather than trying to replace the Eikonal regularization in general, we identify the conditions under which it becomes incompatible with the local control structure of manipulation and develop formulations that explicitly account for this structure.

GCRL for robotic manipulation GCRL has also emerged as a natural framework for robotic manipulation, particularly since language-conditioned policies can be viewed as goal-conditioned policies in which goals are specified through language or language-aligned representations (Stepputtis et al., 2020; Ma et al., 2022b; 2023). Recent work has further incorporated pretrained vision-language models into GCRL (Wang et al., 2025), generated task-specific policy parameters from goal descriptions (Zhou et al., 2026), leveraged automata for temporally extended manipulation (Manganaris et al., 2026a), and developed hierarchical, entity-centric methods for long-horizon robotic tasks (Haramati et al., 2026). In contrast, our work focuses on PDE-constrained, physics-informed goal-conditioned value learning and develops formulations that more explicitly account for the entity-centric structure of manipulation and its state-dependent controllability.

3 PRELIMINARIES

GCRL formulation We consider a finite-horizon discounted goal-conditioned Markov Decision Process (MDP) defined by $(\mathcal{S}, \mathcal{G}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{P}_g, \rho_0, \gamma)$, where $\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_N$ denotes the entity-centric state space, with each $\mathcal{S}_i$ corresponding to the state space of entity i. $\mathcal{G}$ denotes the goal space, $\mathcal{A}$ the action space, and $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ the transition map, with $\mathcal{P}(\mathcal{S})$ denoting the set of probability distributions over $\mathcal{S}$. The goal-conditioned reward function $\mathcal{R} : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ is defined through goal-satisfaction sets $\mathcal{S}_g(g) \subseteq \mathcal{S}$ as $\mathcal{R}(s, g) = 0$ if $s \in \mathcal{S}_g(g)$ and $\mathcal{R}(s, g) = -1$ otherwise. At the beginning of each episode, a goal $g \sim \mathcal{P}_g$ is sampled from the goal distribution $\mathcal{P}_g \in \mathcal{P}(\mathcal{G})$, and an initial state $s_0 \sim \rho_0$ from the initial state distribution $\rho_0 \in \mathcal{P}(\mathcal{S})$. The agent seeks to maximize $J(\pi) = \mathbb{E}_{g\sim\mathcal{P}_g, \tau_\pi(g)} \left[\sum_{t=0}^{T} \gamma^t \mathcal{R}(s_t, g)\right]$, where $\gamma \in (0, 1]$ is the discount factor and $\tau_\pi(g)$ denotes a distribution of trajectories $(g, s_0, a_0, s_1, a_1, \ldots, s_T)$ generated by the policy $\pi : \mathcal{S} \times \mathcal{G} \to \mathcal{P}(\mathcal{A})$ with goal g. The value function induced by π is defined as $V^\pi(s, g) = \mathbb{E}_{\tau_\pi(g)} \left[\sum_{t=0}^{T} \gamma^t \mathcal{R}(s_t, g) \mid S_0 = s,\, G = g\right]$. Finally, we denote parameterized functions by $\pi_{\boldsymbol{\theta}}$, with parameters $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$.

Continuous-time optimal control In continuous time, we consider a dynamical system $\dot{s}(t) = f(s(t), a(t))$, where $s(t)$ denotes the state, $\dot{s}(t)$ its time derivative, and $a(t) \in \mathcal{A}$ the control action. Given an initial state $s(0) = s$ and a goal g, the undiscounted goal-reaching optimal control problem seeks a control trajectory $a(\cdot)$ that drives the system to g while minimizing the cumulative cost $J = \int_0^T c(s(t), a(t))\, dt$, where T denotes the goal-reaching time and $c(s(t), a(t))$ is the instantaneous cost. The optimal cost-to-go is defined as $d^*(s, g) = \inf_{a(\cdot)} \int_0^T c(s(t), a(t))\, dt$, with $s(0) = s$ and $s(T) = g$. By the Bellman principle of optimality, this cost satisfies $d^*(s, g) = \inf_{a\in\mathcal{A}} [c(s, a)\Delta t + d^*(s + f(s, a)\Delta t, g)] + o(\Delta t)$ for a short interval Δt. Under standard regularity conditions, expanding $d^*(s + f(s, a)\Delta t, g)$ to first order and taking the limit as $\Delta t \to 0$ yields

$$\inf_{a\in\mathcal{A}} [c(s, a) + \nabla_s d^*(s, g)^\intercal f(s, a)] = 0. \tag{1}$$

Given $d^*(g, g) = 0$, Eq. (1) represents the HJB PDE for the undiscounted goal-reaching problem.

Physics-informed GCRL Pi-GCRL connects goal-conditioned value learning with continuous-time optimal control by augmenting standard discrete-time value losses with regularizers derived from PDEs. For example, an Eikonal-regularized implicit value-learning objective can be written as

$$\mathbb{E}_{\tau_\pi(g)} \Big[\underbrace{L_2^\iota \left(\mathcal{R}(s, g) + \gamma V_{\bar{\boldsymbol{\theta}}_V}(s', g) - V_{\boldsymbol{\theta}_V}(s, g)\right)}_{\text{value-learning loss}} + \underbrace{\left(\|\nabla_s V_{\boldsymbol{\theta}_V}(s, g)\| S(s) - 1\right)^2}_{\text{Eikonal residual}} \Big], \tag{2}$$

where L_2^ι denotes the expectile regression loss used in implicit value learning (Kostrikov et al., 2022), and $\bar{\boldsymbol{\theta}}_V$ denotes target-network parameters. In Eq. (2), the first term is a discrete-time value-learning loss, while the second penalizes violations of the Eikonal constraint, with $S(s)$ denoting a prescribed speed profile. This constraint is derived by the Eikonal PDE, $\|\nabla_s V_{\boldsymbol{\theta}_V}(s,g)\|^2 = 1/S(s)^2$, which classically describes wave propagation in heterogeneous media (Noack & Clark, 2017). Under the convention $V^* = -d^*$, prior work has shown that the Eikonal PDE and the HJB in Eq. (1) are closely related (Giammarino et al., 2026; Giammarino & Qureshi, 2026). In particular, the two equations become equivalent when the dynamics are isotropic, i.e., when the feasible instantaneous velocity set $\mathcal{F}(s) = \{f(s,a) : a \in \mathcal{A}\}$ is a scaled Euclidean ball $\mathcal{F}(s) = S(s)\mathbb{B}$ for some $S(s) > 0$. As a result, Eikonal-based regularization has proven effective in navigation and object-free goal-reaching as a computationally efficient surrogate for more general HJB PDEs. In manipulation, however, these benefits do not transfer as readily. We next formalize this failure mode.

4 Goal-Conditioned Value Learning under Hybrid Contact Dynamics

This section analyzes the primary failure mode of Eikonal-based regularization for goal-conditioned value learning. We begin by modeling robotic manipulation as a hybrid dynamical system (Hedlund & Rantzer, 1999):

$$\dot{x}(t) = f_{q(t)}(x(t), z(t), u(t)), \;\; \dot{z}(t) = h_{q(t)}(x(t), z(t), u(t)), \;\; q(t) = \nu(x(t), z(t), q(t^-)), \quad (3)$$

where $x(t) \in \mathcal{X} \subset \mathbb{R}^n$ denotes the continuous state of the robot, $z(t) \in \mathcal{Z} \subset \mathbb{R}^m$ denotes the continuous state of the manipulated objects, $q(t) \in \mathcal{Q}$ denotes the discrete interaction mode, with $\mathcal{Q} \subset \mathbb{Z}$ a finite set, and $u(t) \in \mathcal{U}$ is the continuous control input. The map ν captures transitions between interaction modes (e.g., from free-space motion to holding an object), with $q(t^-)$ denoting the mode immediately before the transition. In manipulation, the object dynamics h_q depend strongly on the current interaction mode $q(t)$. For example, before grasping, an object may be locally uncontrollable, so that $h_q(x,z,u) = 0$, whereas during grasping or pushing, the same object coordinates may become controllable through their coupling with the agent dynamics. We consider tasks in which the goal involves controlling object states, such as pick-and-place. In this setting, we want to determine when the Eikonal-based term in Eq. (2) ceases to provide a sound approximation of the local goal-reaching geometry and, consequently, a sound inductive bias for value learning. The key difficulty is that object coordinates may remain relevant to the long-horizon cost-to-go even when they are locally uncontrollable in the current mode. The following theorem formalizes why this creates a structural mismatch between full-state Eikonal regularization and the local HJB in Eq. (1).

Theorem 4.1 (Eikonal mismatch under locally uncontrollable coordinates)**.** *Consider an interaction mode q of the hybrid system in Eq. (3), away from a mode-switching boundary, and suppose the state decomposes as $s = (x,z)$ with dynamics $\dot{x} = f_q(x,z,u)$ and $\dot{z} = 0$, where x denotes the locally controllable coordinates and z denotes the locally uncontrollable coordinates in mode q. Assume unit running cost and define the maximum feasible instantaneous speed as $\mathcal{F}_q^*(x,z) := \sup_{u\in\mathcal{U}} \|f_q(x,z,u)\|$, with $0 < \mathcal{F}_q^*(x,z) < \infty$. Let $d^*(x,z,g)$ be differentiable at (x,z). If the locally uncontrollable coordinates remain relevant to future goal reaching, i.e., the optimal remaining cost-to-go varies with z so that $\nabla_z d^*(x,z,g) \neq 0$, then the full-state Eikonal constraint cannot be simultaneously satisfied with the local HJB characterization.*

Proof. The HJB equation in this mode depends only on gradients along the controllable coordinates, $\inf_{u\in\mathcal{U}} \left\{1 + \nabla_x d^*(x,z,g)^\top f_q(x,z,u)\right\} = 0$. Given $\mathcal{F}_q^*(x,z) := \sup_{u\in\mathcal{U}} \|f_q(x,z,u)\|$, the HJB condition implies $1 = \sup_{u\in\mathcal{U}}\{-\nabla_x d^*(x,z,g)^\top f_q(x,z,u)\}$ and, by Cauchy-Schwarz, $1 \leq \mathcal{F}_q^*(x,z)\|\nabla_x d^*(x,z,g)\|$. However, the full-state Eikonal constraint requires $\mathcal{F}_q^{*2}(x,z)\left(\|\nabla_x d^*(x,z,g)\|^2 + \|\nabla_z d^*(x,z,g)\|^2\right) = 1$. If $\nabla_z d^*(x,z,g) \neq 0$, this implies $\mathcal{F}_q^*(x,z)\|\nabla_x d^*(x,z,g)\| < 1$, contradicting the local HJB condition. Hence, whenever d^* depends on locally uncontrollable coordinates, the full-state Eikonal constraint cannot be satisfied by the HJB solution, as it allocates part of the gradient magnitude to coordinates that cannot contribute to the local dynamics. Refer to Appendix A for the step-by-step proof. □

Theorem 4.1 shows that the failure of Eikonal-based constraints can arise from a mismatch with the local geometry induced by the system dynamics. In particular, object coordinates may remain relevant to future goal reaching even when they are locally uncontrollable in the current interaction

mode. The local HJB equation naturally respects this structure because only gradients along feasible instantaneous motions contribute to the local cost-to-go. In contrast, the full-state Eikonal equation constrains the entire gradient and can therefore penalize valid value variation along locally uncontrollable coordinates. Thus, physics-informed regularization should account not only for the structure of the value function, but also for the directions along which the system can locally evolve.

Notably, prior work identified Lipschitz continuity of the optimal goal-conditioned value function as a sufficient condition for Eikonal-based regularization to provide a useful inductive bias (Giammarino et al., 2026; Giammarino & Qureshi, 2026). Our analysis shows that this condition alone is not sufficient in manipulation and, more generally, in systems with state-dependent controllability where the regularization must also reflect the locally feasible dynamics of each interaction mode.

A natural follow-up question is whether the misalignment identified in Theorem 4.1 is simply a consequence of applying an isotropic Eikonal constraint to anisotropic dynamics. While related, the two cases are fundamentally different. Under full-dimensional anisotropy, feasible motion remains supported in every local state direction, although with direction-dependent speeds. In this case, the HJB equation in Eq. (1) captures this control-induced geometry, whereas the Eikonal equation in Eq. (2) provides an isotropic approximation. By contrast, under local uncontrollability, the feasible velocity set loses dimensionality, so the full-state Eikonal constraint regularizes directions that are not locally realizable. The following theorem formalizes this distinction.

Theorem 4.2 (Eikonal approximation under full-dimensional anisotropic dynamics)**.** *Consider a continuous-time goal-reaching problem with dynamics $\dot{s} = f(s, u)$, unit running cost, and differentiable optimal cost-to-go $d^*(s, g)$. Let $\mathcal{F}(s) = \{f(s, u) : u \in \mathcal{U}\}$ denote the feasible instantaneous velocity set. Assume that the convex hull of the feasible velocity set contains a Euclidean ball of radius $r(s) > 0$ and is contained in a Euclidean ball of radius $R(s) < \infty$. Then, the local HJB characterization implies $\frac{1}{R(s)} \leq \|\nabla_s d^*(s, g)\| \leq \frac{1}{r(s)}$. If $r(s) = R(s) = S(s)$, the feasible geometry is isotropic and the bounds reduce to the Eikonal equation $S(s)\|\nabla_s d^*(s, g)\| = 1$. Hence, when the feasible velocity set remains full-dimensional but anisotropic, the Eikonal equation provides an isotropic approximation of the HJB-induced local geometry.*

Proof. With unit running cost, the HJB equation is $\inf_{u\in\mathcal{U}}\{1 + \nabla_s d^*(s, g)^\top f(s, u)\} = 0$, or equivalently $\sigma_{\mathcal{F}(s)}(-\nabla_s d^*(s, g)) = 1$, where $\sigma_{\mathcal{F}(s)}(w) = \sup_{v\in\mathcal{F}(s)} w^\top v$ denotes the support function of $\mathcal{F}(s)$. Since a set and its convex hull have the same support function, and the support function is monotone under set inclusion, the inner and outer ball assumptions imply $r(s)\|\nabla_s d^*(s, g)\| \leq \sigma_{\mathcal{F}(s)}(-\nabla_s d^*(s, g)) \leq R(s)\|\nabla_s d^*(s, g)\|$. Using $\sigma_{\mathcal{F}(s)}(-\nabla_s d^*(s, g)) = 1$ from the HJB condition gives $r(s)\|\nabla_s d^*(s, g)\| \leq 1 \leq R(s)\|\nabla_s d^*(s, g)\|$, and therefore $\frac{1}{R(s)} \leq \|\nabla_s d^*(s, g)\| \leq \frac{1}{r(s)}$. When $r(s) = R(s) = S(s)$, the bounds coincide and we recover the Eikonal equation $S(s)\|\nabla_s d^*(s, g)\| = 1$. Refer to Appendix A for the step-by-step proof. □

To summarize, Theorem 4.1 and Theorem 4.2 lead to two main observations. First, full-state Eikonal regularization introduces a *structural misalignment* when the dynamics contain locally uncontrollable coordinates, such as object states in no-contact modes. In this case, the Eikonal-based objective constrains value gradients along directions that do not correspond to feasible instantaneous motion, as shown in Theorem 4.1. Second, Theorem 4.2 clarifies why Eikonal regularization can still provide a useful inductive bias in navigation and object-free goal-reaching tasks, even when the true dynamics are not perfectly isotropic. In these settings, anisotropy distorts the local geometry induced by the dynamics but does not eliminate feasible motion along goal-relevant directions, so the Eikonal constraint remains an imperfect but dynamically meaningful approximation.

5 Controllability-Aware and Hierarchical Value Learning

Guided by the analysis above, we develop two complementary formulations that align physics-informed regularization with the local controllability structure of the dynamics: (i) a controllability-aware residual evaluated along observed feasibility transitions, and (ii) a hierarchical decomposition that applies PDE-based regularization within a controllable low-level representation.

Controllability-aware physics-informed value learning Consider an observed transition (s, s') where the state may contain both locally controllable and locally uncontrollable coordinates. Rather than constraining the norm of the full value gradient as in Eq. (2), we evaluate the physics-informed constraint along the empirical state transition $s' - s$. For a fixed discretization interval, this transition

provides a local approximation of the integrated dynamics, i.e., $s' - s \approx f(s,a)\Delta t$, and reflects which state coordinates can change under the current interaction mode. Rather than enforcing the full HJB equation in Eq. (1), which requires optimizing over the action space $\mathcal{A}$, we use the observed transition to construct an *empirical directional HJB residual*. In practice, g is sampled as a future goal along the same trajectory containing (s, s'), so that the observed transition corresponds to a feasible direction along a trajectory that eventually reaches the selected goal. This yields the approximation

$$\mathcal{L}_{\text{CA-HJB}}(\boldsymbol{\theta}) = \mathbb{E}_{\tau_\pi(g)}\left[\left(\nabla_s d_{\boldsymbol{\theta}}(s,g)^\top (s' - s) + 1\right)^2\right], \tag{4}$$

where $d_{\boldsymbol{\theta}}(s,g) = -V_{\boldsymbol{\theta}}(s,g)$ and the unit term is consistent with the unit transition cost used in GCRL. Unlike the full-state Eikonal residual in Eq. (2), the term in (4) evaluates value variation only along observed feasible transitions. Consequently, locally uncontrollable coordinates do not contribute when their displacement is zero, while differences in the magnitude of feasible motion are preserved. This allows the regularizer to reflect the state-dependent local controllability identified in Section 4.

Notably, transition-based HJB regularization of this form has also appeared in prior work (Giammarino et al., 2026); accordingly, our contribution does not lie in the algebraic form of Eq. (4) itself. Rather, the analysis in Section 4 provides a controllability-based justification for its use in manipulation and explains why evaluating the constraint along feasible transitions avoids the mismatch introduced by full-state Eikonal regularization.

Physics-informed Hierarchical Flow We next propose a hierarchical formulation motivated by the same controllability principle. Instead of imposing PDE-based constraints on the full manipulation state, we separate high-level object-centric reasoning from low-level goal reaching, and apply the PDE regularizer only in a low-level representation aligned with the locally controllable dynamics.

Specifically, we define a high-level policy $\pi_{\text{hi}} : \mathcal{S} \times \mathcal{G} \to \mathcal{X}$, which maps the full state $s \in \mathcal{S}$ and goal $g \in \mathcal{G}$ to a subgoal $x_g \in \mathcal{X}$. Here, $\mathcal{S}$ denotes the full manipulation state space, while $\mathcal{X}$ denotes the controllable state representation associated with the continuous state variable x in Eq. (3). The low-level policy $\pi_{\text{lo}} : \mathcal{X} \times \mathcal{X} \to \mathcal{A}$ then operates entirely within $\mathcal{X}$, mapping the current low-level state $x_s \in \mathcal{X}$ and subgoal $x_g \in \mathcal{X}$ to a primitive action. As a result, we obtain the following decomposition:

$$(s,g) \xrightarrow{\pi_{\text{hi}}} x_g, \qquad (x_s, x_g) \xrightarrow{\pi_{\text{lo}}} a.$$

Built on this decomposition, we introduce *Physics-informed Hierarchical Flow* (Pi-H-Flow), which combines a flow-based high-level subgoal generator (Lyu et al., 2026) with low-level physics-informed value learning. Since the low-level learner acts over the controllable representation $\mathcal{X}$, Pi-GCRL objectives are better aligned with the local directions that the policy can influence. The low-level learner can therefore use any physics-informed objectives, including Eik-QRL (Giammarino & Qureshi, 2026), Eik-GCIVL as in Eq. (2), or the transition-based HJB residual in Eq. (4).

We parameterize π_{hi} with a flow model because of its capacity to generate subgoals directly in the controllable representation, making them interpretable and avoiding an additional decoding step (Park et al., 2026). To train the high-level flow, we sample subgoal targets from the replay buffer as future controllable states along the same trajectory. Specifically, for a sampled state s, we let x_1 be the controllable representation of a state reached after a fixed number of environment steps. Given x_1, we sample noise $x_0 \sim \mathcal{N}(0, I)$ and interpolation time $t \sim \mathcal{U}[0, 1)$, define $x_t = (1-t)x_0 + tx_1$ with target velocity $y = x_1 - x_0$, and train the flow field $v_{\boldsymbol{\phi}}$ with

$$\mathcal{L}_{\text{hi-flow}}(\boldsymbol{\phi}) = \mathbb{E}\left[\|v_{\boldsymbol{\phi}}(s, g, x_t, t) - y\|^2\right].$$

We also train a high-level critic $Q^{\text{hi}}_{\boldsymbol{\psi}}(s, g, x_g)$ using temporal-difference updates over temporally extended subgoal transitions. At evaluation time, subgoals are generated by sampling an initial noise vector $x_0 \sim \mathcal{N}(0, I)$ and integrating the learned flow field $v_{\boldsymbol{\phi}}$ from $t = 0$ to $t = 1$ to obtain a candidate subgoal x_g. Following Park et al. (2026), we sample multiple candidate subgoals from the learned flow, score them with the high-level critic, and pass the highest-valued subgoal to the low-level policy. We refer to the supplementary material for a summary of the full pipeline (Appendix B).

6 Experiments

We organize our experiments around four main questions: (1) can we empirically observe the effect of full-state Eikonal regularization when some state coordinates are locally uncontrollable? (2) does this

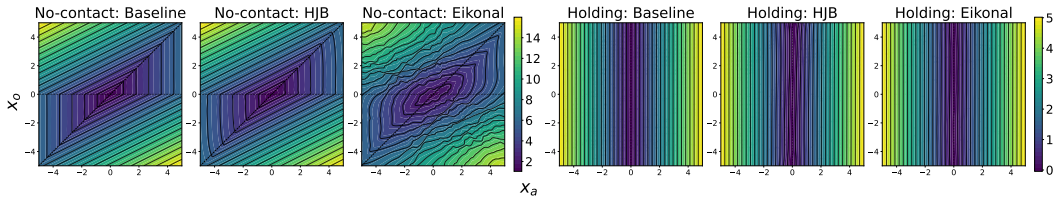


Figure 1: Results for regressing d^* in (5). In the no-contact mode, full-state Eikonal regularization distorts the learned function $d_{\boldsymbol{\theta}}$ due to locally uncontrollable coordinates, as highlighted in Theorem 4.1. The HJB regularizer addresses this issue by constraining only the controllable direction. In the holding mode, where all coordinates are locally controllable, all losses recover similar solutions.

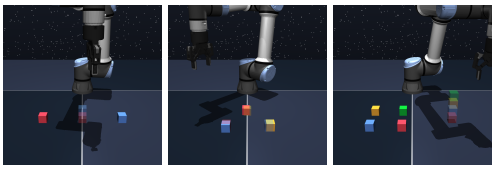
mismatch translate into degraded performance in manipulation? (3) does our controllability-aware formulations address this issue and improve goal-reaching performance? and (4) can the resulting framework be applied to real-world scenarios?

A 1-D illustrative example We introduce a simple example to visualize the failure mode of full-state Eikonal regularization, as discussed in Theorem 4.1. Consider an MDP with $s = (x_a, x_o, h)$, where x_a is the agent position, x_o is the object position, and $h \in \{0, 1\}$ indicates whether the agent is holding the object. The agent can move left or right, and can switch to the holding mode only when it is co-located with the object. Once $h = 1$, the object moves together with the agent. For goals where the object must be held at the target location, i.e., $h_g = 1$ and $x_a^g = x_o^g$, the undiscounted optimal cost-to-go under unit step cost is

$$d^*(x_a, x_o, h; x_o^g, h_g = 1) = \begin{cases} |x_a - x_o^g|, & h = 1, \\ |x_a - x_o| + 1 + |x_o - x_o^g|, & h = 0. \end{cases} \tag{5}$$

Given d^*, we fix $x_o^g = 0$ and train $d_{\boldsymbol{\theta}}$ with three different losses: (i) an unregularized baseline $\mathcal{L}(d_{\boldsymbol{\theta}}, d^*) = \mathbb{E}[(d_{\boldsymbol{\theta}}(s, g) - d^*(s, g))^2]$; (ii) a transition-based HJB-regularized loss $\mathcal{L}_{\text{HJB}}(d_{\boldsymbol{\theta}}, d^*) = \mathcal{L}(d_{\boldsymbol{\theta}}, d^*) + \mathcal{L}_{\text{CA-HJB}}(\boldsymbol{\theta})$, with $\mathcal{L}_{\text{CA-HJB}}(\boldsymbol{\theta})$ defined in Eq. (4); and (iii) an Eikonal-regularized loss $\mathcal{L}_{\text{Eik}}(d_{\boldsymbol{\theta}}, d^*) = \mathcal{L}(d_{\boldsymbol{\theta}}, d^*) + \mathcal{L}_{\text{Eik}}(\boldsymbol{\theta})$, where $\mathcal{L}_{\text{Eik}}(\boldsymbol{\theta})$ is the Eikonal residual in Eq. (2). The results are summarized in Fig. 1, where we plot the learned $d_{\boldsymbol{\theta}}$ for both the no-contact mode ($h = 0$) and the holding mode ($h = 1$). In Fig. 1, for the no-contact mode both the baseline and the transition-based HJB-regularized solution recover the piecewise-linear structure of d^*. In contrast, the Eikonal-regularized solution clearly distorts the value landscape by smoothing and reshaping its dependence on x_o. In the holding mode, where all state coordinates are locally controllable, the three losses recover similar solutions. These results empirically support the analysis in Section 4. We refer to Appendix C for more on this example.

Simulated experiments In the following, we evaluate the impact of physics-informed regularization in goal-conditioned manipulation tasks. Our goal is to assess whether the failure mode identified in Section 4 translates into degraded performance, and whether the proposed formulations can address this limitation. We conduct experiments on the manipulation suite of OGBench (Park et al., 2025), shown in Fig. 2. These environments require a 6-DoF UR5e arm to pick and rearrange multiple cubes into a specified goal configuration. The full MDP state contains robot proprioception, end-effector pose, gripper state, and object poses. Importantly, these tasks are goal-conditioned, i.e., algorithms are not evaluated only on a single task but on their ability to generalize across different goal configurations.



(a) `Double` (b) `Triple` (c) `Quadruple`

Figure 2: OGBench (Park et al., 2025) manipulation environments used in our experiments. The state includes robot proprioception, end-effector pose, gripper state, and object poses, and the goal is to arrange the cubes into a specified target configuration.

We test our controllability-aware formulation by replacing the full-state Eikonal residual in Eq. (2) with the transition-based HJB residual in Eq. (4). We refer to this method as HJB-GCIVL. We also evaluate our hierarchical formulation, Pi-H-Flow, using HJB-GCIVL as the low-level learner; we

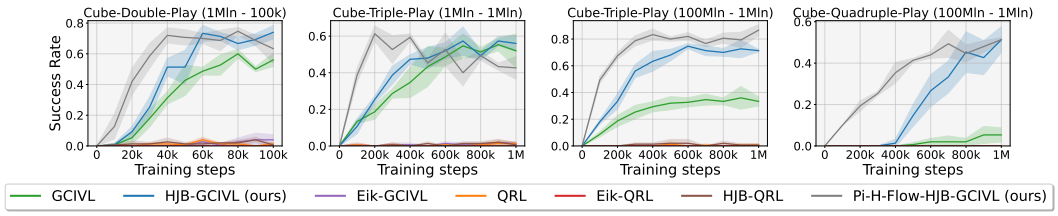


Figure 3: Results on the OGBench cube environments. The notation in each subplot title, e.g., `100Mln-1Mln`, indicates the dataset size and number of training steps, respectively. Evaluation is performed over 5 goal configurations with 10 episodes each and repeated over 3 seeds. We report the mean and standard deviation across seeds.

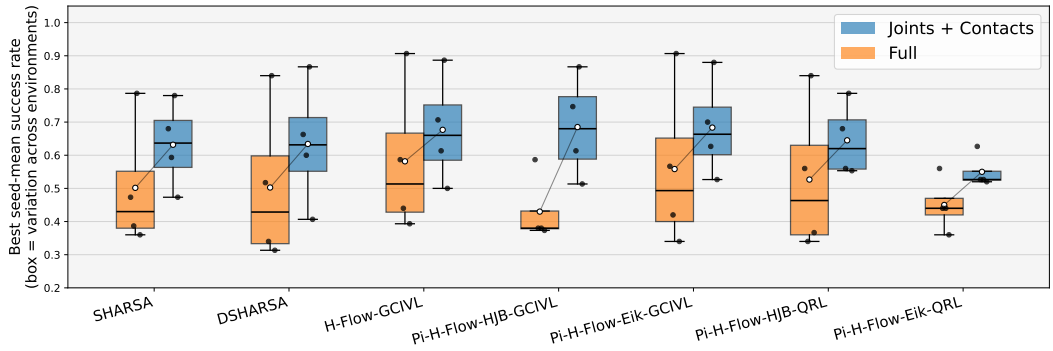


Figure 4: State representation ablation for hierarchical algorithms. For each algorithm, representation, and environment, we first average the evaluation curves across seeds and then take the best checkpoint of the resulting mean curve. Evaluation is performed as in Fig. 3. Each box summarizes these best seed-mean success rates across environments. Black dots denote individual environments, while white dots denote the mean across environments. We report the full training curves in Appendix C.

refer to this variant as Pi-H-Flow-HJB-GCIVL. In Pi-H-Flow-HJB-GCIVL, the low-level learner operates on a controllable representation containing only manipulator-related state variables, excluding object-pose coordinates. We compare against GCIVL (Park et al., 2025), Eik-GCIVL (Giammarino et al., 2026), QRL (Wang et al., 2023), and Pi-QRL variants, including Eik-QRL and HJB-QRL (Giammarino & Qureshi, 2026). The results are summarized in Fig. 3. QRL-based methods do not converge in this setting, highlighting the rigidity of quasimetric parameterizations in high-dimensional state spaces, as also observed in prior work (Giammarino & Qureshi, 2026). Consistent with our analysis, Eik-GCIVL fails to converge as well, reflecting the misalignment between full-state Eikonal regularization and state-dependent controllability. In contrast, GCIVL learns meaningful policies, but both our proposed variants achieve substantially better performance, especially in the more challenging `Triple` and `Quadruple` environments. Finally, Pi-H-Flow-HJB-GCIVL outperforms and converges faster than HJB-GCIVL, supporting the benefit of hierarchy for long-horizon goal-conditioned manipulation.

State representation ablation. An important question is whether the proposed controllability-aware state decomposition is actually useful for hierarchical algorithms. Pi-H-Flow is designed to use different representations at the high and low levels. However, one could argue that, with appropriately defined subgoals, using the full manipulation state or a controllable low-level representation should lead to similar performance, since the subgoals may already restrict the low-level problem to controllable variables or modes. We test this hypothesis using the same environments, tasks, and evaluation protocol as in Fig. 3.

Results are summarized in Fig. 4, where we compare hierarchical algorithms whose low-level MDP is defined either on the full manipulation state or on the proposed controllable representation, denoted as "Joints+Contacts", which excludes object-pose coordinates. We also include SHARSA and DSHARSA (Park et al., 2026) as hierarchical baselines. Across algorithms, using the controllable low-level representation consistently improves performance over using the full state, with gains between $10\% - 20\%$ in best mean success rate across seeds for most methods. This improvement is not limited

to physics-informed objectives, suggesting that goal reaching in controllable robot-centric coordinates is an easier low-level problem than goal reaching in the full object-centric manipulation state. For physics-informed methods, the same decomposition further aligns the PDE-based regularizer with the locally feasible directions of the low-level dynamics. Among the evaluated variants, Pi-H-Flow-HJB-GCIVL achieves the strongest overall performance, supporting the combination of controllability-aware regularization with the proposed hierarchical MDP decomposition.

Real-world experiments. We further evaluate the proposed framework on the real-world pick-and-place task shown in Fig. 5. Due to space constraints, we summarize the main findings here and provide additional details in Appendix D. Given a randomized initial object pose, the robot must place the object at the center of the table. State and action spaces are defined analogously to the OGBench manipulation environments (Park et al., 2025), with object-pose estimates obtained from an RGB-D camera. For these experiments, we construct an offline dataset using a scripted data-collection policy based on an MPC controller (Manganaris et al., 2026b). We collect approximately 60k interaction steps and train four policies for 300k gradient steps: GCIVL, HJB-GCIVL, H-Flow-GCIVL, and Pi-H-Flow-HJB-GCIVL. We then evaluate the trained policies on the physical system over 20 episodes under three object-position thresholds, corresponding to 3cm, 5cm, and 10cm Euclidean distance from the center of the table.

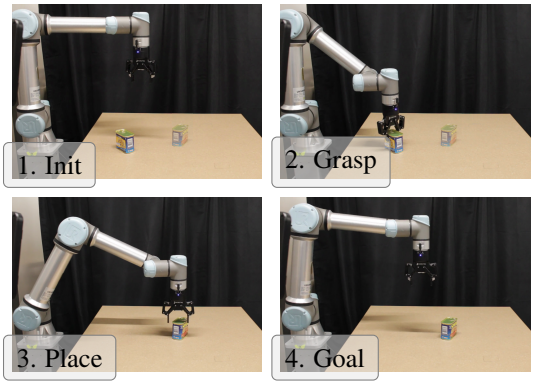


Figure 5: Real-world experimental setup. The task consists of moving an object to the center of the table from randomized initial poses.

GCIVL achieves success rates of 25%, 40%, and 45% at the three thresholds, respectively, while HJB-GCIVL improves to 40%, 50%, and 65%. H-Flow-GCIVL obtains 25%, 40%, and 40%, whereas Pi-H-Flow-HJB-GCIVL achieves the strongest performance, with success rates of 50%, 60%, and 75%. These results show that, even with a limited dataset of only 60k real-world samples, all methods learn meaningful behavior from offline data, while the physics-informed variants generally improve over their non-physics-informed counterparts. These experiments showcase that the proposed controllability-aware formulations remain effective beyond simulation and are compatible with practical robot-learning pipelines.

7 Conclusion

In this paper, we extended and analyzed Pi-GCRL for robotic manipulation. We showed that full-state Eikonal regularization can be misaligned with state-dependent controllability, distorting value functions and degrading performance. This analysis motivated two complementary solutions: a transition-based HJB regularizer that constrains value gradients along dynamically feasible directions, and a hierarchical decomposition that applies physics-informed regularization only in a low-level controllable representation. Across illustrative, simulated, and real-world experiments, these formulations improve over full-state regularization, especially in challenging multi-object tasks.

Limitations and future work. While these results are encouraging, important challenges remain. First, all evaluated algorithms are still data hungry, as is common in GCRL. Although physics-informed objectives improve performance, stronger physics priors are needed to improve sample efficiency, especially during real-world training. Second, our transition-based HJB regularizer shapes value learning according to feasible directions of motion, but it does not by itself guarantee that the resulting policy avoids unsafe contacts, respects actuator or force limits, or adapts its impedance to object properties. Future work will incorporate more explicit physics models into the HJB formulation, rather than relying only on data-driven approximations of feasible transitions. We also plan to extend Pi-GCRL toward safety-aware and compliant manipulation, for example by integrating reachability, constraint satisfaction, contact-force-aware objectives, or impedance-aware value regularization. More broadly, we are interested in using Pi-GCRL as a structured approach for adapting pretrained vision-language-action models. This direction could connect semantic goals to low-level behaviors that encode properties such as fragility, material constraints, or preferred modes of interaction.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of mathematical proofs. We did not use generative AI tools for any other tasks requiring disclosure.

Additionally, we used generative AI tools to create or modify figures or images and to edit portions of the manuscript to improve readability. All AI-assisted mathematical content was independently reviewed and verified by the authors for correctness, and all AI-assisted figures, images, and textual edits were reviewed by the authors before inclusion in the manuscript. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is primarily methodological and studies physics-informed value learning for GCRL under state-dependent controllability. We do not identify direct application-specific risks beyond those generally associated with deploying learning-based robotic manipulation systems. Since robotic manipulation may be used in manufacturing, logistics, and automation, practical deployment would require appropriate safety validation and safeguards for operation near humans. The experiments in this paper are conducted in controlled simulated and laboratory settings, and the proposed methods are not intended to provide formal safety guarantees by themselves.

REPRODUCIBILITY STATEMENT

To promote transparency and reproducibility, we will release the code used for our experiments in the camera-ready version. The appendix further reports implementation details, hyperparameters, additional training curves, and details of the real-world experimental setup and evaluation protocol.

REFERENCES

Hongjoon Ahn, Heewoong Choi, Jisu Han, and Taesup Moon. Option-aware temporally abstracted value for offline goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 38:99833–99861, 2026.

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in Neural Information Processing Systems*, 30, 2017.

Yevgen Chebotar, Karol Hausman, Yao Lu, Ted Xiao, Dmitry Kalashnikov, Jake Varley, Alex Irpan, Benjamin Eysenbach, Ryan Julian, Chelsea Finn, et al. Actionable models: Unsupervised offline reinforcement learning of robotic skills. *arXiv preprint arXiv:2104.07749*, 2021.

Ishan Durugkar, Mauricio Tec, Scott Niekum, and Peter Stone. Adversarial intrinsic motivation for reinforcement learning. *Advances in Neural Information Processing Systems*, 34:8622–8636, 2021.

Benjamin Eysenbach, Tianjun Zhang, Sergey Levine, and Russ R Salakhutdinov. Contrastive learning as goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 35:35603–35620, 2022.

Vittorio Giammarino and Ahmed H Qureshi. Goal reaching with eikonal-constrained hierarchical quasimetric reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5WhsCB0Vty.

Vittorio Giammarino, Matthew F Dunne, Kylie N Moore, Michael E Hasselmo, Chantal E Stern, and Ioannis Ch Paschalidis. Combining imitation and deep reinforcement learning to human-level performance on a virtual foraging task. *Adaptive Behavior*, 32(3):251–263, 2024.

Vittorio Giammarino, Ruiqi Ni, and Ahmed H Qureshi. Physics-informed value learner for offline goal-conditioned reinforcement learning. *Advances in Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=LRYgQuz7kY.

Dan Haramati, Carl Qi, Tal Daniel, Amy Zhang, Aviv Tamar, and George Konidaris. Hierarchical entity-centric reinforcement learning with factored subgoal diffusion. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TimC6hxVHj.

Sven Hedlund and Anders Rantzer. Optimal control of hybrid systems. In *Proceedings of the 38th IEEE Conference on Decision and Control (Cat. No. 99CH36304)*, volume 4, pp. 3972–3977. IEEE, 1999.

Sinan Ibrahim, Mostafa Mostafa, Ali Jnadi, Hadi Salloum, and Pavel Osinenko. Comprehensive overview of reward engineering and shaping in advancing reinforcement learning applications. *IEEE Access*, 12:175473–175500, 2024.

Leslie Pack Kaelbling. Learning to achieve goals. In *International Joint Conference on Artificial Intelligence*, volume 2, pp. 1094–8. Citeseer, 1993.

Hyungkyu Kang, Byeongchan Kim, and Min hwan Oh. Latent representation alignment for offline goal-conditioned reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=oddGER4AzK.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Yun-Hsuan Lien, Ping-Chun Hsieh, Tzu-Mao Li, and Yu-Shuen Wang. Enhancing value function estimation through first-order state-action dynamics in offline reinforcement learning. In *International Conference on Machine Learning*, 2024.

Bo Liu, Yihao Feng, Qiang Liu, and Peter Stone. Metric residual network for sample efficient goal-conditioned reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8799–8806, 2023.

Lei Lyu, Yunfei Li, Yu Luo, Fuchun Sun, Tao Kong, Jiafeng Xu, and Xiao Ma. Flow-based policy for online reinforcement learning. *Advances in Neural Information Processing Systems*, 38: 93967–93990, 2026.

Jason Yecheng Ma, Jason Yan, Dinesh Jayaraman, and Osbert Bastani. Offline goal-conditioned reinforcement learning via f-advantage regression. *Advances in Neural Information Processing Systems*, 35:310–323, 2022a.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022b.

Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. In *International Conference on Machine Learning*, pp. 23301–23320. PMLR, 2023.

Anastasios Manganaris, Vittorio Giammarino, and Ahmed H Qureshi. Automaton constrained q-learning. *Advances in Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=DLt2Ep1S3q.

Anastasios Manganaris, Jeremy Lu, Ahmed H Qureshi, and Suresh Jagannathan. Graph-of-constraints model predictive control for reactive multi-agent task and motion planning. *arXiv preprint arXiv:2603.18400*, 2026b.

Lina Mezghani, Sainbayar Sukhbaatar, Piotr Bojanowski, Alessandro Lazaric, and Karteek Alahari. Learning goal-conditioned policies offline with self-supervised reward shaping. In *Conference on Robot Learning*, pp. 1401–1410. PMLR, 2023.

Vivek Myers, Bill Zheng, Benjamin Eysenbach, and Sergey Levine. Offline goal-conditioned reinforcement learning with quasimetric representations. *Advances in Neural Information Processing Systems*, 38:19654–19679, 2026.

Marcus M Noack and Stuart Clark. Acoustic wave and eikonal equations in a transformed metric space for various types of anisotropy. *Heliyon*, 3(3), 2017.

Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. Hiql: Offline goal-conditioned rl with latent states as actions. *Advances in Neural Information Processing Systems*, 36, 2024.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations*, volume 2025, pp. 94937–94982, 2025.

Seohong Park, Kevin Frans, Deepinder Mann, Benjamin Eysenbach, Aviral Kumar, and Sergey Levine. Horizon reduction makes rl scalable. *Advances in Neural Information Processing Systems*, 38:8350–8389, 2026.

Silviu Pitis, Harris Chan, Kiarash Jamali, and Jimmy Ba. An inductive bias for distances: Neural nets that respect the triangle inequality. In *International Conference on Learning Representations*.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International Conference on Machine Learning*, pp. 1312–1320. PMLR, 2015.

Haruki Settai, Naoya Takeishi, and Takehisa Yairi. A temporal difference method for stochastic continuous dynamics. *Advances in Neural Information Processing Systems*, 2026. URL `https://openreview.net/forum?id=UKFg5yeZeX`.

Harshit Sikchi, Rohan Chitnis, Ahmed Touati, Alborz Geramifard, Amy Zhang, and Scott Niekum. Smore: Score models for offline goal-conditioned reinforcement learning. *arXiv preprint arXiv:2311.02013*, 2023.

Eduardo Sontag. An abstract approach to dissipation. In *Proceedings of 1995 34th IEEE Conference on Decision and Control*, volume 3, pp. 2702–2703. IEEE, 1995.

Simon Stepputtis, Joseph Campbell, Mariano Phielipp, Stefan Lee, Chitta Baral, and Heni Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. *Advances in Neural Information Processing Systems*, 33:13139–13150, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Hrishikesh Viswanath, Juanwu Lu, S Talha Bukhari, Damon Conover, Ziran Wang, and Aniket Bera. Physics informed viscous value representations. *arXiv preprint arXiv:2602.23280*, 2026.

Huiyi Wang, Fahim Shahriar, Alireza Azimi, Gautham Vasan, Rupam Mahmood, and Colin Bellinger. Versatile and generalizable manipulation via goal-conditioned reinforcement learning with grounded object detection. *arXiv preprint arXiv:2507.10814*, 2025.

Tongzhou Wang, Antonio Torralba, Phillip Isola, and Amy Zhang. Optimal goal-reaching reinforcement learning via quasimetric learning. In *International Conference on Machine Learning*, pp. 36411–36430. PMLR, 2023.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879, 2024. doi: 10.1109/CVPR52733.2024.01692.

Rui Yang, Yiming Lu, Wenzhe Li, Hao Sun, Meng Fang, Yali Du, Xiu Li, Lei Han, and Chongjie Zhang. Rethinking goal-conditioned supervised learning and its connection to offline rl. *arXiv preprint arXiv:2202.04478*, 2022.

Rui Yang, Lin Yong, Xiaoteng Ma, Hao Hu, Chongjie Zhang, and Tong Zhang. What is essential for unseen goal generalization of offline goal-conditioned rl? In *International Conference on Machine Learning*, pp. 39543–39571. PMLR, 2023.

Pei Zhou, Wanting Yao, Qian Luo, Xunzhe Zhou, and Yanchao Yang. Hyper-goalnet: Goal-conditioned manipulation policy learning with hypernetworks. *Advances in Neural Information Processing Systems*, 2026. URL `https://openreview.net/forum?id=aWWRPyGMie`.

A STEP-BY-STEP DERIVATIONS

A.1 THEOREM 4.1 RESTATED

Theorem A.1 (Eikonal mismatch under locally uncontrollable coordinates)**.** *Consider an interaction mode q of the hybrid system in Eq. (3), away from a mode-switching boundary, and suppose the state decomposes as $s = (x, z)$ with dynamics*

$$\dot{x} = f_q(x, z, u), \qquad \dot{z} = 0,$$

where x denotes the locally controllable coordinates and z the locally uncontrollable coordinates in mode q. Assume unit running cost and define the maximum feasible instantaneous speed as

$$\mathcal{F}_q^*(x, z) := \sup_{u \in \mathcal{U}} \|f_q(x, z, u)\|,$$

with $0 < \mathcal{F}_q^*(x, z) < \infty$. Let $d^*(x, z, g)$ be differentiable at (x, z). If the locally uncontrollable coordinates remain relevant to future goal reaching, such that

$$\nabla_z d^*(x, z, g) \neq 0,$$

then the full-state Eikonal constraint cannot be simultaneously satisfied with the local HJB characterization.

Proof. The HJB equation in this mode depends only on gradients along the controllable coordinates,

$$\inf_{u \in \mathcal{U}} \left\{1 + \nabla_x d^*(x, z, g)^\top f_q(x, z, u)\right\} = 0.$$

Given

$$\mathcal{F}_q^*(x, z) := \sup_{u \in \mathcal{U}} \|f_q(x, z, u)\|.$$

From the local HJB condition, we obtain

$$\inf_{u \in \mathcal{U}} \nabla_x d^*(x, z, g)^\top f_q(x, z, u) = -1.$$

Equivalently,

$$1 = -\min_{u \in \mathcal{U}} \nabla_x d^*(x, z, g)^\top f_q(x, z, u) = \sup_{u \in \mathcal{U}} \left\{-\nabla_x d^*(x, z, g)^\top f_q(x, z, u)\right\}.$$

For any $u \in \mathcal{U}$, Cauchy-Schwarz gives

$$-\nabla_x d^*(x, z, g)^\top f_q(x, z, u) \leq \|\nabla_x d^*(x, z, g)\| \|f_q(x, z, u)\|.$$

Taking the supremum over u on both sides,

$$\sup_{u \in \mathcal{U}} \left\{-\nabla_x d^*(x, z, g)^\top f_q(x, z, u)\right\} \leq \|\nabla_x d^*(x, z, g)\| \sup_{u \in \mathcal{U}} \|f_q(x, z, u)\|.$$

Using the definition

$$\mathcal{F}_q^*(x, z) := \sup_{u \in \mathcal{U}} \|f_q(x, z, u)\|,$$

we therefore obtain

$$1 \leq \mathcal{F}_q^*(x, z) \|\nabla_x d^*(x, z, g)\|.$$

However, the full-state Eikonal constraint requires

$$\mathcal{F}_q^{*2}(x, z) \left(\|\nabla_x d^*(x, z, g)\|^2 + \|\nabla_z d^*(x, z, g)\|^2\right) = 1.$$

If $\nabla_z d^*(x, z, g) \neq 0$, this implies

$$\mathcal{F}_q^*(x, z) \|\nabla_x d^*(x, z, g)\| < 1,$$

contradicting the local HJB condition. Hence, whenever d^* depends on locally uncontrollable coordinates, the full-state Eikonal constraint cannot be satisfied by the HJB solution, as it allocates part of the gradient magnitude to coordinates that cannot contribute to the local dynamics. □

Refer also to the one-dimensional illustrative example in Section 6. Together with its extended version in Section C.2, they provide an intuitive and graphical interpretation of the mismatch characterized by Theorem 4.1.

A.2 THEOREM 4.2 RESTATED

Theorem A.2 (Eikonal approximation under full-dimensional anisotropic dynamics)**.** *Consider a continuous-time goal-reaching problem with dynamics $\dot{s} = f(s, u)$, unit running cost, and differentiable optimal cost-to-go $d^*(s, g)$. Let*

$$\mathcal{F}(s) = \{f(s, u) : u \in \mathcal{U}\}$$

denote the feasible instantaneous velocity set, and consider its convex hull with respect to the origin of the local velocity reference frame. Define

$$\mathbb{B}_{r(s)}(0) := \{v : \|v\| \le r(s)\}, \qquad \mathbb{B}_{R(s)}(0) := \{v : \|v\| \le R(s)\},$$

with $r(s) > 0$ and $R(s) < \infty$. Assume that

$$\mathbb{B}_{r(s)}(0) \subseteq \operatorname{conv} \mathcal{F}(s) \subseteq \mathbb{B}_{R(s)}(0).$$

Then, the local HJB characterization implies

$$\frac{1}{R(s)} \le \|\nabla_s d^*(s, g)\| \le \frac{1}{r(s)}.$$

If $r(s) = R(s) = S(s)$, the feasible geometry is isotropic and the bounds reduce to the Eikonal equation

$$S(s)\|\nabla_s d^*(s, g)\| = 1.$$

Hence, when the feasible velocity set remains full-dimensional but anisotropic, the Eikonal equation provides an isotropic approximation of the HJB-induced local geometry.

Proof. With unit running cost, the HJB equation is

$$\inf_{u \in \mathcal{U}} \left\{1 + \nabla_s d^*(s, g)^\top f(s, u)\right\} = 0.$$

Therefore,

$$\inf_{u \in \mathcal{U}} \nabla_s d^*(s, g)^\top f(s, u) = -1,$$

or equivalently,

$$1 = \sup_{u \in \mathcal{U}} \left\{-\nabla_s d^*(s, g)^\top f(s, u)\right\}.$$

Since

$$\mathcal{F}(s) = \{f(s, u) : u \in \mathcal{U}\},$$

this can be written as

$$1 = \sup_{v \in \mathcal{F}(s)} \left\{-\nabla_s d^*(s, g)^\top v\right\} = \sigma_{\mathcal{F}(s)} \left(-\nabla_s d^*(s, g)\right),$$

where $\sigma_{\mathcal{F}(s)}(w) := \sup_{v \in \mathcal{F}(s)} w^\top v$ is the support function of the feasible velocity set. A set and its convex hull have the same support function, so

$$\sigma_{\mathcal{F}(s)}(w) = \sigma_{\operatorname{conv} \mathcal{F}(s)}(w).$$

By assumption we have that

$$\mathbb{B}_{r(s)}(0) \subseteq \operatorname{conv} \mathcal{F}(s) \subseteq \mathbb{B}_{R(s)}(0),$$

and since the support function is monotone under set inclusion we obtain

$$\sigma_{\mathbb{B}_{r(s)}(0)}(w) \le \sigma_{\mathcal{F}(s)}(w) \le \sigma_{\mathbb{B}_{R(s)}(0)}(w).$$

For a Euclidean ball centered at the origin we can write

$$\sigma_{\mathbb{B}_\rho(0)}(w) = \rho\|w\|,$$

which replaced in the inequality above yields

$$r(s)\|w\| \le \sigma_{\mathcal{F}(s)}(w) \le R(s)\|w\|.$$

Setting

$$w = -\nabla_s d^*(s, g)$$

and using the HJB condition

$$\sigma_{\mathcal{F}(s)}\left(-\nabla_s d^*(s, g)\right) = 1,$$

we obtain

$$r(s)\|\nabla_s d^*(s, g)\| \le 1 \le R(s)\|\nabla_s d^*(s, g)\|.$$

Finally, since $r(s) > 0$ and $R(s) > 0$ we can write

$$\frac{1}{R(s)} \le \|\nabla_s d^*(s, g)\| \le \frac{1}{r(s)}.$$

When $r(s) = R(s) = S(s)$, the lower and upper bounds coincide:

$$S(s)\|\nabla_s d^*(s, g)\| = 1,$$

which is precisely the Eikonal equation. □

B Algorithms and Implementation details

B.1 Training and inference pipeline for Pi-H-Flow-HJB-GCIVL

We denote the full manipulation state by s, the low-level controllable representation by x, the task goal by g, and a high-level subgoal by x_g. Pi-H-Flow-HJB-GCIVL consists of a high-level value function $V^{\mathrm{hi}}_{\boldsymbol{\theta}}$, a double high-level critic $Q^{\mathrm{hi}}_{\boldsymbol{\psi}_1}, Q^{\mathrm{hi}}_{\boldsymbol{\psi}_2}$, a high-level flow actor $v_{\boldsymbol{\phi}}$, a double low-level value function $V^{\mathrm{lo}}_{\boldsymbol{\omega}_1}, V^{\mathrm{lo}}_{\boldsymbol{\omega}_2}$, and a low-level actor $\pi^{\mathrm{lo}}_{\boldsymbol{\eta}}$. Target-network parameters are denoted by adding a bar to the corresponding parameter vector, following the notation used in Eq. (2) in the main text. We first summarize the main loss functions used to train these components, before presenting the full training and inference pipelines. These pipelines are respectively summarized in Algorithm 1 and Algorithm 2. For additional implementation details, including the dataset processing pipeline, model architectures, loss implementations, and training scripts, we refer to our GitHub repository.

High-level critic and value losses. The high-level critic is trained with temporally extended temporal-difference (TD) targets (Sutton & Barto, 2018). For a subgoal transition of length k, we define

$$y^{\mathrm{hi}} = \mathcal{R}^{\mathrm{hi}} + \gamma^k m^{\mathrm{hi}} V^{\mathrm{hi}}_{\boldsymbol{\theta}}(s_{t+k}, g), \tag{6}$$

where m^{hi} is the high-level continuation mask. The high-level critic loss is

$$\mathcal{L}^{\mathrm{hi}}_Q(\boldsymbol{\psi}) = \mathbb{E}_{\mathcal{D}}\left[\ell\big(Q^{\mathrm{hi}}_{\boldsymbol{\psi}_1}(s_t, g, x_g), y^{\mathrm{hi}}\big) + \ell\big(Q^{\mathrm{hi}}_{\boldsymbol{\psi}_2}(s_t, g, x_g), y^{\mathrm{hi}}\big)\right], \tag{7}$$

where ℓ is the value loss used in implementation, e.g., binary cross-entropy for the reported experiments, and $\mathcal{D}$ is a replay buffer. The high-level value function is trained from the target high-level critic:

$$\mathcal{L}^{\mathrm{hi}}_V(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}\left[\ell\left(V^{\mathrm{hi}}_{\boldsymbol{\theta}}(s_t, g), \min_{i=1,2} Q^{\mathrm{hi}}_{\bar{\boldsymbol{\psi}}_i}(s_t, g, x_g)\right)\right]. \tag{8}$$

Low-level HJB-regularized value loss. The low-level value function operates on the controllable representation x. Let

$$y^{\mathrm{lo}}_i = \mathcal{R}^{\mathrm{lo}} + \gamma_{\mathrm{lo}} m^{\mathrm{lo}} V^{\mathrm{lo}}_{\bar{\boldsymbol{\omega}}_i}(x_{t+1}, x_g), \qquad i \in \{1, 2\}, \tag{9}$$

and

$$y^{\mathrm{lo}} = \mathcal{R}^{\mathrm{lo}} + \gamma_{\mathrm{lo}} m^{\mathrm{lo}} \min_{i=1,2} V^{\mathrm{lo}}_{\bar{\boldsymbol{\omega}}_i}(x_{t+1}, x_g). \tag{10}$$

The expectile advantage is computed using the target value estimate,

$$A^{\mathrm{lo}} = y^{\mathrm{lo}} - \frac{1}{2}\sum_{i=1}^{2} V^{\mathrm{lo}}_{\bar{\boldsymbol{\omega}}_i}(x_t, x_g). \tag{11}$$

The implicit value-learning term is

$$\mathcal{L}^{\mathrm{lo}}_{\mathrm{IVL}}(\boldsymbol{\omega}) = \sum_{i=1}^{2} \mathbb{E}_{\mathcal{D}} \left[L_2^{\iota} \left(A^{\mathrm{lo}}, y_i^{\mathrm{lo}} - V^{\mathrm{lo}}_{\boldsymbol{\omega}_i}(x_t, x_g) \right) \right]. \tag{12}$$

We add the transition-based HJB residual from the main text. Defining $d_{\boldsymbol{\omega}_i}(x, x_g) = -V^{\mathrm{lo}}_{\boldsymbol{\omega}_i}(x, x_g)$ yields

$$\mathcal{L}^{\mathrm{lo}}_{\mathrm{HJB}}(\boldsymbol{\omega}) = \sum_{i=1}^{2} \mathbb{E}_{\mathcal{D}} \left[\left(\nabla_x d_{\boldsymbol{\omega}_i}(x_t, x_g)^{\top} (x_{t+1} - x_t) + 1 \right)^2 \right]. \tag{13}$$

The complete low-level value loss is

$$\mathcal{L}^{\mathrm{lo}}_{V}(\boldsymbol{\omega}) = \mathcal{L}^{\mathrm{lo}}_{\mathrm{IVL}}(\boldsymbol{\omega}) + \lambda_{\mathrm{HJB}} \mathcal{L}^{\mathrm{lo}}_{\mathrm{HJB}}(\boldsymbol{\omega}), \tag{14}$$

with $\lambda_{\mathrm{HJB}} = 1$ in our implementation.

High-level flow actor loss. The high-level actor generates subgoals in the controllable representation. Given a future controllable state x_1 sampled from the same trajectory, we sample $x_0 \sim \mathcal{N}(0, I)$ and $t \sim \mathcal{U}[0, 1]$, define

$$x_t = (1 - t)x_0 + t x_1, \qquad y = x_1 - x_0, \tag{15}$$

and train the high-level flow field with

$$\mathcal{L}^{\mathrm{hi}}_{\mathrm{flow}}(\boldsymbol{\phi}) = \mathbb{E}_{\mathcal{D}, x_0, t} \left[\| v_{\boldsymbol{\phi}}(s, g, x_t, t) - y \|^2 \right]. \tag{16}$$

Low-level actor loss. The low-level actor is trained with advantage-weighted regression. We define

$$A^{\mathrm{actor}}_{\mathrm{lo}} = \frac{1}{2} \sum_{i=1}^{2} V^{\mathrm{lo}}_{\boldsymbol{\omega}_i}(x_{t+1}, x_g) - \frac{1}{2} \sum_{i=1}^{2} V^{\mathrm{lo}}_{\boldsymbol{\omega}_i}(x_t, x_g). \tag{17}$$

The low-level actor loss is

$$\mathcal{L}^{\mathrm{lo}}_{\pi}(\boldsymbol{\eta}) = -\mathbb{E}_{\mathcal{D}} \left[\min \left(\exp(\alpha A^{\mathrm{actor}}_{\mathrm{lo}}), 100 \right) \log \pi^{\mathrm{lo}}_{\boldsymbol{\eta}}(a_t | x_t, x_g) \right]. \tag{18}$$

Algorithm 1 Training Pi-H-Flow-HJB-GCIVL

Input: Offline dataset $\mathcal{D}$, high-level value $V^{\mathrm{hi}}_{\boldsymbol{\theta}}$, high-level critic $Q^{\mathrm{hi}}_{\boldsymbol{\psi}}$, high-level flow actor $v_{\boldsymbol{\phi}}$, low-level value $V^{\mathrm{lo}}_{\boldsymbol{\omega}}$, low-level actor $\pi^{\mathrm{lo}}_{\boldsymbol{\eta}}$, target networks, target update rate τ
while not converged **do**
  Sample a batch from $\mathcal{D}$
  Parse full states s and controllable states x
  Update $V^{\mathrm{hi}}_{\boldsymbol{\theta}}$ using Eq. (8)
  Update $Q^{\mathrm{hi}}_{\boldsymbol{\psi}}$ using Eq. (7)
  Update $V^{\mathrm{lo}}_{\boldsymbol{\omega}}$ using Eq. (14)
  Update $v_{\boldsymbol{\phi}}$ using Eq. (16)
  Update $\pi^{\mathrm{lo}}_{\boldsymbol{\eta}}$ using Eq. (18)
  Update target networks: $\bar{\boldsymbol{\psi}} \leftarrow (1 - \tau)\bar{\boldsymbol{\psi}} + \tau \boldsymbol{\psi}, \quad \bar{\boldsymbol{\omega}} \leftarrow (1 - \tau)\bar{\boldsymbol{\omega}} + \tau \boldsymbol{\omega}$
end while

Algorithm 2 Inference with Pi-H-Flow-HJB-GCIVL

Input: Current full state s, current controllable state x, goal g, high-level flow actor $v_{\boldsymbol{\phi}}$, high-level critic $Q^{\mathrm{hi}}_{\boldsymbol{\psi}}$, low-level actor $\pi^{\mathrm{lo}}_{\boldsymbol{\eta}}$, number of samples N, flow steps M
Sample N initial subgoals $x^i_0 \sim \mathcal{N}(0, I)$
for $m = 0, \ldots, M-1$ **do**
  Set $t_m = m/M$
  Update each candidate: $x^i_{m+1} = x^i_m + \frac{1}{M} v_{\boldsymbol{\phi}}(s, g, x^i_m, t_m)$
end for
Score candidates using the high-level critic: $q_i = \min_{j=1,2} Q^{\mathrm{hi}}_{\boldsymbol{\psi}_j}(s, g, x^i_M)$
Select the best subgoal: $x_g = x_M^{\arg\max_i q_i}$
Sample low-level action: $a \sim \pi^{\mathrm{lo}}_{\boldsymbol{\eta}}(\cdot|x, x_g)$
Clip a to the action bounds
return a

B.2 Hyperparameters

All methods use the hyperparameters reported in Table 1. For Pi-H-Flow-HJB-GCIVL, the high-level flow actor uses $M = 10$ integration steps at inference and samples $N = 32$ candidate subgoals, which are ranked using the high-level critic. The low-level discount is set automatically as $1 - 1/K$, where K is the number of subgoal steps.

Table 1: Hyperparameters used for Pi-H-Flow-HJB-GCIVL.

Hyperparameter	**Value**
Learning rate	3×10^{-4}
Batch size	1024
Actor hidden dimensions	$(1024, 1024, 1024, 1024)$
Value hidden dimensions	$(1024, 1024, 1024, 1024)$
Expectile ι	0.9
AWR temperature α	10.0
Layer normalization	True
Constant actor standard deviation	True
Discount factor γ	0.999
Low-level discount γ_{lo}	$1 - 1/K$
Target update rate τ	0.005
Flow integration steps	10
Candidate subgoals at inference	32
Low-level observation type	Joints+Contacts
Subgoal horizon K	25
Value goal: current state probability	0.2
Value goal: trajectory future probability	0.5
Value goal: random state probability	0.3
Actor goal: current state probability	0.0
Actor goal: trajectory future probability	1.0
Actor goal: random state probability	0.0

C Experiments: additional details and training curves

C.1 Computational resources

All experiments follow a standard offline RL pipeline, where datasets are collected before training and algorithms are trained only on fixed offline data. Training was performed on a workstation equipped with four NVIDIA GeForce RTX 3090 GPUs, each with 24GB of memory. We parallelized independent training runs across GPUs, with each run using a single GPU. In our implementation,

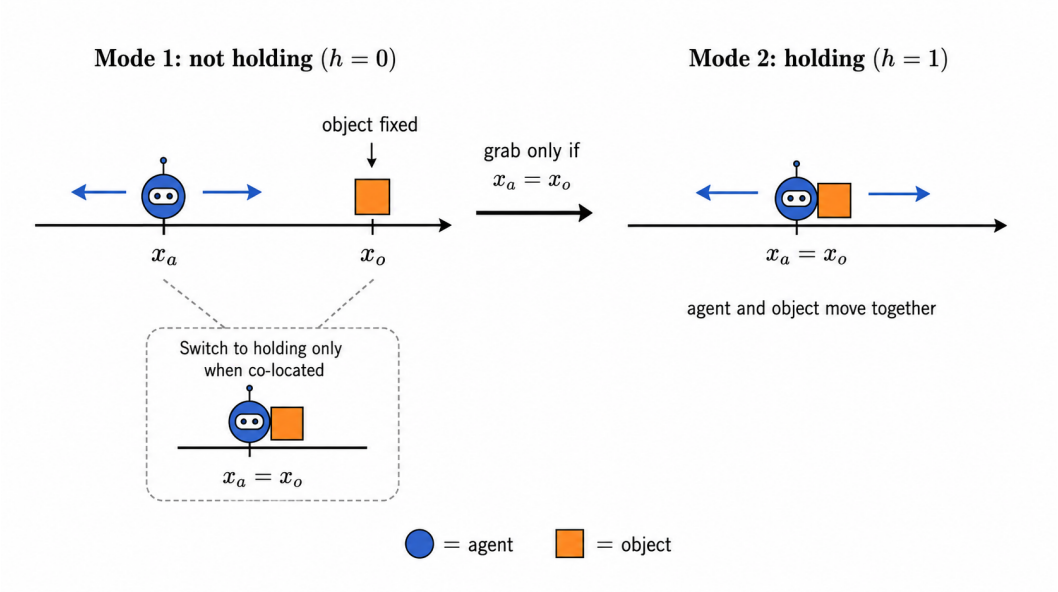


Figure 6: 1-D toy example of a contact-rich manipulation task with state-dependent dynamics.

the GPU memory usage of each algorithm remained below approximately 4GB, and training for one million gradient steps required approximately five hours per run.

C.2 Extended analysis of the hybrid toy example

In Section 6, we used a simple hybrid manipulation example to illustrate the failure mode predicted by Theorem 4.1. Here, we provide additional details on the construction of the example, the closed-form optimal cost-to-go, and the regression losses used to generate Fig. 1. The goal is to isolate the role of local uncontrollability from other sources of approximation error and to show explicitly how full-state Eikonal regularization behaves differently across contact modes.

The state is $s = (x_a, x_o, h)$, where x_a denotes the agent position, x_o denotes the object position, and $h \in \{0, 1\}$ indicates whether the agent is holding the object. When $h = 0$, the agent can move left or right while the object remains fixed. The agent can switch to the holding mode only when it is co-located with the object, i.e., when $x_a = x_o$. Once $h = 1$, the object is attached to the agent and moves together with it. We refer to Fig. 6 for a graphical illustration of this toy example.

For completeness, we now write the optimal cost-to-go in an explicit decomposed form. We consider goals in which the object must be held at the target location, so that $h_g = 1$ and $x_a^g = x_o^g$. Under unit step cost, if the agent is already holding the object, it only needs to transport it to the target. If the agent is not holding the object, it must first reach the object, switch contact mode, and then transport the object to the target:

$$d^*(x_a, x_o, h; x_o^g, h_g = 1) = \begin{cases} \underbrace{|x_a - x_o^g|}_{\text{transport object to target}}, & h = 1, \\ \underbrace{|x_a - x_o|}_{\text{reach object}} + \underbrace{1}_{\text{grasp/switch mode}} + \underbrace{|x_o - x_o^g|}_{\text{transport object to target}}, & h = 0. \end{cases}$$

Given this ground-truth cost-to-go, we fix $x_o^g = 0$ and train a neural approximation $d_{\boldsymbol{\theta}}$ with three losses. The first is an unregularized regression baseline,

$$\mathcal{L}_{\text{base}}(d_{\boldsymbol{\theta}}, d^*) = \mathbb{E}_s \left[\left(d_{\boldsymbol{\theta}}(s, g) - d^*(s, g) \right)^2 \right]. \tag{19}$$

The second augments this objective with the transition-based HJB residual in Eq. (4),

$$\mathcal{L}_{\text{HJB}}(d_{\boldsymbol{\theta}}, d^*) = \mathcal{L}_{\text{base}}(d_{\boldsymbol{\theta}}, d^*) + \mathcal{L}_{\text{CA-HJB}}(\boldsymbol{\theta}). \tag{20}$$

The third augments the same regression objective with the full-state Eikonal residual in 2,

$$\mathcal{L}_{\mathrm{Eik}}(d_{\boldsymbol{\theta}}, d^*) = \mathcal{L}_{\mathrm{base}}(d_{\boldsymbol{\theta}}, d^*) + \mathcal{L}_{\mathrm{Eik}}(\boldsymbol{\theta}). \tag{21}$$

Fig. 7 provides a more detailed view of the behavior summarized in the main text. In the no-contact mode, the ground-truth cost-to-go depends on both x_a and x_o, but only x_a is locally controllable. The baseline and HJB-regularized models recover the main piecewise-linear structure of d^*, with errors mostly localized near nondifferentiable boundaries such as $x_a = x_o$ and $x_o = x_o^g$. This is expected, since the value function has kinks at mode-switching and goal-alignment sets, where a smooth neural approximation cannot exactly match the true nonsmooth geometry.

The full-state Eikonal regularizer behaves differently. In the no-contact mode, it produces a visibly distorted value landscape and structured errors over broad regions of the state space, not only near the nondifferentiable boundaries. This supports Theorem 4.1, which states that the issue is not merely that nonsmoothness of the value function, but that the Eikonal penalty constrains gradients along the object coordinate even though the object is locally uncontrollable. In the holding mode, where the object moves with the agent, this mismatch disappears and all three losses produce similar fitted landscapes with comparatively small errors. Thus, this extended visualization confirms that the failure of full-state Eikonal regularization is state-dependent and specifically tied to local uncontrollability.

C.3 State representation ablation curves

Fig. 8 provides the full training curves corresponding to the ablation summarized in Fig. 4. The left and middle columns report the performance of each hierarchical method when the low-level MDP is defined on the full manipulation state and on the proposed controllable "Joints+Contacts" representation, respectively. The latter excludes object-state coordinates from the low-level problem. The right column averages performance across algorithms to isolate the effect of the low-level state representation.

Across environments, the controllable representation generally leads to faster learning and higher final success rates. This improvement is visible across all the environments, where using the full manipulation state often slows learning or plateaus at lower success rates. These trends support the claim that removing object-pose coordinates from the low-level problem simplifies goal-reaching and better aligns the low-level learner with locally controllable dynamics.

D Real-world Experiments

In this section, we provide additional details on the real-world experiments presented in the main text in Section 6, Fig. 5. We first describe the experimental setup and data-collection procedure, and then present qualitative results that complement the quantitative evaluation reported in the main paper.

D.1 Setup and data collection

We consider a tabletop pick-and-place task in which a UR5e robot must grasp an object and move it between randomly sampled configurations and a target goal region. Object pose estimation is performed from RGB-D observations captured by an Intel RealSense D415 camera using FoundationPose (Wen et al., 2024). The policy input consists of robot proprioception, consisting of the end-effector position, yaw, and gripper state, and the estimated object pose.

To collect offline interaction data, we design a scripted behavior policy based on a task and motion planner (Manganaris et al., 2026b). The planner uses the estimated state of the robot-object system to generate pick-and-place trajectories between random object configurations and the goal state. An example rollout of this data-collection policy is shown in Fig. 9. We collect approximately 188k transition samples and subsequently filter the dataset during post-processing, removing object resetting steps, failed executions, and poor-quality demonstrations. This yields a final offline dataset of approximately 60k samples used for training.

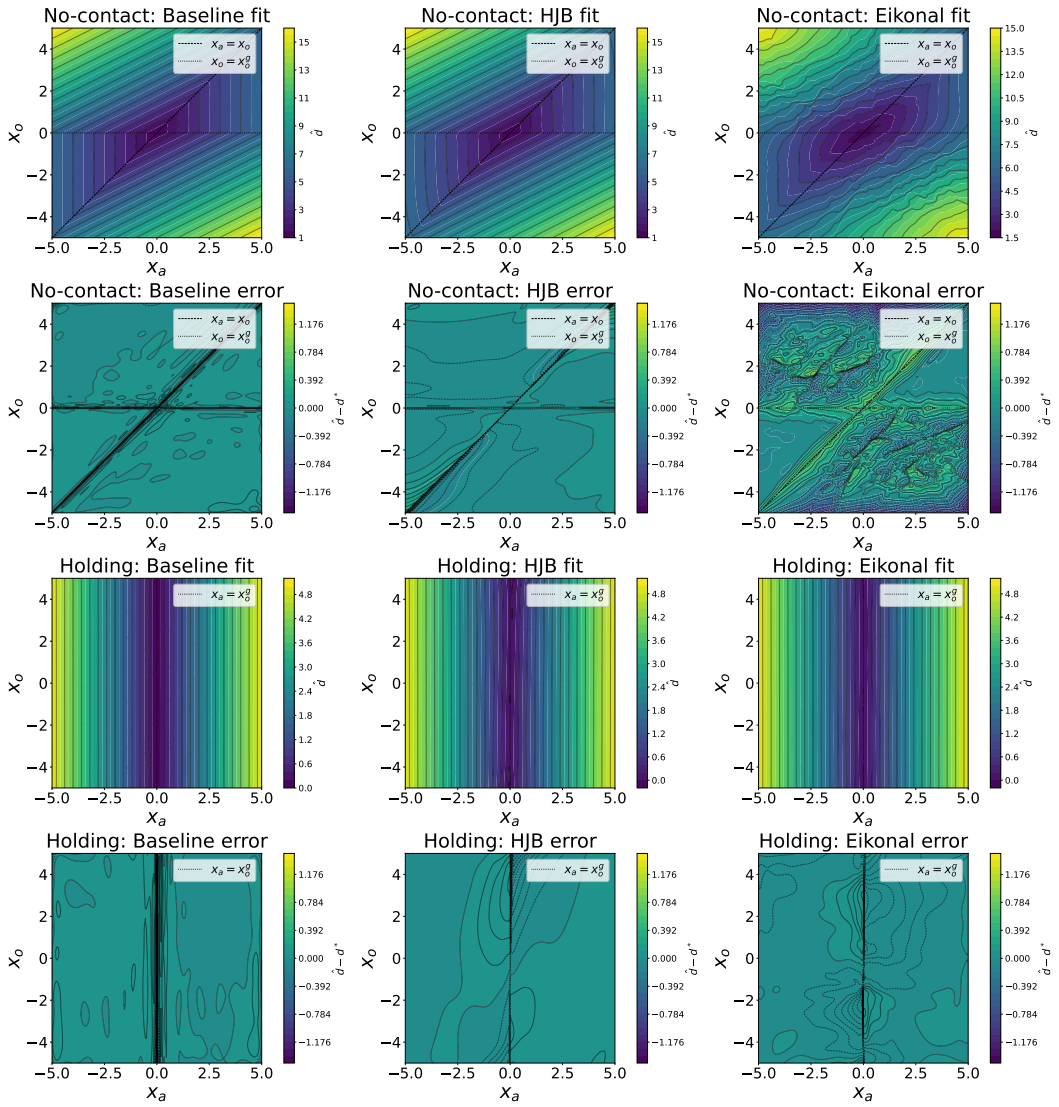


Figure 7: Extended visualization of the hybrid toy example in Fig. 6. Rows show the fitted cost-to-go $\hat{d}$ and the error $\hat{d} - d^*$ for the no-contact and holding modes. Columns compare the unregularized baseline in Eq. (19), the transition-based HJB regularizer in Eq. (20), and the full-state Eikonal regularizer in Eq. (21). In the no-contact mode, the Eikonal penalty introduces larger structured errors because the object coordinate affects the long-horizon cost-to-go while being locally uncontrollable. In contrast, the baseline and HJB-regularized solutions more closely recover the piecewise-linear structure of d^*. In the holding mode, where the object moves with the agent and the relevant coordinates are locally controllable, all methods recover similar value landscapes. Across methods, residual errors concentrate near the nondifferentiable switching and goal-alignment boundaries, where gradient-based regularization is naturally harder to satisfy.

D.2 Extended summary of the results

Table 2 provides a compact summary of the real-world evaluation results discussed in Section 6 of the main text. Policies are trained on approximately 60k interaction steps for 300k gradient steps and evaluated over 20 episodes on the physical system. Success rates are reported under three object-position thresholds, corresponding to 3cm, 5cm, and 10cm Euclidean distance from the center of the table. Overall, the results show that the HJB-style regularizer improves performance over the GCIVL baseline, and that performance improves further when physics-informed value learning

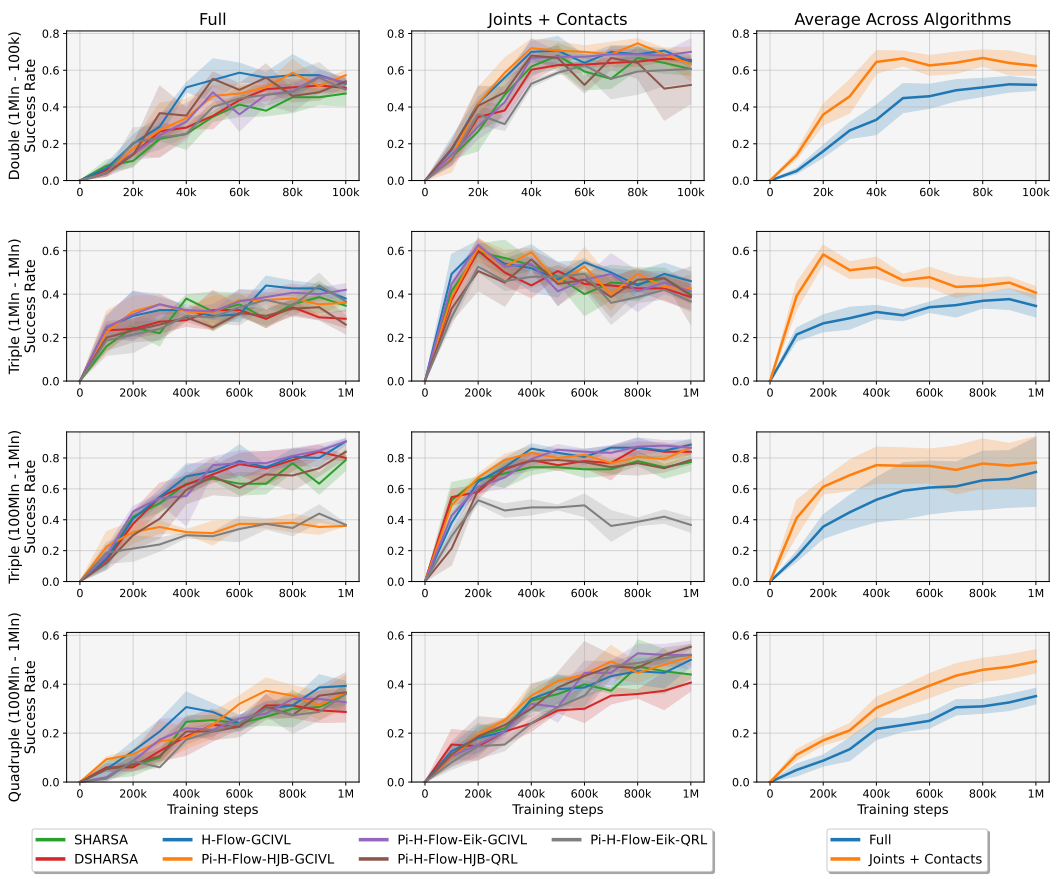


Figure 8: State representation ablation. We compare hierarchical algorithms using the full manipulation state at the low level against variants using only the controllable manipulator-related state. The left column reports results with the full low-level state, the middle column reports results with the proposed "Joints+Contacts" representation, and the right column averages performance across algorithms for each representation. The notation in each row title, e.g., `100Mln-1Mln`, denotes the dataset size and number of training steps, respectively. Evaluation is performed as in Fig. 3. Curves report the mean and standard deviation across seeds.

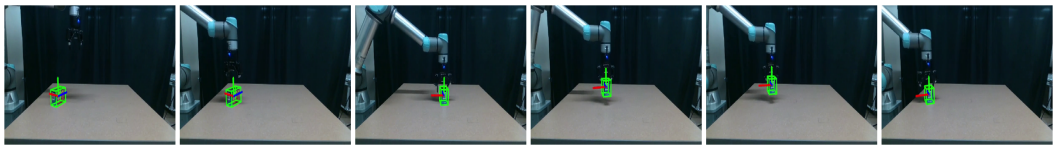


Figure 9: Images from data collection on the real-world pick-and-place setup, showing an initial pick, final placement, and object resetting. The task is performed with a UR5e robot, while RGB-D observations are recorded using an Intel RealSense D415 camera. The pose of the manipulated object is estimated using FoundationPose (Wen et al., 2024).

is combined with the proposed hierarchical decomposition. In particular, Pi-H-Flow-HJB-GCIVL achieves the strongest performance across all evaluation thresholds.

We further provide qualitative examples of the learned behavior in Fig. 10. These rollouts are collected from an HJB-GCIVL policy and are intended to complement the quantitative results by illustrating both successful behaviors and representative failure modes.

The rollout in Fig. 10a shows a recovery behavior: the agent initially misses the object due to poor end-effector alignment, but subsequently repositions, grasps the object, and completes the task.

Table 2: Real-world pick-and-place success rates over 20 evaluation episodes. Policies are trained on approximately 60k interaction steps for 300k gradient steps.

Method	3cm	5cm	10cm
GCIVL	25	40	45
HJB-GCIVL	40	50	65
H-Flow-GCIVL	25	40	40
Pi-H-Flow-HJB-GCIVL	**50**	**60**	**75**

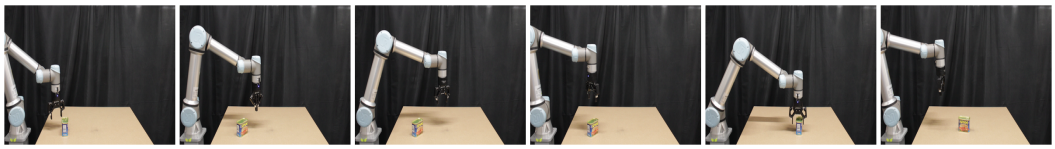
Success rates are reported in percent for object-position thresholds of 3cm, 5cm, and 10cm.

Fig. 10b shows a successful rollout in which the agent moves the object by sliding it along the table rather than directly lifting and placing it. This behavior was not explicitly present in the demonstrated dataset, suggesting that the learned policy can exploit alternative interaction modes.

Fig. 10c shows a representative failure case in which the agent gets stuck in mid-air before placing the object. In this configuration, we observed that the policy sometimes recovers, while in other cases external intervention is required.

Finally, Fig. 10d and Fig. 10e show two successful rollouts. The first corresponds to a smooth pick-and-place behavior similar to the demonstrated trajectories, whereas the second shows a less typical strategy in which the robot grasps the object near its rim due to imperfect object–end-effector alignment, but still recovers and completes the task.

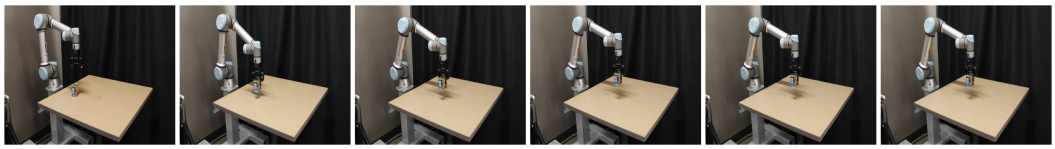
These real-world results provide evidence that the proposed approach can produce robust and adaptive manipulation behaviors on a physical system, while also highlighting remaining failure modes that motivate future work on safety, compliance, and more structured contact-aware control.



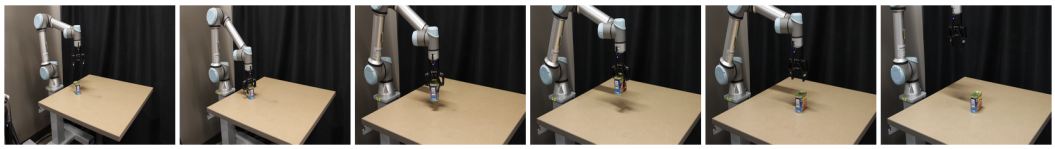
(a) Recovery after an initially misaligned grasp attempt.



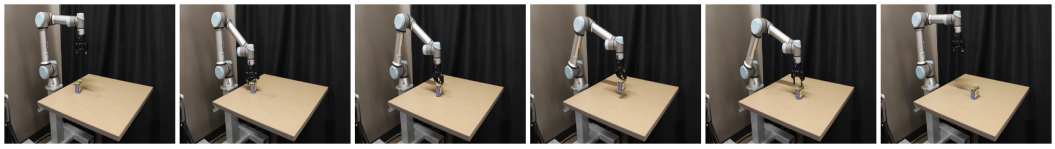
(b) Successful object transport through sliding.



(c) Failure case where the policy stalls before placement.



(d) Smooth successful pick-and-place rollout.



(e) Successful recovery after grasping near the object rim.

Figure 10: Qualitative examples of real-world rollouts collected from the learned policy. Each row shows temporally ordered frames from a representative trajectory, illustrating successful behaviors, recovery from imperfect grasps, alternative interaction modes, and representative failure cases.