When to Trust the Teacher? Adaptive Coupling in Interactive Robot Learning

Nikolaus Feith
Chair of Cyber Physical Systems
Technical University Leoben
Leoben, Austria
nikolaus.feith@unileoben.ac.at

Elmar Rückert
Chair of Cyber Physical Systems
Technical University Leoben
Leoben, Austria
elmar.rueckert@unileoben.ac.at

Abstract

Interactive robot learning methods typically treat the human teacher as an infallible oracle, limiting the agent's ability to surpass the expert or reject adversarial advice. We introduce MAGIC (Modulated Asymmetric Games for Interactive Control), a framework that formulates interactive learning as an asymmetric leader–follower game between a Teacher and a Learner. The Teacher is an inverse reward field—instantiated with energy-based and flow-matching heads—that scores trajectory segments in SE(3) via contrastive learning on expert demonstrations. The Learner is a hierarchical flow-matching policy (Eye, Brain, Muscle) that maximizes a shaped reward mixing environment reward and Teacher signal. A gradient-agreement coupling determines state-dependent trust: when the Teacher's directional signal agrees with the task critic's gradient, the Teacher is trusted; otherwise it is ignored. We prove that the alternating update satisfies the regularity conditions of two-timescale stochastic approximation. The core pipeline is implemented and unit-tested; we present the framework, its theoretical grounding, and the planned experimental evaluation on 9 ManiSkill3 manipulation tasks, LIBERO with noisy human demonstrations, and real-robot transfer on UR3e and SO-101 arms.

1 Introduction

Interactive robot learning (IntRL) combines human guidance with autonomous skill acquisition, yet most methods treat the teacher as a black-box oracle: the agent imitates demonstrated actions [Ross et al., 2011] or follows a fixed shaped reward [Knox and Stone, 2009, Brys et al., 2015]. This makes it difficult for the learner to surpass the expert or reject misleading advice from noisy or adversarial teachers.

We introduce MAGIC, which formulates IntRL as an asymmetric *leader–follower* bi-level game. The Teacher (leader) shapes the reward landscape via inverse reinforcement learning (IRL) on demonstrations, while the Learner (follower) optimizes a hierarchical policy under a shaped reward that mixes environment reward and the Teacher's signal. A state-dependent coupling weight β_t modulates the Teacher's influence, allowing the Learner to down-weight the Teacher when its advice would reduce task success.

Our contributions are: (1) a formalization of interactive learning as an asymmetric bi-level game with formal regularity guarantees under two-timescale stochastic approximation; (2) an inverse reward field Teacher with modular energy-based (EBM) and flow-matching (FM) heads operating on the same trajectory representations as the Learner; (3) a threshold-free gradient-agreement coupling that compares Teacher, critic, and policy vector fields in subgoal space. This paper presents the framework design, theoretical analysis, and implementation status. Experimental results are forthcoming; Section 4 details current progress and planned evaluation.

The Second Austrian Symposium on AI and Vision (AIROV25).

2 The MAGIC Framework

2.1 Asymmetric Bi-Level Formulation

MAGIC operates on $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}^L, \mathcal{A}^T, P, R^L, R^T, \gamma \rangle$, where only the Learner's actions enter the environment transition kernel. The Teacher's "actions" are interventions on the expert dataset $\mathcal{D}_E$. The Learner maximizes a shaped reward:

$$r_L(s_t, g_t) = \mathbb{E}_{\tau_t \sim f_\psi(\cdot|s_t,g_t)} \big[R_{\text{env}}(s_t, \tau_t) + \beta_t \, \hat{R}_T(s_t, \tau_t) \big], \tag{1}$$

where $\hat{R}_T$ is the Teacher's inverse-reward estimate and $\beta_t \in [0, 1]$ is the coupling weight. The inner loop updates Teacher parameters Θ via contrastive IRL; the outer loop updates Learner parameters via policy gradient on r_L. Timescale separation (N_T:1 update ratio) ensures the Teacher remains approximately stationary from the Learner's perspective [Borkar, 2008].

2.2 Hierarchical Learner (Eye, Brain, Muscle)

The Learner decomposes into three modules: **Eye** (ϕ): a VC-1-based [Majumdar et al., 2023] multi-view encoder mapping images and proprioception to state s_t; **Brain** (π_θ): a flow-matching policy in a 10D subgoal space (3D translation, 6D rotation [Zhou et al., 2020], 1D gripper) using importance-sampling flow-matching actor–critic [Zhang et al., 2025]; **Muscle** (f_ψ): an ActionFlow [Funk et al., 2024] model in SE(3) generating 16-step trajectory segments conditioned on (s_t, g_t). The Brain uses a *dual critic*: Q_{task} trained on R_{env} only, and Q_{shaped} on the full shaped reward. The Muscle retains g_t in the computational graph, making the Jacobian $J = \partial \tau_t / \partial g_t$ available via a single VJP—essential for the gradient-agreement coupling.

2.3 Inverse Reward Field (Teacher)

The Teacher scores trajectory segments via a weighted combination: $\hat{R}_T = \lambda_{\text{EBM}} \hat{R}_T^{\text{EBM}} + \lambda_{\text{FM}} \hat{R}_T^{\text{FM}}$. The **EBM head** maps (s_t, τ_t) to a scalar energy trained with an InfoNCE [Oord et al., 2018] objective using $K{=}4$ structured negatives. The **FM head** learns a velocity field via conditional flow matching [Lipman et al., 2022]; its cosine similarity with the target velocity serves as the scalar reward, and the velocity field itself provides a directional signal.

2.4 Gradient-Agreement Coupling

Our primary coupling is threshold-free, exploiting the fact that Teacher, Brain, and Critic all define vector fields in subgoal space. The Teacher's trajectory-space directional signal $\tilde{v}_T$ is projected to subgoal space via the Muscle's VJP: $v_T(s_t, g_t) = \left(\frac{\partial \tau_t}{\partial g_t} \right)^\top \tilde{v}_T(s_t, \tau_t) \in \mathbb{R}^{10}$. The coupling weight is:

$$\beta_t^{\text{agree}} = \max\big(0, \ \cos\big(\nabla_g Q_{\text{task}}(s_t, g_t), \ v_T(s_t, g_t)\big)\big) \, , \tag{2}$$

so the Teacher is trusted when its direction agrees with increasing task value, and ignored otherwise. A warm-up phase ($\beta_t = \beta_0$ for N_{warm} steps) prevents premature Teacher rejection before the critic is informative.

2.5 Convergence Guarantee

Proposition 1 (Regularity of MAGIC)**.** *Under spectral normalization on the EBM head, reward clipping $\hat{R}_T \in [-R_{\max}, R_{\max}]$, entropy-regularized policy optimization with $\alpha > 0$, and Robbins–Monro step sizes with $\alpha_n^L / \alpha_n^T \to 0$, the MAGIC iterates satisfy the Lipschitz, bounded-iterate, and martingale-noise conditions of Borkar [2008] (Ch. 6, Thm. 2) and track the corresponding two-timescale ODE almost surely.*

3 Planned Experimental Evaluation

All experiments use synthetic teachers (trained on offline demonstrations; no live human at interaction time). Results will report mean $\pm$ std over 3 seeds.

Table 1: ManiSkill3 evaluation tasks (Franka Panda, motion-planning demos).

Easy	Medium	Hard
PickCube-v1	StackCube-v1	PegInsertionSide-v1
PushCube-v1	LiftPegUpright-v1	PlugCharger-v1
PullCube-v1	PokeCube-v1	StackPyramid-v1

ManiSkill3 (primary benchmark). ManiSkill3 [Tao et al., 2024] provides GPU-accelerated parallel environments (>30K FPS), enabling large-scale ablation sweeps. We evaluate on 9 Franka Panda manipulation tasks spanning three difficulty levels (Table 1), with 100 motion-planner demonstrations per task and 2M training steps per run. For non-interactive baselines (BC, SAC, PPO), we cite ManiSkill3's published numbers.

LIBERO (noisy human demonstrations). LIBERO [Liu et al., 2023] provides tasks with 50 human-teleoperated demonstrations each—noisy and suboptimal compared to motion planners. We select 3 tasks from LIBERO-OBJECT to test whether MAGIC's coupling correctly down-weights a Teacher trained on imperfect data.

Real-world experiments (planned). (1) A UR3e (6-DOF) with three RealSense D435i cameras performs Duplo block assembly using 20 kinesthetic demonstrations, with sim-to-real transfer followed by real-world fine-tuning. (2) A low-cost SO-101 arm (5-DOF, ∼$300) with two USB webcams performs Sort-by-Color and Cup-on-Saucer using 30 teleoperated demonstrations, trained entirely on the real robot. Real-robot experiments are proof-of-concept; full human-in-the-loop evaluation is future work.

Ablation conditions. Full MAGIC (A0) is compared against: no Teacher (A1), EBM-only (A2), FM-only (A3), oracle Teacher (A4), threshold coupling (A5), advantage coupling (A6), naive negatives (A7), DAgger (A9), and ThriftyDAgger [Hoque et al., 2021] (A10). Full ablations run on all 9 ManiSkill3 tasks with 3 seeds; a demo efficiency sweep (10/25/50/100 demos) on PickCube-v1 characterizes data requirements.

Hypotheses. (H1) Teacher improves sample efficiency (A0 vs. A1). (H2) Combined EBM+FM Teacher outperforms either head alone (A0 vs. A2/A3). (H3) Gradient-agreement coupling outperforms threshold and advantage baselines (A0 vs. A5/A6). (H4) Structured negatives improve Teacher quality (A0 vs. A7). (H5) MAGIC enables the Learner to surpass expert demonstrations on ≥3 tasks.

4 Implementation Status

The MAGIC pipeline is fully implemented and unit-tested (346+ tests passing). All core components are complete: Eye (VC-1), Brain (ISFM + dual critic), Muscle (ActionFlow SE(3) with differentiable mode), Teacher (EBM with InfoNCE + FM head), all three couplings, and the full training loop with W&B logging. The ManiSkill3 wrapper supports all 9 tasks (0.4s/ep with VC-1) and 100 motion-planner demos have been collected per task.

The immediate next steps are: (1) training the Muscle on ManiSkill3 PickCube-v1, (2) running the no-Teacher baseline (A1) and full MAGIC (A0) for first learning curves, (3) scaling to all 9 tasks and the full ablation suite. LIBERO integration and real-robot experiments follow after simulation results are established.

5 Conclusion

We have presented MAGIC, a bi-level framework for interactive robot learning with gradient-agreement coupling and formal regularity guarantees. The pipeline is fully implemented; experiments on 9 ManiSkill3 tasks, LIBERO, and real-robot transfer are underway.

Acknowledgments

This work was created as part of the research project, MUTAVIA (FO999922732), which are funded by the Österreichische Forschungsförderungsgesellschaft mbH (FFG).

References

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 100. Springer, 2008.

Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E Taylor, and Ann Nowé. Reinforcement learning from demonstration through shaping. In *IJCAI*, pages 3352–3358, 2015.

Niklas Funk, Julen Urain, Joao Carvalho, Vignesh Prasad, Georgia Chalvatzaki, and Jan Peters. Actionflow: Equivariant, accurate, and efficient policies with spatially symmetric flow matching. *arXiv preprint arXiv:2409.04576*, 2024.

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*, 2021.

W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16, 2009.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791, 2023.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, et al. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024.

Yuyang Zhang, Yang Hu, Bo Dai, and Na Li. Max-entropy reinforcement learning with flow matching and a case study on lqr, 2025. URL https://arxiv.org/abs/2512.23870.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks, 2020. URL https://arxiv.org/abs/1812.07035.