HINTS: HUMAN-INTUITED CUES FOR REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

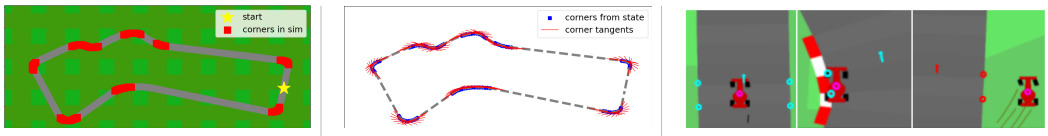
1 OVERVIEW

We present revisions to our recent submission. These include: additional implementation details (S2) and related work (S3), and reformatted results (S4).

2 IMPLEMENTATION

Our framework[1] augments the agent with a hint generator that outputs cues, or groundings of (human-provided) conceptual hints in the environment. We present two designs for the generator:

Programmatic generator G A program generates cues as a function of action sequences, from the simulated scene, the environment state, or rewards. Action-based cues are embeddings computed from sequences of joint states. State-based cues can be subsets of the environment state or derived from scene information. As an example, goal-distance used in Classic Control tasks and height, speed, max-velocity in Locomotion are derived from internal state (Section 3). The Car Racing environment has no state, so all state-based cues are derived from the track geometry. The program provided an indicator ℓ representing whether the car is on-track and a curvature κ measure computed from scene geometry in the immediate vicinity of the car hull (Figure 1). In particular, we approximate curvature as the change in neighbouring track tangent vectors. We also set a speed value in $[0, 1]$ based on the sharpness of the corner. There is an opportunity to scale up such programs using off-the-shelf code generators.



track corner (red) identification | (*state-based) cue estimation | *on/off, *curvature, speed

Figure 1: Visualisations of programmatic generation of grounded hints (cues) for Car Racing.

Learnt generator G_ϕ A neural generator takes only image observations as input to generate cues. Our implementation uses a ResNet architecture [He+15] which we adapted for regression tasks [Lathuilière+18]. The generator consists of a 34-layer backbone and two output heads for regression and classification. The backbone has (in order) an 1-layer CNN input layer, `MaxPool` operation, 16 ResNet2 blocks (each with 2-layer CNNs), `AveragePool` operation, and a fully connected output layer. We applied `BatchNorm` to all CNN layers, followed by `ReLU` activation. Regression and classification heads are 2-layer MLPs with `SoftMax` applied only to the classification head. For the input, the generator maintains a c-sized context window of observations. This context is processed into a c-channel input observation; specifically, a grey-scale image with zero-centered ($[-0.5, 0.5]$) pixels. The final output is the concatenation of the regression and classification outputs.

We pre-trained the generator using supervised learning to predict target cues for an environment. The training and test data were generated by rolling out a choice policy over H-horizon steps, with a 4:5 ratio split. We primarily used a non-uniform random policy with sampling concentrated away

[1] available at https://github.com/charlie-the-brave/anon-hints/tree/reviews

from small actions. We trained G_ϕ using a mean-squared-error loss for regression targets and cross-entropy loss for classification. Once trained, we freeze the generator weights before training agents conditioned on its output.

Scaling up More complex designs may be necessary for tasks with a non-trivial structure. For example tasks involving multiple objects might require object-specific cues that are co-dependent. In long horizon settings, cues might need to adapt to variations in the environment or task phases. An architecture that balances simplicity and flexibility together with a semi-supervised learning algorithm could enable HINTs to scale to challenging domains like dexterous manipulation.

3 Related Work

Human-in-the-loop RL incorporates human input into the learning process so that an agent can satisfy multiple objectives (e.g., efficiency, safety, adaptability, etc.) [Najar+21],[Retzlaff+24]. This paradigm has at least two schools of thought in which the human interacts dynamically with the agent or supervises the agent in an offline sense. Approaches with an iterative dynamic use human input to shape policies [Knox+09],[Ross+11], estimate reward functions and value [Christiano+17],[Liu+19],[Guan+21],[Ouyang+22],MacGlashan+23, and more [Abel+17]. On the other hand, human input can be integrated into learning without repeated interactions, as in reward inference from preferences, demonstrations [Finn+16],[Szot+23],[Kim+23], or suboptimal data [Muslimani+25]. In our approach, humans can target more than the reward function through the hint generator, contrary to the latter approaches. We thereby minimise the need for humans to provide expert demonstrations or accurate preferences by requesting conceptual hints which the generator grounds in the environment.

4 Presentation

Correction in Tab. 2 (shown as T1) The column showing 'HINTs-$\mathbf{x}$' for state-conditioning should be relabelled to 'HINTs-$\{h\}$' for full-conditioning, where $\{h\}$ the set of all cues to the left of 'composite'; e.g., goal distance together with angular velocity and action cue. The results for conditioning on raw state for Acrobot, Pendulum, Inverted Double Pendulum were affected by an illusive bug in the training data collection pipeline that swapped $\{h\}$ for $\mathbf{x}$. The bug has since been resolved in the published code.

Reformatting Tab. 2 & 3 We have reformatted Table 2 (shown as T1) as an example of modifications we plan to make. The cells containing hint label have been highlighted in grey.

Reformatting Fig. 3 We have restructured Figure 3 (shown as F2), focussing on the relative performance of HINTs-FC against vision-only and state-only baselines. The text has been enlarged for improved visualisation of the axes.

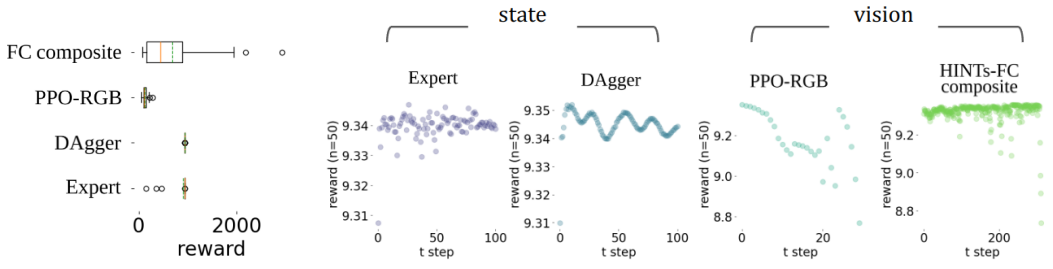


Figure 2: Inverted Double Pendulum reward over n=50 seeds. (Left) HINTs-FC outperforms PPO-RGB while approaching DAGGER. (Right) Instantaneous reward vs trajectory time step.

Table 1: Performance on image-based Classic Control tasks: Pendulum, Acrobot, and Inverted Double Pendulum (Inv-Double Pendulum) over n=50 random seeds. Baseline agents take state inputs. Grounded hints (shortened to hints here) are grey; ($\star$) denotes constituents of 'composite' hints. $\{h\}$ denotes the set of available environment cues. HINTs-FC agents (dark blue) significantly outperform PPO-RGB agents (orange) by conditioning on a variety of hints (via Eqn 3 (**??**)).

Avg reward ↑	No training budget			Tight train budget							
Agents	Expert	DAGGER	GAIL	PPO			HINTs-FC		(conditioning w/ more info →)			
	$\mathbf{x}$	$\mathbf{x}$	$\mathbf{x}$	$\mathbf{x}$	$\mathbf{x}+\mathcal{N}(0,1)$	RGB	$\star\, d\theta_1/dt$	$\star\, d\theta_2/dt$	joint	composite	$\mathbf{x}_{\text{partial}}$	$\{h\}$
Acro	-72.44 ±13.1	-67.30	-500.00 ±0.0	-197.90 ±36.4	-201.60 ±41.1	-500.00 ±0.0	-212.38 ±52.2	-228.98 ±63.2	-302.24 ±75.7	-197.40 ±41.4	-241.66 ±46.5	-500.00 ±0.0
Inv-Double Pendulum	906.34 ±150.2	943.85 ±0.8	943.26 ±0.7	944.67 ±0.2	109.98 ±25.9	130.99 ±41.8	$\star\, d\theta_1/dt$ 281.14 ±99.6	$\star\, d\theta_2/dt$ 400.34 ±215.6	D_{goal} 40.07 ±9.7	composite 680.84 ±659.8	$\mathbf{x}_{\text{partial}}$ 70.35 ±19.0	$\{h\}$ 88.56 ±14.4
Pendulum *Swingup*	-174.77 ±96.7	-253.93 ±216.4	-191.99 ±118.9	-1270.86 ±77.3	-1100.96 ±66.8	-1352.82 ±169.6	$\star\, d\theta/dt$ -250.31 ±164.0	$a=\tau(\theta)$ -1236.58 ±256.8	$\star\, D_{\text{goal}}$ -1203.08 ±307.6	composite -241.14 ±178.5	$\mathbf{x}_{\text{partial}}$ -1304.77 ±180.1	$\{h\}$ -1248.46 ±150.5
Pendulum *Swingto deploy*		111.82 ±137.8	-	120.15 ±85.3	115.20 ±44.5	21.26 ±138.9	$\star\, d\theta/dt$ 111.08 ±186.4	composite -118.48 ±72.0	$\{h\}$ 18.76 ±105.5			
Pendulum *Swingto train/eval*		-	-	121.89 ±65.1	-	22.81 ±130.1	$\star\, d\theta/dt$ 127.07 ±72.7	composite 246.14 ±99.1	$\{h\}$ 13.90 ±126.9			

5 REFERENCES

He et al. Deep Residual Learning for Image Recognition. CVPR 2015. URL https://arxiv.org/abs/1512.03385.

Lathuilière et al. A Comprehensive Analysis of Deep Regression. CVPR 2018. URL https://arxiv.org/abs/1803.08450.

Retzlaff et al. Human-in-the-Loop Reinforcement Learning: A Survey and Position on Requirements, Challenges, and Opportunities. JAIR 2024. URL https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2021.584075/full.

Najar et al. Reinforcement Learning With Human Advice: A Survey. FRAI 2021. URL https://dl.acm.org/doi/10.1613/jair.1.15348.

Muslimani et al. Leveraging Sub-Optimal Data For Human-Inthe-Loop Reinforcement Learning. ICLR 2025. URL https://arxiv.org/abs/2405.00746.

Szot et al. BC-IRL: Learning Generalizable Reward Functions From Demonstrations. ICLR 2023. URL https://arxiv.org/abs/2303.16194.

Kim et al. Learning Shared Safety Constraints from Multi-task Demonstrations. NIPS 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/124dde499d62b58e97e42a45b26d7369-Paper-Conference.pdf.

Ouyang et al. Training language models to follow instructions with human feedback. 2022. URL https://arxiv.org/abs/2203.02155.

Guan et al. Reinforcement Learning with Explanation and Context-Aware Data Augmentation. NIPS 2021. URL https://papers.nips.cc/paper_files/paper/2021/file/b6f8dc086b2d60c5856e4ff517060392-Paper.pdf.

Finn et al. Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization. 2016. URL https://proceedings.mlr.press/v48/finn16.pdf.

MacGlashan et al. Interactive Learning from Policy-Dependent Human Feedback ICML 2017.

Abel et al. Agent-Agnostic Human-in-the-Loop Reinforcement Learning. NIPS 2017. URL https://arxiv.org/abs/1701.04079.

Knox and Stone. Interactively Shaping Agents via Human Reinforcement The TAMER Framework. K-CAP 2009. URL https://www.cs.utexas.edu/ bradknox/papers/kcap09-knox.pdf.

Ross et al. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. AISTATS 2011. URL https://arxiv.org/abs/1011.0686.

Christiano et al. Deep Reinforcement Learning from Human Preferences. NIPS 2017. URL https://arxiv.org/abs/1706.03741.

Liu et al. Deep Reinforcement Active Learning for Human-in-the-Loop Person Re-Identification. ICCV 2019.
URL https://openaccess.thecvf.com/content_ICCV_2019/html/Liu_Deep_Reinforcement_Active_Learning_for_Human-in-the-Loop_Person_Re-Identification_ICCV_2019_paper.html.