Intent-Based Reward Inference for Value-Aligned Reinforcement Learning

Md Masudur Rahman and Juan P. Wachs
Edwardson School of Industrial Engineering, Purdue University, West Lafayette, IN 47907, USA
rahman64@purdue.edu, jpwachs@purdue.edu

Abstract

AI systems must act in ways that reflect human values and intentions, yet defining suitable reward signals remains a major challenge. Reinforcement learning enables agents to learn through trial and error, powering systems such as AlphaGo to superhuman performance. However, the common assumption that agents learn from a single reward function provided by the environment is often unrealistic beyond controlled benchmarks, and hand-crafted rewards can be brittle or misaligned with human intent. To address these alignment challenges, we propose Inference-Based Reinforcement Learning (InfeRL), a framework for training agents with rewards inferred to match human goals. InfeRL allows an agent to infer its own reward by comparing its behavior to a high-level goal. Goals can be expressed in natural language and interpreted through a vision–language model. This removes the need for explicit environment rewards and instead emphasizes semantic alignment with human-described success. We evaluate InfeRL on standard Gymnasium environments which provide clear ground-truth rewards for comparison. InfeRL achieves performance close to agents trained with environment rewards, while following tasks described in natural language rather than relying on handcrafted signals. It supports novel instructed behaviors, such as rotating or walking, purely from language goals, and demonstrates its capacity to handle multi-objective instructions involving spatial reasoning. This work represents a step toward reinforcement learning agents that are transparent, adaptable, and aligned with human values.

1 Introduction

Ensuring that AI systems behave in ways consistent with human values and intentions remains a central challenge in the development of safe and effective artificial agents. As learning agents are increasingly deployed in open-ended and safety-critical environments, the risks posed by misalignment grow significantly. Poorly specified objectives can be exploited by agents, leading to phenomena such as reward hacking, unsafe exploration, or superficially successful behavior that ultimately violates human expectations. A critical aspect of this problem lies in how agents are trained to evaluate success. Specifically, the challenge is to define and deliver reward signals that accurately capture human intent.

Reinforcement learning (RL) has traditionally operated within the Markov Decision Process (MDP) framework, wherein an agent maximizes a scalar reward signal provided by the environment. Although this setup has demonstrated effectiveness in many benchmark domains, it often proves unrealistic in real-world applications. In practical scenarios, ground-truth reward signals are typically brittle, hand-crafted, and may reflect proxy objectives rather than true human goals. Moreover, in real deployments such as delivery robots or household assistants, dense per-step reward feedback is rarely available, and when it is, it often fails to reflect nuanced human preferences such as safety, efficiency, or comfort.

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

In this work, we propose a different perspective. Rather than relying solely on externally provided reward signals, agents should be able to infer their own reward functions based on observations and goal specifications that are interpretable by humans. We introduce **Inference-Based Reinforcement Learning (InfeRL)**, a framework in which the reward function is computed internally by the agent rather than being specified by the environment. The agent evaluates its behavior against a high-level goal expressed in natural language, thereby framing reinforcement learning as a process of interpreting intent rather than maximizing a predefined numerical objective.

To investigate this idea, we conduct experiments in three Gymnasium environments: CartPole, MuJoCo Ant, and MuJoCo Walker2D. We introduce several modifications to the classic CartPole environment to explore a range of settings under our framework. The Ant and Walker2D environments, with their high-dimensional state and action spaces, serve to demonstrate both the generality of the framework and its ability to produce novel instructed behaviors. These environments provide ground-truth rewards that enable direct comparison with the semantically inferred rewards produced by InfeRL. Using pretrained vision-language models such as CLIP Radford et al. [2021], we score agent behavior against natural language prompts, for example, "keep the pole upright and the cart centered." We then assess whether agents trained using these inferred rewards exhibit behaviors aligned with the intended goals, and we evaluate robustness across variations in prompt phrasing.

We evaluate InfeRL using both PPO Schulman et al. [2017] and DQN Mnih et al. [2015] to demonstrate compatibility with continuous and discrete action spaces. Our experiments show that InfeRL matches ground-truth performance in standard tasks, enables new instructed behaviors such as rotation and walking, and partially solves multi-objective instructions with semantically rich specifications. These results highlight both the promise and current limitations of reward inference in handling abstract and compositional goals. Our contributions are summarized as follows:

- We introduce **InfeRL**, a framework where agents infer rewards by aligning behavior with high-level, language-based goals.
- We evaluate InfeRL on Gymnasium benchmarks, comparing inferred rewards with ground-truth signals and analyzing robustness to prompt and task variations.
- We show that InfeRL matches ground-truth performance in standard tasks, enables novel instructed behaviors, and highlights limitations in handling semantically complex instructions—offering insights into how RL algorithms can be improved for better alignment.

By reframing reward design as a problem of semantic alignment, we offer a path toward reinforcement learning agents that more effectively understand and pursue human-defined success. We argue that such frameworks are essential for building AI systems whose behaviors are transparent, adaptable, and aligned with human values.

2 Background and Related Work

Reinforcement Learning and Reward Design. Reinforcement learning (RL) is commonly formulated under the Markov Decision Process (MDP) framework, where an agent interacts with an environment defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$. At each timestep t, the agent observes a state $s_t \in \mathcal{S}$, takes an action $a_t \in \mathcal{A}$, and receives a *scalar reward* $r_t = r(s_t, a_t)$ from the environment. The agent then learns a policy $\pi(a_t|s_t)$ that maximizes the expected discounted return over time. While this formulation is elegant, it assumes that the environment can always provide a meaningful reward signal. In many real settings this assumption is unrealistic. Designing such reward signals is difficult. Reward functions are often hand crafted and reflect brittle heuristics or proxy objectives that do not capture the true intent of a task. Agents can exploit loopholes in these rewards, leading to reward hacking or unintended behaviors. In real-world deployments such as robotics or interactive systems, reward signals may be unavailable, delayed, or unobservable He et al. [2024], Anderson et al. [2021], Kadian et al. [2020], Chang et al. [2025], Hsu et al. [2023], Rusu et al. [2017], Peng et al. [2018], James et al. [2017], Hsu et al. [2023], Levine et al. [2018], Pinto and Gupta [2016].

Approaches Beyond Hand-Crafted Rewards. Several approaches have been explored to overcome these limitations. Inverse reinforcement learning (IRL) Ng and Russell [2000] seeks to infer reward functions from demonstrations or comparisons, but it depends on large amounts of human input or curated datasets. Intrinsic motivation methods such as curiosity driven exploration Pathak et al. [2017], Burda et al. [2018] provide task agnostic signals that encourage exploration, yet they do not ensure

alignment with human specified goals. Goal conditioned reinforcement learning Andrychowicz et al. [2017] introduces goal vectors as inputs to policies but still relies on external reward functions.

Vision-Language Models and Supervised Approaches. Other research relies on supervised learning. Vision-language-action (VLA) models Zitkovich et al. [2023], Brohan et al. [2022], Kim et al. [2024], O'Neill et al. [2024] train agents using paired image, text, and action data. These models can follow prompts but depend on large scale human labeled data and do not involve trial and error learning. They lack the flexibility of reward based adaptation.

Recent progress in pretrained vision-language models (VLMs) such as CLIP Radford et al. [2021], VideoCLIP Xu et al. [2021], Flamingo Alayrac et al. [2022], GPT-4 Achiam et al. [2023], LLAVA Liu et al. [2023, 2024], and Qwen Yang et al. [2025] shows that such models can ground semantics across modalities. These models encode rich visual and linguistic features and can generalize to new inputs. They have been used in classification, retrieval, and control, but mostly as perception modules or auxiliary scorers.

Reward Inference with VLMs. Building on these ideas, recent studies explore using VLMs as reward providers. Rocamonde et al. Rocamonde et al. [2024] show that VLMs can serve as zero-shot reward models, enabling agents to learn from natural language prompts without hand designed rewards. Baumli et al. Baumli et al. [2023] analyze how VLM capacity affects reward quality for visual tasks. Wang et al. [2024] integrate VLM feedback into RL pipelines to generate task specific reward functions. These works demonstrate feasibility but still treat the reward model as an external scoring module.

Reward-Free Frameworks. Other perspectives, such as reward free frameworks, suggest that agents can learn without explicit rewards by discovering skills or predicting representations. For example, DIAYN Eysenbach et al. [2019] learns diverse skills by maximizing behavioral distinguishability, and curiosity driven exploration Burda et al. [2018] encourages visiting novel states. These methods avoid brittle rewards but do not offer direct alignment with high level goals.

Position of This Work. Inference-Based Reinforcement Learning (InfeRL) builds on these developments and introduces a different perspective. Instead of relying on environment provided rewards, InfeRL lets the agent compute its own reward internally. A pretrained semantic model evaluates how well recent behavior aligns with a goal expressed in natural language or other modalities. This makes the reward an interpretable signal based on intent rather than a fixed numerical function. InfeRL keeps the trial and error nature of reinforcement learning but moves reward generation inside the agent. This design connects to prior work on semantic reward modeling while addressing alignment concerns. It enables agents to adapt to new tasks by changing the goal description, and it provides a clearer path for human oversight and adjustment.

3 Inference-Based RL Framework

In standard reinforcement learning, an agent operates within a Markov Decision Process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where r is a scalar reward provided by the environment. The agent aims to learn a policy $\pi(a|s)$ that maximizes the expected cumulative reward:

$$\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$$

This formulation assumes that the environment provides a well-defined, informative reward signal at each step. In many real-world applications this assumption does not hold. Reward signals are often incomplete or even misaligned with the true intent of the task, which can lead to behaviors that technically optimize the reward while violating human expectations.

Reward Inference from Goals. InfeRL defines the modified RL problem as (Figure 1):

$$\mathcal{M}' = (\mathcal{S}, \mathcal{A}, P, \mathcal{G}, f_{\text{inf}}, \gamma)$$

where:

- $\mathcal{G}$ is the space of high-level task goals (e.g., language prompts or images),
- $f_{\text{inf}} : (\tau, g) \to \mathbb{R}$ is a reward inference function comparing recent behavior τ to a goal g,

Standard Reinforcement Learning (MDP)
Formalism:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$$

- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $P(s'|s,a)$: transition dynamics
- $r(s,a)$: reward from environment
- γ: discount factor

Objective:

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$$

Reward: Environment-provided
Policy: $\pi(a_t|s_t)$

Inference-Based RL (InfeRL)
Formalism:

$$\mathcal{M}' = (\mathcal{S}, \mathcal{A}, P, \mathcal{G}, f_{\text{inf}}, \gamma)$$

- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $P(s'|s,a)$: transition dynamics
- $\mathcal{G}$: goal space (e.g., text)
- $f_{\text{inf}}(\tau, g)$: inferred reward
- γ: discount factor

Objective:

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t f_{\text{inf}}(\tau_t, g)\right]$$

Reward: Inferred by agent
Policy: $\pi(a_t|s_t, f_{\text{text}}(g))$

Figure 1: InfeRL replaces externally defined rewards with internally inferred signals based on semantic alignment between behavior and goals.

- $(\mathcal{S}, \mathcal{A}, P, \gamma)$ follow the standard MDP definition.

InfeRL removes the dependency on externally provided reward signals and instead aligns behavior through goal-driven reward inference. Rather than blindly following a hand-crafted reward, the agent interprets high-level goals expressed by a human and generates its own reward signals based on semantic alignment. This design encourages agents to optimize for what humans actually intend, reducing the risk of reward hacking and unintended side effects.

At each step, the agent receives a goal $g \in \mathcal{G}$ and collects a short trajectory $\tau_t = [I_{t-k+1}, \ldots, I_t]$ of observations. A pretrained perceptual model then computes alignment:

$$r_t = f_{\text{inf}}(\tau_t, g) = \alpha \cdot \cos\big(f_{\text{vision}}(\tau_t), f_{\text{text}}(g)\big),$$

where α is a scaling parameter. f_{vision} and f_{text} map visual clips and goal descriptions into a shared embedding space. The cosine similarity between these embeddings provides a reward that reflects how well the agent's behavior aligns with the goal.

Goal-Conditioned Policies. We condition the policy on a goal embedding $u = f_{\text{text}}(g)$, allowing a single agent to generalize across a range of tasks and goals:

$$\pi(a_t \mid s_t, u)$$

At test time, the agent can receive a new instruction in natural language without any retraining and continue to infer reward signals that align with the newly specified intent.

Alignment-Centric Interpretation. InfeRL preserves the reinforcement learning loop of exploration, credit assignment, and policy improvement. The key difference is that the reward signal is not a fixed external number but a flexible, interpretable alignment score. By grounding rewards in human-readable goals, InfeRL promotes transparency, enables oversight, and encourages behaviors that better reflect human values. This shift supports the broader aim of building agents whose objectives are not only optimized but also aligned with the intent behind their design.

4 Theoretical Analysis of Inference-Based Rewards in RL

We present an analysis of reinforcement learning with *inferred rewards* $\hat{r}$ obtained from models such as VLMs. Our goal is to characterize the precise conditions under which standard RL algorithms converge to policies that are near-optimal under the true reward r.

4.1 Preliminaries

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ denote the true discounted infinite-horizon MDP, where $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the ground-truth reward function and $\gamma \in (0, 1)$ is the discount factor. The value of a policy π under r is defined as

$$V_r^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s \right].$$

We assume access only to an inferred reward $\hat{r}$—possibly stochastic—generated from observation of trajectories by an inference model (e.g., a VLM). The optimal policies under r and $\hat{r}$ are denoted π_r^* and $\pi_{\hat{r}}^*$, respectively.

4.2 Temporal Consistency and Markovianity

Assumption 1 (Temporal Consistency)**.** *There exists $k \geq 1$ such that for all (s_t, a_t),*

$$\hat{r}(s_t, a_t) = f(r(s_t, a_t), s_{t:t+k}),$$

where $s_{t:t+k}$ denotes the k-length trajectory segment, and f is a deterministic mapping.

This assumption requires that the inferred reward $\hat{r}$ is Markovian, or at least quasi-Markovian, when conditioned on a bounded temporal window. In other words, while a single state-action pair may be insufficient to determine the reward unambiguously, a short trajectory segment can disambiguate situations that appear similar at the level of instantaneous observations but are semantically distinct. Without this property, the inferred reward may conflate states corresponding to qualitatively different outcomes, leading to systematic misalignment between $\hat{r}$ and the true reward r.

4.3 Quasi-Markov Rewards via Augmented MDPs

We formalize the notion of "quasi-Markov" rewards by showing that if the inferred reward depends on a bounded-length trajectory window, one can construct an *augmented MDP* in which the same reward is Markovian, allowing standard RL guarantees to apply.

Definition 1 (Quasi-Markov reward: **past-window** form)**.** *Fix $k \in \mathbb{N}$. An inferred reward $\hat{r}$ is* k-quasi-Markov (past-window) *if there exists a measurable function*

$$g : \mathcal{S}^k \times \mathcal{A} \times \mathcal{U} \to \mathbb{R}$$

such that for all t,

$$\hat{r}_t \equiv \hat{r}(s_t, a_t, u) \;=\; g\big((s_{t-k+1}, \ldots, s_t),\, a_t,\, u\big),$$

where $u \in \mathcal{U}$ is a fixed goal or condition (e.g., a language embedding) that remains constant within an episode.[1]

Definition 2 (Quasi-Markov reward: **future-window (look-ahead)** form)**.** *Fix $k \in \mathbb{N}$. An inferred reward $\hat{r}$ is* k-quasi-Markov (future-window) *if there exists a measurable function*

$$h : \mathcal{S}^{k+1} \times \mathcal{A} \times \mathcal{U} \to \mathbb{R}$$

such that for all t,

$$\hat{r}_t \equiv \hat{r}(s_t, a_t, u) \;=\; h\big((s_t, \ldots, s_{t+k}),\, a_t,\, u\big).$$

Remark. Definition 1 corresponds to computing rewards from a short *history* (e.g., a clip ending at time t). Definition 2 corresponds to rewards that require a short *look-ahead* (e.g., verifying stability over the next k frames). The latter can be made causal by delaying the reward by k steps.

5 Experiments

We evaluate the InfeRL framework in standard continuous control environments to demonstrate that agents can learn meaningful behaviors from natural language prompts and visual observations alone, without relying on environment-supplied rewards. To investigate this, we focus on three Gymnasium

[1] Padding for $t < k$ can be handled via sentinel states or an initial distribution over $\mathcal{S}^{k-1}$.

environments: CartPole, MuJoCo Ant, and Walker2D. We design several variations of the classic CartPole environment to showcase our framework across different settings. The Ant and Walker2D experiment highlights the generalizability of our method, as it involves a more complex environment with high-dimensional action and observation spaces. These tasks also provide clear ground-truth rewards, allowing for direct comparison with inferred semantic rewards. We employ pretrained vision-language models (e.g., CLIP Radford et al. [2021]) to evaluate agent behavior against textual prompts, for instance, a prompt like "keep the pole upright and the cart centered."

5.1 Experimental Setup

To evaluate the effectiveness of InfeRL, we design experiments that isolate and test the agent's ability to infer rewards from natural language prompts and visual feedback. Our goal is to demonstrate that agents can learn meaningful and goal-aligned behaviors without access to manually engineered reward functions. We consider two widely used continuous control domains and introduce multiple task variations within each to assess generalization, interpretability, and robustness to prompt shifts. Details environments and task setups are in Appendix.

Reward Inference Mechanism. At each timestep, a pretrained vision–language model (e.g., CLIP) encodes a short window of visual observations into an embedding. This is compared to the goal embedding using cosine similarity, producing an inferred reward signal. The inferred reward replaces the environment's native reward and is used directly for training.

Reinforcement Learning Algorithm. We use Proximal Policy Optimization (PPO) Schulman et al. [2017] and Deep Q-Networks (DQN) Mnih et al. [2015] for reinforcement learning, with DQN applied in environments with discrete action spaces. Implementations are based on the Stable-Baselines3 library Raffin et al. [2021]. Both algorithms are trained solely using inferred reward signals. The algorithm implementation is used without modification, treating the similarity score from the vision–language model as a per-step reward. This illustrates how InfeRL can be readily integrated into standard RL pipelines without changes to the learning algorithm itself.

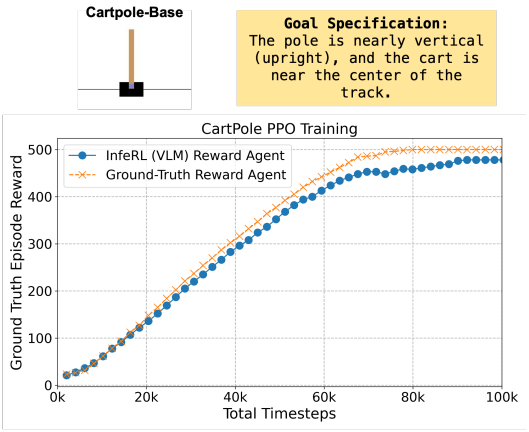

Figure 2: **CartPole PPO training results.** InfeRL achieves performance comparable to a ground-truth reward agent, demonstrating that inferred rewards based on natural language goals can effectively guide policy learning.

Alignment Focus. Our experiments are designed not only to benchmark performance but also to probe alignment. By observing how agents interpret and act upon high-level natural language instructions, we assess whether inferred rewards lead to behaviors consistent with human intent.

Evaluation Metrics. We evaluate performance both quantitatively and qualitatively. Quantitatively, we compare episode returns between InfeRL and baseline agents. Qualitatively, we assess behavioral alignment with prompts, and identify any misaligned or unintended behaviors. Evaluation was conducted across five training runs per environment. For each trained policy, we performed ten rollouts and manually labeled behaviors as either successful or unsuccessful based on alignment with the natural language goal. Due to the absence of a ground-truth reward function in these settings, we did not report learning curves but instead relied on qualitative evaluation of observed behaviors.

5.2 Results

5.2.1 Matching Ground-Truth Reward Performance

A core objective of InfeRL is to evaluate whether agents can learn effective behaviors by relying solely on reward signals inferred from natural language instructions and visual observations. In this

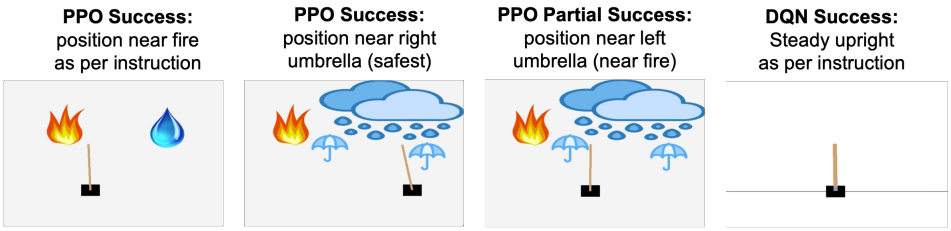

Figure 3: **CartPole instruction following with PPO and DQN.** Visualizations of final agent behaviors under different prompt types and environments. **Left to Right:** (1) PPO-trained agent in the FireWater setting learns to position itself near the fire icon, in line with the provided goal. (2) In the MultiCue environment, the PPO agent successfully navigates to the rightmost umbrella, avoiding fire and rain, as instructed. (3) A partial success in the same MultiCue environment, where the agent stops near the left umbrella, satisfying some but not all constraints. (4) DQN agent, operating with a **discrete action space**, is evaluated on the base CartPole setup and learns to stay upright and centered as specified in the instruction. These results highlight InfeRL's ability to support multi-objective goals under both continuous (PPO) and discrete (DQN) control regimes (video in supplementary).

section, we examine whether such agents can match the performance of traditional reinforcement learning agents trained with direct access to ground-truth rewards.

In the CartPole task, shown in Figure 2, the agent is instructed to balance the pole while keeping the cart near the center of the track. Results are averaged over five runs, with low variance omitted from the figure. The ground-truth reward in this environment is defined by episode length, which corresponds to the number of steps the agent maintains balance before termination. Our results show that the InfeRL agent achieves return curves comparable to the baseline trained with environment-provided rewards. This demonstrates that the inferred reward signal, computed from the alignment between goal instructions and observed behavior, is sufficient to support stable and effective policy learning.

In the Ant-Balance setting, the goal is to walk while staying balanced, as defined in the semantic goal specification listed in Table 1. Despite the complexity of the MuJoCo Ant environment, which requires coordination across a high-dimensional action space, we observed that the agent trained with InfeRL learns to walk consistently away from its starting position which is the ground-truth expected behavior Towers et al. [2024], Brockman et al. [2016]. The learned behavior aligns well with the intended instruction, even in the absence of a handcrafted reward function. This further supports the generalizability of InfeRL beyond simple control tasks.

These two experiments establish that inferred reward signals can match the performance of ground-truth signals in both simple and complex continuous control domains. The ability to achieve behaviorally aligned outcomes without relying on manually engineered rewards is an important step toward building agents that can generalize across diverse goals in real-world scenarios.

5.2.2 Generalization to Multi-Objective and Compositional Goals

Beyond matching ground-truth rewards in standard control settings, we investigate whether InfeRL can generalize to more complex scenarios involving multi-objective and compositional goals. In particular, we evaluate two modified versions of the CartPole environment: **CartPole-FireWater** and **CartPole-MultiCue**, both of which require the agent to interpret and act upon symbolic visual cues in accordance with multi-part natural language instructions (see Table 1).

In both environments, InfeRL successfully learns to align with the goal instruction within 100K timesteps. For example, in the CartPole-FireWater setting, the agent is prompted to move the cart toward the fire icon and away from the blue water droplet while maintaining an upright pole. The agent most often moves toward the fire region, suggesting that the vision–language model correctly aligns the visual scene with the prompt. However, due to the multi-objective nature of the task, the agent also needs to keep the pole upright. This requirement can at times conflict with positional goals, causing the agent to move away or lose stability when attempting to satisfy both conditions simultaneously.

In the CartPole-MultiCue setting, the agent is given an even more complex instruction involving fire, umbrellas, and rain. We observe that the agent more reliably aligns with the correct visual targets in this case (for the multi-objective setting). One possible explanation is that the visual representation of water droplets and umbrellas in this environment is more salient and semantically distinct compared to the more stylized or cartoonish water droplet in CartPole-FireWater. The fire icon, on the other hand, remains visually prominent across both settings. This highlights a limitation of vision–language models like CLIP, which may struggle with less prototypical visual representations when computing semantic similarity.

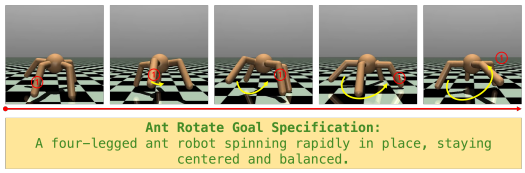

Figure 4: **Ant-Rotate behavior guided by a language-specified goal.** A sequence of frames showing the Ant agent rotating counterclockwise in place. Red circles mark a front leg for orientation; yellow arrows indicate the direction of rotation. The behavior is learned solely from natural language-based reward inference, without any handcrafted shaping or environment-provided reward (video in supplementary).

To further evaluate multi-objective and instruction-following behavior, we qualitatively analyze agent performance across the CartPole FireWater and MultiCue variants (Figure 3). In the FireWater task, the agent consistently positions itself near the fire icon, in accordance with the given prompt. In the more complex MultiCue setting, the agent often navigates to the rightmost umbrella, which is far from both fire and rain, reflecting successful multi-constraint alignment. However, in some runs, the agent stops under the left umbrella, leading to a partial success. This highlights sensitivity to spatial distinctions in the visual scene and to how well semantic embeddings capture relative object configurations.

Additionally, we validate the framework's compatibility with **discrete action spaces** by training a DQN Mnih et al. [2015] agent in the standard CartPole setting. The agent achieves stable upright behavior, matching the goal prompt. These results demonstrate that InfeRL supports both continuous and discrete control, and can generalize across multi-objective goals when the visual and linguistic cues are sufficiently expressive.

5.2.3 Learning Instructed Novel Behavior from Language Goals

In this experiment, we evaluate whether InfeRL can guide the agent to perform a novel but explicitly instructed behavior within the **MuJoco Ant** environment. The agent receives the goal description: *"a four-legged ant robot spinning in place while staying balanced."* This task requires a significant departure from the default forward locomotion typically learned in standard Ant environments and instead involves producing symmetric leg movements that achieve rotation without translation.

As illustrated in Figure 4, the agent successfully learns to rotate in place while maintaining balance. Visual evidence shows consistent angular displacement across frames, with red leg markers and yellow arrows confirming stable in-place spinning. Importantly, this behavior was learned without any handcrafted reward shaping or motion specification, relying entirely on the inferred reward signal produced by the vision–language model.

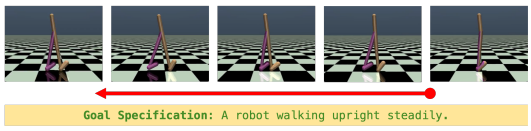

Figure 5: **MuJoCo Walker2D behavior under inferred reward.** A sequence of frames showing the agent walking upright steadily in the backward direction. The agent learns to maintain balance and upright posture but chooses to walk in reverse, highlighting a partial success and the impact of ambiguous language instructions (video in supplementary).

This result confirms that InfeRL can guide agents to execute non-default behaviors that align with human-intended goals, derived purely from natural language instructions. It demonstrates the flexibility of the framework to generalize beyond typical reward structures present in the environment.

We further examine the generality of InfeRL by testing its ability to guide motion in the **MuJoCo Walker2D** environment. Here, the agent is instructed using the prompt: *"A robot walking upright*

steadily." As shown in Figure 5, the agent learns a stable walking gait that maintains upright posture and leg coordination. However, the learned behavior consistently moves in the backward direction.

This outcome highlights an important challenge in reward inference from natural language. While the instruction successfully drives coordinated walking, the absence of explicit directional language leads the agent to adopt a gait that is semantically plausible but misaligned with implicit human expectations. The vision–language model appears to judge backward walking as consistent with the phrase "walking upright," which lacks positional or goal-oriented constraints.

5.2.4 Limitations of Reward Inference in Complex Goal Settings

While InfeRL demonstrates strong performance in both single- and multi-objective tasks, it encounters limitations in environments that require nuanced semantic reasoning. For instance, in the CartPole-MultiCue setting, the agent is instructed to stay upright while positioning itself under the rightmost umbrella, far from fire and rain. Although the agent maintains balance as required, it often stops at a nearer umbrella or stays close to its initial position, leading to only partial success in achieving the full instruction (Figure 3).

This outcome highlights two primary challenges. First, vision–language models used for reward inference may lack sufficient expressivity to distinguish between subtle, stylized visual features—particularly in settings where semantic meaning is spatially distributed or symbolically abstract. Second, inferred rewards are often computed over short temporal windows and may not reflect the long-term goals described in natural language. Such rewards can be informative in a local context but fail to guide behavior globally, especially when the semantics of the task are non-Markovian or require delayed credit assignment. As a result, standard reinforcement learning algorithms may converge to locally optimal but globally misaligned behaviors.

5.2.5 Implications for Algorithm Design

The limitations above suggest several design considerations for algorithms operating under inferred reward settings. InfeRL alters the fundamental assumption of reward observability in reinforcement learning by replacing ground-truth signals with rewards inferred from vision–language alignment. Because inferred rewards are temporally aggregated and can vary in scale or consistency, existing algorithms like PPO or DQN may require modification to handle reward sparsity and ambiguity.

First, stabilizing training may require normalization, smoothing, or segment-level averaging of reward signals to improve signal-to-noise ratio. Second, policies should be explicitly conditioned on goal representations to enable prompt-based generalization. Incorporating a goal embedding into the policy allows a single agent to adapt to different tasks and instructions without retraining. Third, because inferred rewards can be recomputed post hoc from stored trajectories, off-policy learning and data reuse become more powerful in this setting.

Finally, there is a broader need to explore architectures that move beyond the traditional MDP framework. Hierarchical or memory-augmented agents, for example, could better interpret and align with high-level, temporally extended goals. These directions may help close the gap between short-term semantic feedback and long-term behavioral alignment, paving the way for more instruction-aware and robust learning systems.

6 Conclusion

We introduced Inference-Based Reinforcement Learning (InfeRL), a framework in which agents learn behaviors by inferring rewards from the alignment between natural language goals and visual observations. Our results demonstrate that InfeRL can match the performance of agents trained with ground-truth rewards, support goal-directed generalization, and enable novel behaviors based solely on high-level instructions. While effective in many scenarios, challenges remain in handling semantically rich or multi-objective goals, where inferred rewards may be locally informative but insufficient to guide long-horizon behavior. These findings highlight the need for learning algorithms that better integrate temporal structure, semantic grounding, and goal representations. InfeRL offers a step toward more interpretable and flexible reinforcement learning, advancing the use of language as a tool for behavior alignment.

Acknowledgments and Disclosure of Funding

This work was supported by NSF under Award No. 521982.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. Sim-to-real transfer for vision-and-language navigation. In *Conference on Robot Learning*, pages 671–681, 2021.

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.

Kate Baumli, Satinder Singh, Feryal Behbahani, Harris Chan, Gheorghe Comanici, Sebastian Flennerhag, Maxime Gazeau, Kristian Holsheimer, Dan Horgan, Michael Laskin, et al. Vision-language models as a source of rewards. In *Second Agent Learning in Open-Endedness Workshop*, 2023.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. In *arXiv preprint arXiv:2212.06817*, 2022.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Wei-Jer Chang, Francesco Pittaluga, Masayoshi Tomizuka, Wei Zhan, and Manmohan Chandraker. Safe-sim: Safety-critical closed-loop traffic simulation with diffusion-controllable adversaries. In *European Conference on Computer Vision*, pages 242–258, 2025.

Benjamin Eysenbach, Julian Ibarz, Abhishek Gupta, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Haoran He, Peilin Wu, Chenjia Bai, Hang Lai, Lingxiao Wang, Ling Pan, Xiaolin Hu, and Weinan Zhang. Bridging the sim-to-real gap from the information bottleneck perspective. In *8th Annual Conference on Robot Learning*, 2024. URL `https://openreview.net/forum?id=Bq4XOaU4sV`.

Kai-Chieh Hsu, Allen Z Ren, Duy P Nguyen, Anirudha Majumdar, and Jaime F Fisac. Sim-to-lab-to-real: Safe reinforcement learning with shielding and generalization guarantees. *Artificial Intelligence*, 314:103811, 2023. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2022.103811. URL `https://www.sciencedirect.com/science/article/pii/S0004370222001515`.

Stephen James, Andrew J Davison, and Edward Johns. Transferring end-to-end visuomotor control from simulation to real world for a multi-stage task. In *Conference on Robot Learning*, pages 334–343, 2017.

Abhishek Kadian, Joanne Truong, Aaron Gokaslan, Alexander Clegg, Erik Wijmans, Stefan Lee, Manolis Savva, Sonia Chernova, and Dhruv Batra. Sim2real predictivity: Does evaluation in simulation predict real-world performance? *IEEE Robotics and Automation Letters*, 5:6670–6677, 2020.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *8th Annual Conference on Robot Learning*, 2024.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37:421–436, 2018.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=w0H2xGHlkw`.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, and Georg Ostrovski. Human-level control through deep reinforcement learning. *nature*, 518:529–533, 2015.

Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3803–3810, 2018. doi: 10.1109/ICRA.2018.8460528.

Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 3406–3413, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL `http://jmlr.org/papers/v22/20-1364.html`.

Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=N0I2RtD8je`.

Andrei A Rusu, Matej Vecerik, Thomas Rothörl, Nicolas Heess, Razvan Pascanu, and Raia Hadsell. Sim-to-real robot learning from pixels with progressive nets. In *Conference on robot learning*, pages 262–270, 2017.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. Rl-vlm-f: Reinforcement learning from vision language foundation model feedback. In *Proceedings of the 41th International Conference on Machine Learning*, 2024.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 2165–2183. PMLR, 06–09 Nov 2023. URL `https://proceedings.mlr.press/v229/zitkovich23a.html`.

A Experiment Settings

A.1 Environments

We evaluate InfeRL across diverse control environments to assess its ability to generalize inferred rewards from natural language instructions. Our chosen domains span both simple and high-dimensional tasks, enabling us to test semantic alignment under varying levels of complexity.

CartPole: We design three variants of the classic CartPole environment to evaluate semantic reward inference under natural language instructions (Figure 6):

- *CartPole-Base*: The standard setup where the agent balances a pole on a moving cart.
- *CartPole-FireWater*: A modified environment with symbolic visual cues—fire on the left and a water droplet on the right—used to evaluate direction-sensitive instructions such as "move toward water" or "avoid fire."
- *CartPole-MultiCue*: A more visually complex setting with fire icons on the left, and water droplets, umbrellas, and clouds in the center and right. This configuration supports richer goal specifications like "stay under the umbrella" or "avoid hazardous areas."

MuJoCo Ant: We evaluate InfeRL in two variants of the Ant environment:

- *Ant-Balance*: The agent is instructed to rotate in place while maintaining balance.

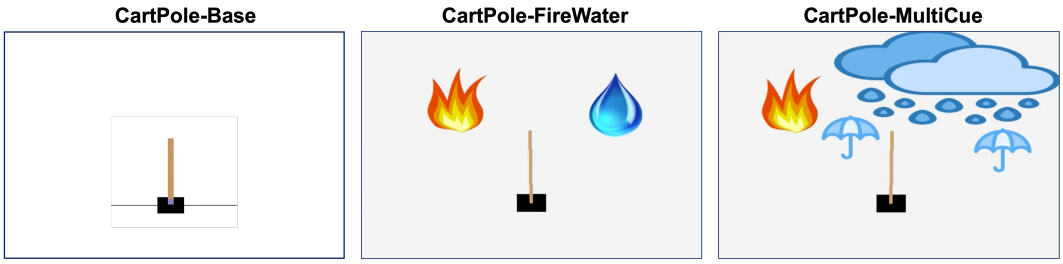

Figure 6: **Variants of the CartPole environment used in our experiments.** From left to right: (1) **CartPole-Base**: the standard task where the agent balances a pole on a cart. (2) **CartPole-FireWater**: background includes a fire icon on the left and a water droplet on the right, enabling directional prompts. (3) **CartPole-MultiCue**: background includes fire, water droplets, umbrellas, and clouds, supporting more complex instructions such as "stay under the umbrella" or "avoid fire." These variations preserve the original task dynamics while introducing symbolic visual cues that enable natural language instruction grounding. They are designed to evaluate the agent's ability to infer rewards from semantics, generalize across goal specifications, and follow increasingly abstract or context-dependent instructions.

- *Ant-Rotate*: A more specific task where the agent must spin in place without moving forward. This behavior deviates from the default locomotion behavior of Ant. Designing a handcrafted reward for such motion is non-trivial, but we demonstrate that a simple instruction such as "rotate" allows the agent to learn this behavior through reward inference (see Figure 4).

The Ant environment is significantly more complex than CartPole, with high-dimensional continuous action and observation spaces. **MuJoCo Walker2D**: We evaluate InfeRL in a variant of the Walker2D environment where the agent is instructed to walk while maintaining balance. Unlike standard Walker2D policies that are optimized for forward locomotion, our setup does not specify the direction of walking. As a result, the agent learns to walk backward—a valid solution under the instruction and inferred reward. This demonstrates that InfeRL can generalize to new behaviors without explicit motion direction cues, guided solely by language-based goal specification (see Figure 5).

These experiments serve to evaluate the scalability and generalization of InfeRL beyond simple control tasks.

These variations preserve the underlying task dynamics of environments while enabling more expressive, interpretable natural language goals. RGB renderings of the environment are used exclusively for computing inferred rewards via the vision–language model. The reinforcement learning agent itself receives only the standard low-dimensional state vector (e.g., position, velocity) from the CartPole environment during training.

While this framework can, in principle, be extended to visual RL settings where the agent also receives raw RGB observations, we deliberately adopt a *vector-based setup to isolate the impact of visual reward inference from the challenges of high-dimensional state learning*.

A.2 Goal Prompts

Each task is specified using a natural language prompt that captures the intended behavior. For example, in **CartPole-Base**, the prompt is: "The pole remains upright and the cart stays near the center." In **Ant-Rotate**, the prompt is: "A four-legged ant robot spins in place while staying balanced." We also explore variations of each prompt to evaluate robustness.

Table 1: Natural language goal specifications used for reward inference across different environment settings.

Environment	Setting	Task Type	Goal Specification
CartPole	Base	Single-objective	The pole is nearly vertical (upright), and the cart is near the center of the track.
CartPole	FireWater	Multi-objective	A cart with an upright pole is positioned directly under a red and yellow fire icon, far away from the blue water droplet.
CartPole	MultiCue	Multi-objective + Complex	A cart with an upright pole is positioned directly under a red and yellow fire icon, far away from the blue water droplet.
		Multi-objective + Complex	The pole is upright and stable, with both the cart and pole positioned under the right umbrella, far from the fire and out of the rain.
MuJoCo Ant	Balance	Single-objective	A four-legged robot walking and balanced.
MuJoCo Ant	Rotate	Novel behavior	A four-legged ant robot spinning rapidly in place, staying centered and balanced.
MuJoCo Walker2D	Walk	Single-objective + Ambiguous	A robot walking upright steadily.