SHADOW: Dynamic-Aware Credit Assignment Against Long-Horizon Tasks

Yuze Liu[1,2], **Chaochao Lu**[2], **Chao Yang**[2*]

[1]Zhejiang University
[2]Shanghai AI Laboratory
{liuyujie, luchaochao, yangchao}@pjlab.org.cn

Abstract

Reinforcement learning (RL) has emerged as the predominant paradigm for training large language model (LLM) agents to solve complex, multi-step tasks through environmental interaction. A fundamental challenge in such long-horizon scenarios is credit assignment, as delayed rewards provide inadequate signals for evaluating individual action contributions. Existing methods typically neglect trajectory transition dynamics, which leads to coarse-grained or biased credit assignment. To address these limitations, we introduce SHADOW, a novel framework that systematically incorporates transition dynamics for improved credit assignment. Our framework makes two primary contributions: (i) a dynamics-aware state grouping mechanism that mitigates misleading action comparisons between dynamically inconsistent states, and (ii) a local dynamic advantage estimator that leverages Generalized Advantage Estimation (GAE) to precisely quantify individual action contributions through a fine-grained analysis of transition patterns. Comprehensive experiments conducted with the Qwen2.5-1.5/7B-Instruct agent model demonstrate that our method achieves success rate improvements of 9.4%/7.6% on the *ALFworld* benchmark and a performance gain of over 5% on *WebShop*.

Code — https://github.com/yzliu7776/SHADOW

Introduction

Large Language Models (LLMs) (Team et al. 2023; Achiam et al. 2023; Guo et al. 2025; Liu et al. 2024; Hui et al. 2024; Bai et al. 2025; Jaech et al. 2024) have evolved into dynamic agents capable of sophisticated behaviors such as planning (Wang et al. 2023), tool use (Schick et al. 2023; Wang et al. 2025b), and sustained environmental interaction (Hong et al. 2024; Putta et al. 2024). Modern LLM agents autonomously navigate diverse interfaces including mobile (Gou et al. 2024; Wang et al. 2024a; Zhang and Zhang 2023; Hong et al. 2024; Hu et al. 2024), operating systems (Tan et al. 2024b; Xie et al. 2024; Zhang et al. 2024a), and web environments (Gur et al. 2023; Furuta et al. 2023; Zheng et al. 2024). They further excel at multi-step coding tasks (Zhang et al. 2024b) and control embodied systems in virtual worlds (Li et al. 2024). These advanced capabilities critically depend on the LLMs' ability to chain actions over extended horizons, a paradigm where reinforcement learning (RL) has proven indispensable. By optimizing for task completion through environmental feedback, RL aligns LLMs with goal-directed behaviors, underpinning cutting-edge AI systems (Wei et al. 2025; Zitkovich et al. 2023; Tan et al. 2024a; Zhai et al. 2024; Bai et al. 2024; Wang et al. 2024b).

However, RL's efficacy in long-horizon agent training is fundamentally constrained by the credit assignment problem (Feng et al. 2025b; Wen et al. 2024; Feng et al. 2025a; Kazemnejad et al. 2024; Wang et al. 2025a; Zheng et al. 2025; Zhang et al. 2025). Unlike single-turn tasks, agentic workflows span numerous steps with delayed rewards, where success signals often emerge only at episode termination. This sparsity obscures individual action contributions, forcing agents to learn from weak, globally attributed feedback. While group-based RL methods (Shao et al. 2024; Yu et al. 2025; Lin et al. 2025) mitigate this via trajectory-level advantage estimation, they lack stepwise granularity: actions are evaluated only within coarse rollout groups, neglecting their incremental progress toward task completion. Consequently, agents experience inefficient exploration, suboptimal grounding in environmental constraints, and diminished performance in complex interactive settings.

While Traditional Group-based RL (Shao et al. 2024) utilizes only trajectory indices for trajectory-level credit assignment, GiGPO (Feng et al. 2025b) further incorporates observational information to achieve step-level credit assignment by grouping states with identical observations (represented by hash values) and comparing actions stemming from these identical observations. However, a critical limitation of GiGPO lies in its disregard for transition dynamic information, which introduces two significant drawbacks: its observation-based state grouping can lead to misleading credit assignment signals due to comparisons of actions from dynamically inconsistent states under partial observability, and its action comparison metric, based solely on discounted episodic returns, results in coarse step-level credit assignment by neglecting local transition dynamic patterns.

These shortcomings underscore the necessity of integrating transition dynamics for enhanced accuracy and granu-

*Chao Yang is the corresponding author

larity in credit assignment, especially under partial observability. Our focus is therefore on exploiting the transition dynamic information within trajectory groups to generate more robust credit assignment signals. This approach allows for the identification of states with inconsistent dynamic behavior, even when they share identical observations; for instance, when the same action leads to divergent successor states or distinct rewards, or when the states themselves possess dissimilar admissible action spaces. Furthermore, leveraging local transition dynamic patterns provides a finer-grained action evaluation, as the state evaluation of consequent successor states offers more immediate and informative insight into action effects than relying solely on long-horizon episodic returns.

Based on these insights, we propose **SHADOW** (**S**tep-**H**onored **A**dvantage with **D**ynamic-aware **O**bservation **W**eight), a novel step-level, group-based credit assignment framework that provides finer-grained credit assignment through two synergistic components:

- A dynamic-aware state grouping mechanism that mitigates action comparisons between dynamically inconsistent states by utilizing a dynamic-aware observation weight matrix extracted from the transition dynamics of the trajectory group.
- A local dynamic advantage estimator that quantifies the contribution of a single action using Generalized Advantage Estimation (Schulman et al. 2015) for more precise, fine-grained credit assignment.

Experiments conducted with the Qwen2.5-1.5B/7B-Instruct LLMs demonstrate that SHADOW outperforms GiGPO and other baseline methods, achieving consistent performance gains 9.4% and 7.6% on the *ALFworld* benchmark, and an improvement of over 5% in *WebShop*.

Related Works

Reinforcement Learning for LLMs

RL has become an indispensable paradigm for enhancing LLMs, particularly for tasks demanding sophisticated reasoning and decision-making. Initial research primarily focused on aligning LLM capabilities with human preferences through techniques like Reinforcement Learning from Human Feedback (RLHF) (Ziegler et al. 2019; Stiennon et al. 2020; Ouyang et al. 2022), establishing a crucial foundation for behavior shaping. More recently, group-based RL algorithms have emerged as robust alternatives to conventional policy optimization strategies such as Proximal Policy Optimization (PPO) (Schulman et al. 2017). These collective learning approaches, including methods like RLOO (Kool, van Hoof, and Welling 2019; Ahmadian et al. 2024), GRPO (Shao et al. 2024), Dr. GRPO (Liu et al. 2025), DAPO (Yu et al. 2025), and CPPO (Lin et al. 2025), circumvent the need for explicit value functions. Instead, they estimate advantages by analyzing collections of samples from identical initial states, enabling efficient and scalable RL training for large models. This methodology has shown impressive empirical performance across diverse applications, notably in mathematical reasoning (Guo et al. 2025; Shao et al. 2024), leveraging search engines (Jin et al. 2025; Sun et al. 2025), and facilitating effective tool use (Qian et al. 2025; Wang et al. 2025b). Despite their efficacy, a persistent challenge for existing group-based RL techniques lies in their limited capacity to provide fine-grained credit assignment signals (Feng et al. 2025b), which is crucial for optimizing long-horizon and complex agent behaviors.

Credit Assignment for LLM Reinforcement Learning

Effective credit assignment mechanisms are increasingly vital for enhancing the sample efficiency of RL in LLMs, particularly when addressing long-horizon tasks. Existing credit assignment methods can be broadly categorized by their scope. Step-level credit assignment aims to provide fine-grained feedback for individual environmental steps, often deriving signals from hashable environmental states (Feng et al. 2025b), meta-reasoning types of the thinking process (Zhang et al. 2025), or learned progress reward models (Wang et al. 2025a), enabling precise optimization at each decision point. Chunk-level credit assignment segments long Chain-of-Thought (Wei et al. 2022) generations into smaller conceptual units or "chunks," utilizing Monte Carlo estimations to provide feedback at this intermediate granularity. Chunk delineation can be determined by rule-based heuristics (Kazemnejad et al. 2024) or by analyzing token-entropy patterns (Zheng et al. 2025). Lastly, token-level credit assignment focuses on modeling the specific contribution of each token to assign highly granular feedback. Such detailed feedback can be derived from the causal relationships of generated environmental actions (Feng et al. 2025a), through the analysis of token-entropy patterns (Wang et al. 2025c), or via group comparison (Wen et al. 2024). Our proposed method aligns with the principles of step-level credit assignment, specifically focusing on leveraging richer information from environmental transition dynamics against long-horizon tasks.

Dynamics-Aware Representation Learning in RL

Dynamics-aware representation learning leverages transition dynamic information from replay buffers to learn latent representations that facilitate state aggregation for downstream tasks. Common approaches include forward dynamics, which predict next states given current states and actions, and inverse dynamics, which predict actions from consecutive states (Pathak et al. 2017). These methods effectively partition the state space by distinguishing transition patterns, enabling meaningful implicit state grouping. Such dynamics-aware representations form the foundation for intrinsic reward computation (Pathak et al. 2017; Raileanu and Rocktäschel 2020; Badia et al. 2020; Guo et al. 2022; Henaff et al. 2022; Saade et al. 2023; Wan et al. 2023), providing principled measures of novelty and diversity through their state grouping properties (Pathak et al. 2017). Reward prediction, another dynamics-aware approach that estimates rewards from state-action pairs, learns state representations crucial for policy learning efficiency via their implicit state grouping properties (Schrittwieser et al. 2020; Hafner et al. 2019, 2020, 2023). Our SHADOW method employs

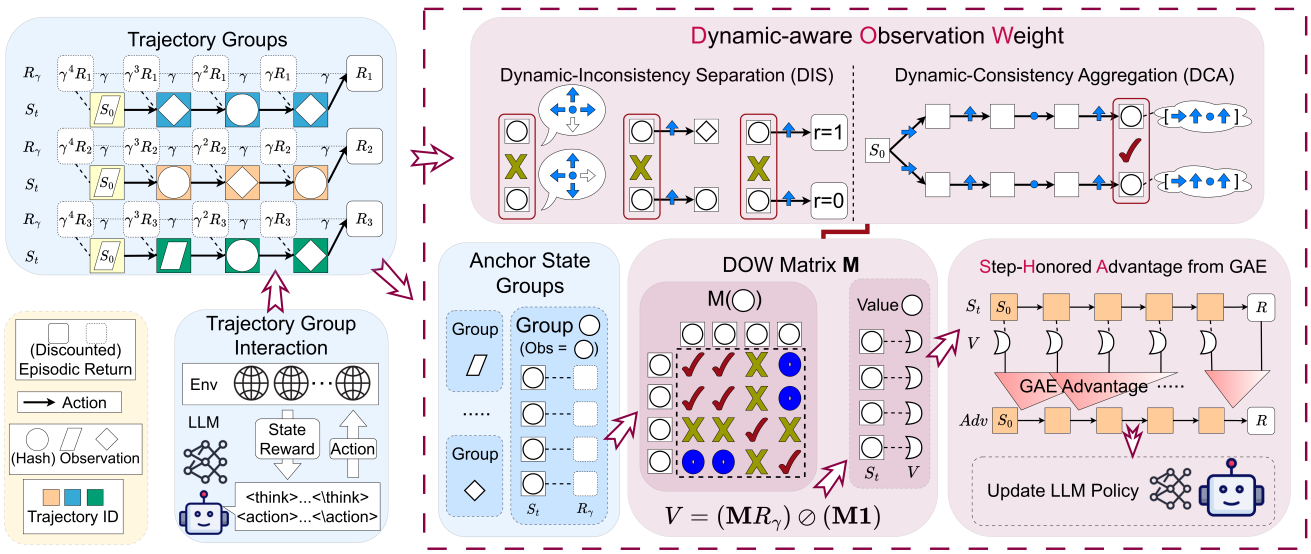

Figure 1: Overview of the SHADOW framework, which features a dynamic-aware state grouping mechanism and a local dynamic advantage estimator. The dynamic-aware state grouping mechanism employs a Dynamic-Aware Observation Weight (DOW) matrix, derived from trajectory group transition dynamics. This matrix implicitly performs Dynamic-Inconsistent Separation (DIS) for states with inconsistent dynamics (e.g., disparate admissible action spaces, forward dynamics, or reward dynamics) and Dynamic-Consistent Aggregation (DCA) for guaranteed identical states (e.g., those with identical action histories from identical initial states in deterministic environments). State value estimation are group-averaging calculated with the weight matrix, which are subsequently fed to a GAE advantage estimator for the final step-honored credit assignment signal.

this dynamics-aware paradigm to mitigate action comparisons between dynamic-inconsistent states for more accurate credit assignment.

Preliminaries

Problem Setup We focus on training LLM-based agents for long-horizon tasks. At each time step t, the agent observes a state s_t, generates a textual action a_t from its policy $\pi_\theta(a_t|s_t)$ (encompassing both the thinking process and the selected environmental action), receives a reward r_t, and transitions to s_{t+1}. A full episode forms a trajectory $\tau = \{(s_1, a_1, r_1), \ldots, (s_T, a_T, r_T)\}$.

Group-based Reinforcement Learning For a given task, LLM agents generate a group of N trajectories $G = \{\tau_1, \ldots, \tau_N\}$ under π_θ, each with a scalar return $R(\tau_i)$. Group-based RL methods, such as GRPO (Shao et al. 2024), compute the episode-level advantage from these group statistics:

$$Adv_{eps}(\tau_i, G) = GN(R(\tau_i), \{R(\tau_j)\}_{j=1}^{N}), \qquad (1)$$

where $GN(x, \mathcal{X}) = \frac{x-\mu(\mathcal{X})}{\sigma(\mathcal{X})}$ or $GN(x, \mathcal{X}) = x - \mu(\mathcal{X})$ is the group normalization operator.

Step-level Credit Assignment in Group-based RL This paper considers step-level credit assignment within the group-based RL setting. This involves calculating the advantage function for each generated textual action at each timestep, formulated as $Adv(a_t|s_t, G)$.

Group-in-Group Policy Optimization (GiGPO) GiGPO (Feng et al. 2025b) employs an anchor state grouping mechanism: states with identical observations are grouped together, and group-wise action comparisons are performed for step-level advantage:

$$Adv_{gigpo}(a_i|s_i, G) = GN(R_\gamma(s_i), \{R_\gamma(s)\}_{s \in G_{anc}(s_i)}), \qquad (2)$$

where $R_\gamma(\cdot)$ is the discounted episodic return, G is the trajectory group and $G_{anc}(s_i) = \{s \in G \mid s = s_i\}$ is the anchor state group.

Dynamic-aware Credit Assignment with SHADOW

Analysis of State Grouping

Despite its advancements, GiGPO's overlook of transition dynamic information introduces two critical limitations for fine-grained credit assignment. First, its observation-based state grouping mechanism can lead to misleading credit assignment signals. This occurs because states with identical observations may possess inconsistent underlying transition dynamics. In such cases, grouping and comparing actions based solely on observation identity can incorrectly assign negative advantage to an optimal action or positive advantage to a suboptimal one if it happens to be grouped with a state having a much higher true optimal return. For instance, if s_1 and s_2 have identical observations but the optimal return $R(s_1, a^*_{s_1}) > R(s_2, a^*_{s_2})$, comparing $R(s_1, a^*_{s_1})$ and $R(s_2, a^*_{s_2})$ might assign a negative advantage to $a^*_{s_2}$ even

Algorithm 1: DOW Matrix Calculation

Require: Trajectory group G with all n states indexed by position with function idx(), anchor state grouping function G_{anc}, admissible action space function $\mathcal{A}$, action history checker function H, weight hyperparameter $K > 0$.

Ensure: DOW Matrix $M \in \mathbb{R}^{n \times n}$

Initialize M:
- if $H(s_i) = H(s_j)$ set $M[i, j] = K + 1$
- else if $(G_{anc}(s_i) = G_{anc}(s_j)) \wedge (\mathcal{A}(s_i) = \mathcal{A}(s_j))$ set $M[i, j] = 1$
- otherwise set $M[i, j] = 0$

$M' \leftarrow M$
while True **do**
  for all (i, j) where $M'[i, j] = 1$ **do**
    Get transition (s_i, a_i, r_i, s_i') and (s_j, a_j, r_j, s_j') from G.
    if a_i=a_j **then**
      if $M[\text{idx}(s_i'), \text{idx}(s_j')] = 0 \vee r_i \neq r_j$ **then**
        for all (k, l) where $(M'[i, k] = K + 1) \wedge (M'[j, l] = K + 1)$ **do**
          $M[k, l] \leftarrow 0$
        end for
      end if
    end if
  end for
  if $M' \equiv M$ **then**
    Break
  end if
  $M' \leftarrow M$
end while
return M

if it is optimal for s_2. Second, GiGPO's reliance on the discounted episodic return, gained after a long horizon, as the sole metric for action evaluation results in coarse-grained advantages. This metric fails to disentangle the precise impact of individual actions within a sequence, thus limiting the granularity of credit assignment.

Dynamic-aware State Grouping Mechanism for Value Estimation

To mitigate misleading credit assignment signals arising from action comparisons between dynamically inconsistent states, SHADOW leverages available trajectory transition dynamic information to refine the grouping of observation-identical states. We formally define dynamically identical states through four fundamental properties, ensuring consistency in their behavioral outcomes:

1. **Identical Forward Dynamics:** Identical states must exhibit the same probabilistic transitions to next states given any identical action, i.e., $p(s'|s_1, a) = p(s'|s_2, a)$ for all a, s'. In deterministic environments, this simplifies to identical states transitioning to identical next states for any identical action.
2. **Identical Admissible Action Spaces:** Dynamically identical states possess the same set of permissible actions. This is a natural extension of the identical forward dynamics, ensuring that not only do actions lead to the same next states, but the set of *available* actions is also consistent.
3. **Identical Reward Dynamics:** Identical states must yield the same probabilistic single-step rewards given any identical action, i.e., $p(r|s_1, a) = p(r|s_2, a)$ for all a, r. In deterministic environments, this means an identical action from dynamically identical states results in an identical single-step reward.
4. **Guaranteed Identicality (Deterministic Environments):** For deterministic environments, states derived from performing identical action sequences from an identical initial state are guaranteed to be truly identical.

Building on these four properties, our framework performs two key operations on observation-identical states to refine their grouping. The first three properties (Identical Forward Dynamics, Identical Admissible Action Spaces, and Identical Reward Dynamics) are utilized for *Dynamic-Inconsistent Separation (DIS)* which identifies and separates state pairs exhibiting dynamically inconsistent behaviors by violating any of these three properties. The fourth property, *Guaranteed Identicality*, forms the basis for *Dynamic-Consistent Aggregation (DCA)*. Leveraging this strong guarantee in deterministic environments, we can further aggregate states that are confirmed to be identical based on their shared action history.

Given a trajectory group containing n states, initially partitioned into anchor state groups based on observation identity, we introduce a Dynamic-Aware Observation Weight (DOW) matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ to implicitly perform DIS and DCA. The DOW matrix $\mathbf{M}$ is initialized block-diagonal, with entries set to 1 for state pairs within the same anchor state group and 0 otherwise. Subsequently, DIS is applied by setting the weight of all dynamic-inconsistent state pairs to 0 within $\mathbf{M}$. Concurrently, DCA is applied by setting the weight of guaranteed identical state pairs to $K + 1$ (where $K > 0$), significantly increasing their influence. Since verifying identical forward dynamics involves identity information backpropagation, this process is performed iteratively, as summarized in Algorithm 1.

The value estimation for each state is then derived from a weighted average using the DOW matrix, which implicitly separates or further aggregates dynamic-inconsistent and guaranteed-identical states, respectively:

$$V(\mathbf{s}, G) = (\mathbf{M} R_\gamma(\mathbf{s})) \oslash (\mathbf{M}\mathbf{1}), \tag{3}$$

where $\oslash$ denotes element-wise division, $\mathbf{1}$ represents the unit vector, and $R_\gamma(\mathbf{s})$ is a vector of discounted episodic returns corresponding to each state in $\mathbf{s}$.

Step-Honored Advantage from GAE

To precisely attribute credit to individual steps, we adapt the Generalized Advantage Estimation (GAE) (Schulman et al. 2015). This adaptation extends GAE's conventional application from neural value approximation settings to our DOW-averaged value estimation, thus providing a refined advantage signal.

Type	Method	ALFWorld							WebShop	
		Pick	Look	Clean	Heat	Cool	Pick2	All	Score	Succ.
Base: Closed-Source Model										
Prompting	GPT-4o	75.3	60.8	31.2	56.7	21.6	49.8	48.0	31.8	23.7
Prompting	Gemini-2.5-Pro	92.8	63.3	62.1	69.0	26.6	58.7	60.3	42.5	35.9
Base: Qwen2.5-1.5B-Instruct										
Prompting	Qwen2.5	5.9	5.5	3.3	9.7	4.2	0.0	4.1	23.1	5.2
Prompting	ReAct	17.4	20.5	15.7	6.2	7.7	2.0	12.8	40.1	11.3
Prompting	Reflexion	35.3	22.2	21.7	13.6	19.4	3.7	21.8	55.8	21.9
RL Training	PPO (with critic)	$64.8_{\pm 3.5}$	$40.5_{\pm 6.9}$	$57.1_{\pm 4.9}$	$60.6_{\pm 6.6}$	$46.4_{\pm 4.0}$	$47.4_{\pm 1.9}$	$54.4_{\pm 3.1}$	$73.8_{\pm 3.0}$	$51.5_{\pm 2.9}$
RL Training	RLOO	$88.3_{\pm 3.0}$	$52.8_{\pm 8.6}$	$71.0_{\pm 5.9}$	$62.8_{\pm 8.7}$	$66.4_{\pm 5.5}$	$56.9_{\pm 4.7}$	$69.7_{\pm 2.5}$	$73.9_{\pm 5.6}$	$52.1_{\pm 6.7}$
RL Training	GRPO	$85.3_{\pm 1.5}$	$53.7_{\pm 8.0}$	$84.5_{\pm 6.8}$	$78.2_{\pm 7.9}$	$59.7_{\pm 5.0}$	$53.5_{\pm 5.6}$	$72.8_{\pm 3.6}$	$75.8_{\pm 3.5}$	$56.8_{\pm 3.8}$
RL Training	GiGPO	$98.9_{\pm 1.6}$	$77.8_{\pm 20.8}$	$79.7_{\pm 8.5}$	$80.6_{\pm 6.4}$	$\mathbf{90.1_{\pm 7.1}}$	$83.6_{\pm 5.1}$	$86.7_{\pm 5.1}$	$83.9_{\pm 2.1}$	$71.4_{\pm 1.8}$
RL Training	SHADOW (Ours)	$\mathbf{99.0}_{\pm 1.4}$	$\mathbf{90.9}_{\pm 7.4}$	$\mathbf{99.0}_{\pm 1.5}$	$\mathbf{100.0}_{\pm 0.0}$	$88.5_{\pm 4.7}$	$\mathbf{94.7}_{\pm 0.4}$	$\mathbf{96.1}_{\pm 0.7}$	$\mathbf{88.1}_{\pm 1.1}$	$\mathbf{76.6}_{\pm 0.6}$
Base: Qwen2.5-7B-Instruct										
Prompting	Qwen2.5	33.4	21.6	19.3	6.9	2.8	3.2	14.8	26.4	7.8
Prompting	ReAct	48.5	35.4	34.3	13.2	18.2	17.6	31.2	46.2	19.5
Prompting	Reflexion	62.0	41.6	44.9	30.9	36.3	23.8	42.7	58.1	28.8
RL Training	PPO (with critic)	$92.3_{\pm 4.0}$	$64.0_{\pm 8.4}$	$92.5_{\pm 2.4}$	$89.5_{\pm 7.0}$	$80.3_{\pm 2.0}$	$68.8_{\pm 8.3}$	$80.4_{\pm 2.7}$	$81.4_{\pm 3.1}$	$68.7_{\pm 5.1}$
RL Training	RLOO	$87.6_{\pm 4.3}$	$78.2_{\pm 8.3}$	$87.3_{\pm 5.8}$	$81.3_{\pm 7.6}$	$71.9_{\pm 5.2}$	$48.9_{\pm 8.4}$	$75.5_{\pm 4.6}$	$80.3_{\pm 3.2}$	$65.7_{\pm 4.0}$
RL Training	GRPO	$90.8_{\pm 5.1}$	$66.1_{\pm 6.7}$	$89.3_{\pm 5.4}$	$74.7_{\pm 6.9}$	$72.5_{\pm 5.4}$	$64.7_{\pm 7.3}$	$77.6_{\pm 5.2}$	$79.3_{\pm 2.8}$	$66.1_{\pm 3.7}$
RL Training	GiGPO	$97.8_{\pm 3.0}$	$89.5_{\pm 8.7}$	$86.2_{\pm 7.3}$	$88.6_{\pm 5.7}$	$\mathbf{85.6}_{\pm 5.9}$	$88.6_{\pm 2.6}$	$90.1_{\pm 1.3}$	$86.5_{\pm 3.7}$	$76.3_{\pm 6.5}$
RL Training	SHADOW (Ours)	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{85.6}_{\pm 10.6}$	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{97.7}_{\pm 1.7}$	$\mathbf{89.1}_{\pm 2.5}$	$\mathbf{81.5}_{\pm 2.9}$

Table 1: Performance on ALFWorld and WebShop. For ALFWorld, the metric is the average success rate (%) for each subtask and the overall result. For WebShop, the metric is the average environmental score and the average success rate (%). Results are averaged over 3 random seeds.

Consider a trajectory $\tau = \{(s_i, a_i, r_i)\}_{i=0}^{N}$ with corresponding state values $\{V(s_i, G)\}_{i=0}^{N}$ obtained from our DOW matrix. We set $V(s_{N+1}, G) = 0$ as a termination condition. The advantage for an action a_i taken at state s_i within trajectory τ is computed as:

$$Adv_{shadow}(a_i|s_i, \tau) = \sum_{k=i}^{N} (\lambda\gamma)^{k-i}\delta_k, \tag{4}$$

where $\delta_k = r_k + \gamma V(s_{k+1}, G) - V(s_k, G)$ represents the temporal difference (TD) error at step k. Here, $\lambda \in [0, 1]$ is a hyperparameter of GAE that continuously interpolates between the local step credit assignment ($\lambda \to 0$) and the trajectory-level credit assignment ($\lambda = 1$).

GiGPO: A Simplified Instance of SHADOW

Our framework integrates the Dynamics-Aware State Grouping with GAE to derive step-level advantages. Beyond this fine-grained step-level credit assignment, SHADOW also incorporates episode-level advantages (i.e., $Adv = Adv_{shadow} + Adv_{eps}$), analogous to GiGPO, to encourage long-range coherence.

The relationship between SHADOW and GiGPO can be understood as follows: GiGPO can be regarded as a simplified instance of SHADOW. Specifically, when the GAE hyperparameter λ is set to 1, and the Dynamic-Aware Observation Weight (DOW) matrix is exclusively initialized in a block-diagonal fashion based on observation identity without further updates from the Dynamic-Inconsistent Separation (DIS) and Dynamic-Consistent Aggregation (DCA) objectives, SHADOW becomes equivalent to GiGPO. This highlights that SHADOW generalizes GiGPO by explicitly exploiting transition dynamic information for improved credit assignment.

Experiments

Experiment Settings

Benchmarks Our LLM agents were rigorously trained and evaluated on two widely recognized and complex long-horizon benchmarks: ALFWorld (Shridhar et al. 2020) and WebShop (Yao et al. 2022). ALFWorld provides an embodied text-based setting for multi-step task planning in simulated household environments across 6 categories of household tasks: Pick & Place (Pick), Examine in Light (Look), Clean & Place (Clean), Heat & Place (Heat), Cool & Place (Cool), and Pick Two & Place (Pick2). WebShop presents a web-based environment from realistic online shopping scenarios where the objective is to search, navigate, and purchase specified items. Crucially, both ALFWorld and WebShop are partially observable environments, where their text-based observations are inherently insufficient to convey global state information.

Baselines For comprehensive performance context, we compared SHADOW against competitive baselines. Our analysis primarily focused on GiGPO (Feng et al. 2025b), and also included closed-source LLMs (GPT-4o (Achiam et al. 2023), Gemini-2.5-Pro (Team et al. 2023)), prompting agents (ReAct (Yao et al. 2023), Reflexion (Shinn et al. 2023)), traditional RL (PPO (Schulman et al. 2017)), and other group-based approaches (RLOO (Kool, van Hoof, and Welling 2019), GRPO (Shao et al. 2024)).

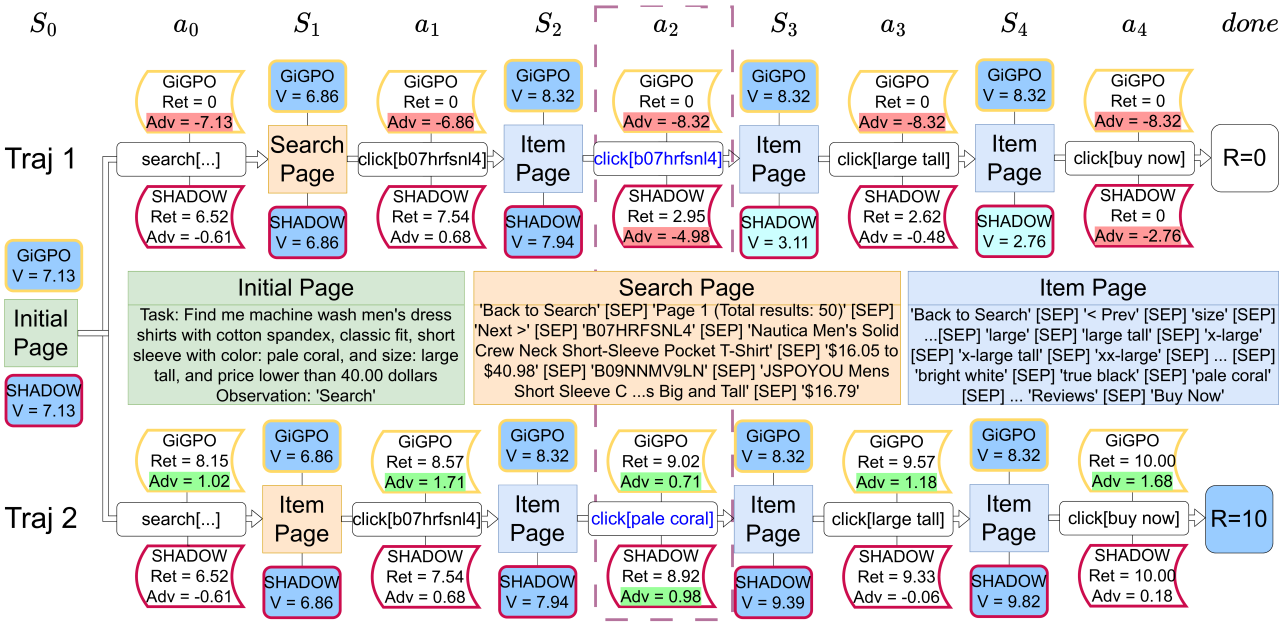

Figure 2: A case analysis illustrating the advantage signals from SHADOW (magenta thick-edged boxes) and GiGPO (yellow thick-edged boxes) within a WebShop trajectory group. GiGPO assigns a uniform positive/negative advantage value to all actions within successful/failing trajectories, respectively. In contrast, SHADOW provides finer-grained credit assignment, specifically focusing on attributing credit or blame to the individual actions that directly contribute to success or failure. States marked with the same page indicator share identical observations.

Training Setup We utilized Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct (Hui et al. 2024) as the base LLMs instructed to generate a chain-of-thought with the `<think> </think>` and `<action> </action>` format. The environmental reward function for RL training comprised a sparse task success reward at episode termination, complemented by a penalty for illegal output formats. For SHADOW, we set the GAE-λ to 0.5 for ALFWorld and 0 for WebShop, and the DCA weight K to 6 for ALFWorld and 30 for WebShop. More granular specifics concerning the full training settings and hyperparameter details are comprehensively documented in the source code.

Experiment Performance

Our empirical evaluation, summarized in Table 1, unequivocally demonstrates the superior performance of our proposed SHADOW across both the ALFWorld and WebShop benchmarks. The results reveal a decisive advantage for SHADOW over all baselines, particularly in comparison to GiGPO (Feng et al. 2025b), which serves as a key structural baseline for this work.

In the ALFWorld environment, SHADOW consistently achieved near-optimal performance with Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct LLMs, exhibiting overall success rates of 96.1% and 97.7%, respectively. These figures significantly surpass those of GiGPO (86.7% and 90.1%) and other baselines. Across the six subtasks, SHADOW outperformed GiGPO in "Pick", "Look", "Clean", "Heat", and "Pick2", while maintaining competent performance in the "Cool" task. Furthermore, SHADOW, utilizing the 1.5B model, achieved a perfect success rate in "Pick", "Look", "Clean", "Heat", and "Pick2", with SHADOW leveraging the 7B model achieving a 100% success rate across all subtasks except "Cool", observed consistently over three different random seeds. Similarly, on the WebShop benchmark, SHADOW delivered remarkable performance, outperforming GiGPO and all other baselines in terms of both average score and overall success rate. Specifically, SHADOW achieved a 5% performance gain over GiGPO for both the 7B and 1.5B models, alongside significantly outperforming other baselines. These results underscore the effectiveness of SHADOW in managing long-horizon tasks through precise and fine-grained credit assignment, which is attributed to its modeling of transition dynamic information.

Case Analysis

Figure 2 illustrates an example of the credit assignment signal derived from a WebShop trajectory group. The objective in this task is to complete an online web shopping process, which involves searching for items, selecting specific attributes, and finalizing the purchase. Successful task completion hinges on clicking the "buy now" button only after all required attributes have been selected; premature clicking results in episode termination without reward. In this particular trajectory group, Trajectory 2 successfully completed the task with the correct sequence of actions, whereas Trajectory 1 performed an invalid action at step 3, with all

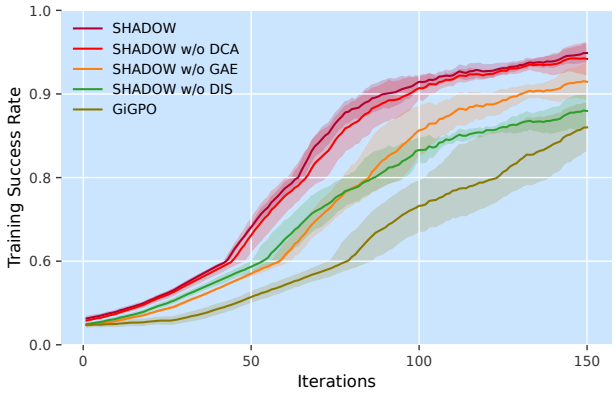

Figure 3: Learning curves of the SHADOW component ablations on Qwen2.5-1.5B-Instruct and ALFWorld. Results are averaged over 3 random seeds.

other actions identical to Trajectory 2. The group also includes an additional six successful trajectories that mirrored Trajectory 2.

Considering these two trajectories, states $S_{4,\tau 1}$ and $S_{4,\tau 2}$ are classified as non-identical due to inconsistent reward dynamics, meaning that identical actions led to disparate rewards. Consequently, states $S_{3,\tau 1}$ and $S_{3,\tau 2}$ are also deemed non-identical due to inconsistent forward dynamics, as the same action yielded non-identical subsequent states ($S_{4,\tau 1}$ and $S_{4,\tau 2}$). In contrast, states $S_{2,\tau 1}$ and $S_{2,\tau 2}$ are guaranteed to be identical, a consequence of identical action histories from identical initial states within this deterministic setting. Action a_2 for both trajectories effectively bifurcates their paths and is, unbiasedly, responsible for the respective success or failure outcomes, given only the transition dynamic information available within the trajectory group. A key distinction arises when comparing SHADOW to GiGPO. GiGPO's advantage computation equally penalizes all actions of the failed Trajectory 1, including the initial two actions that were identical to those in the successful Trajectory 2 and originated from identical states. This can introduce noise into the training process and potentially provide misleading guidance for chain-of-thought (CoT) generation policies due to varying advantages for identical actions from identical states. In contrast, SHADOW's advantage correctly focuses on a_2 for both trajectories. It is also important to note that SHADOW assigns a penalty to $a_{4,\tau 1}$, which coincidentally represents an action leading to irreversible failure within the WebShop domain, given the domain-dependent prior knowledge transition dynamics. This is because SHADOW leverages identical observation information for value estimation when observation pairs exhibit no dynamic-inconsistent behaviors, thereby benefiting in the WebShop setting, albeit with a relatively cautious scaling due to its domain-dependent nature.

Ablation Study

In this subsection, we conduct an ablation study of the three key components of SHADOW using Qwen2.5-1.5B-Instruct on the ALFWorld benchmark: Dynamic-Inconsistent Separation (DIS), Dynamic-Consistent Aggregation (DCA), and the Generalized Advantage Estimation (GAE) component. The results, presented in Table 2 and Figure 3, indicate distinct contributions from each. The GAE component significantly accelerates convergence speed during the early training stages, with a relatively small influence on the final success rates. This suggests that the finer-grained credit assignment signals provided by GAE primarily serve to expedite the learning process. Conversely, the Dynamic-Inconsistency Separation (DIS) component emerges as the dominant factor for the observed boost in final success rates. This highlights that the precise credit assignment signals derived from separating dynamic-inconsistent states for action comparison improve the upper-bound of the agent's performance. Finally, the Dynamic-Consistency Aggregation (DCA) component yields a slight improvement in the final success rate. This minor impact may be attributed to the relatively rare occurrence of truly identical states with identical action sequences within the ALFWorld environments.

Method	DCA	GAE	DIS	Succ % (ALL)
SHADOW (full)	✓	✓	✓	$96.1_{\pm 0.7}$
SHADOW w/o DCA	×	✓	✓	$95.1_{\pm 2.1}$
SHADOW w/o GAE	✓	×	✓	$93.5_{\pm 1.6}$
SHADOW w/o DIS	✓	✓	×	$89.8_{\pm 1.9}$
GiGPO	×	×	×	$86.7_{\pm 6.2}$

Table 2: Component Ablations of SHADOW on Qwen2.5-1.5B-Instruct and ALFWorld, where the overall success rate % is reported. Results are averaged over 3 random seeds.

Conclusion

We present SHADOW, a novel framework that systematically incorporates transition dynamics for improved credit assignment in long-horizon LLM agent training. Our key contributions include a dynamics-aware state grouping mechanism that mitigates misleading action comparisons between dynamically inconsistent states for precise credit assignment and a local dynamic advantage estimator that leverages Generalized Advantage Estimation (GAE) for finer-grained credit assignment. Extensive experiments conducted with Qwen2.5-1.5/7B-Instruct LLMs demonstrated consistent and substantial improvements in performance on challenging long-horizon benchmarks, achieving success rate improvements of 9.4%/7.6% in ALFWorld and a 5% performance gain in WebShop.

While our method effectively addresses critical credit assignment challenges, it inherits the requirement of GiGPO for noise-free hashable observation representations. Furthermore, our current framework introduces an additional constraint: deterministic environment dynamics are assumed for our DOW matrix computation. Future work will explore extending SHADOW to stochastic environments by incorporating probabilistic modeling of state transition equivalences, thereby broadening the applicability of our framework.

Acknowledgements

We sincerely thank Professor Fei Wu for his invaluable guidance and support throughout this research. And this research is funded by the Shanghai Artificial Intelligence Laboratory. In addition, it was partially supported by the National Key R&D Program of China (NO.2022ZD0160102).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahmadian, A.; Cremer, C.; Gallé, M.; Fadaee, M.; Kreutzer, J.; Pietquin, O.; Üstün, A.; and Hooker, S. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Badia, A. P.; Sprechmann, P.; Vitvitskyi, A.; Guo, D.; Piot, B.; Kapturowski, S.; Tieleman, O.; Arjovsky, M.; Pritzel, A.; Bolt, A.; et al. 2020. Never give up: Learning directed exploration strategies. *arXiv preprint arXiv:2002.06038*.

Bai, H.; Zhou, Y.; Pan, J.; Cemri, M.; Suhr, A.; Levine, S.; and Kumar, A. 2024. Digirl: Training in-the-wild device-control agents with autonomous reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 12461–12495.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Feng, L.; Tan, W.; Lyu, Z.; Zheng, L.; Xu, H.; Yan, M.; Huang, F.; and An, B. 2025a. Towards Efficient Online Tuning of VLM Agents via Counterfactual Soft Reinforcement Learning. *arXiv preprint arXiv:2505.03792*.

Feng, L.; Xue, Z.; Liu, T.; and An, B. 2025b. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*.

Furuta, H.; Lee, K.-H.; Nachum, O.; Matsuo, Y.; Faust, A.; Gu, S. S.; and Gur, I. 2023. Multimodal web navigation with instruction-finetuned foundation models. *arXiv preprint arXiv:2305.11854*.

Gou, B.; Wang, R.; Zheng, B.; Xie, Y.; Chang, C.; Shu, Y.; Sun, H.; and Su, Y. 2024. Navigating the digital world as humans do: Universal visual grounding for gui agents. *arXiv preprint arXiv:2410.05243*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Guo, Z.; Thakoor, S.; Pîslar, M.; Avila Pires, B.; Altché, F.; Tallec, C.; Saade, A.; Calandriello, D.; Grill, J.-B.; Tang, Y.; et al. 2022. Byol-explore: Exploration by bootstrapped prediction. *Advances in neural information processing systems*, 35: 31855–31870.

Gur, I.; Furuta, H.; Huang, A.; Safdari, M.; Matsuo, Y.; Eck, D.; and Faust, A. 2023. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Hafner, D.; Lillicrap, T.; Norouzi, M.; and Ba, J. 2020. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*.

Hafner, D.; Pasukonis, J.; Ba, J.; and Lillicrap, T. 2023. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.

Henaff, M.; Raileanu, R.; Jiang, M.; and Rocktäschel, T. 2022. Exploration via elliptical episodic bonuses. *Advances in Neural Information Processing Systems*, 35: 37631–37646.

Hong, W.; Wang, W.; Lv, Q.; Xu, J.; Yu, W.; Ji, J.; Wang, Y.; Wang, Z.; Dong, Y.; Ding, M.; et al. 2024. Cogagent: A visual language model for gui agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14281–14290.

Hu, S.; Ouyang, M.; Gao, D.; and Shou, M. Z. 2024. The dawn of gui agent: A preliminary case study with claude 3.5 computer use. *arXiv preprint arXiv:2411.10323*.

Hui, B.; Yang, J.; Cui, Z.; Yang, J.; Liu, D.; Zhang, L.; Liu, T.; Zhang, J.; Yu, B.; Lu, K.; et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jin, B.; Zeng, H.; Yue, Z.; Yoon, J.; Arik, S.; Wang, D.; Zamani, H.; and Han, J. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Kazemnejad, A.; Aghajohari, M.; Portelance, E.; Sordoni, A.; Reddy, S.; Courville, A.; and Le Roux, N. 2024. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment.

Kool, W.; van Hoof, H.; and Welling, M. 2019. Buy 4 reinforce samples, get a baseline for free!

Li, M.; Zhao, S.; Wang, Q.; Wang, K.; Zhou, Y.; Srivastava, S.; Gokmen, C.; Lee, T.; Li, E. L.; Zhang, R.; et al. 2024. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37: 100428–100534.

Lin, Z.; Lin, M.; Xie, Y.; and Ji, R. 2025. Cppo: Accelerating the training of group relative policy optimization-based reasoning models. *arXiv preprint arXiv:2503.22342*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, Z.; Chen, C.; Li, W.; Qi, P.; Pang, T.; Du, C.; Lee, W. S.; and Lin, M. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions

with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, 2778–2787. PMLR.

Putta, P.; Mills, E.; Garg, N.; Motwani, S.; Finn, C.; Garg, D.; and Rafailov, R. 2024. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv preprint arXiv:2408.07199*.

Qian, C.; Acikgoz, E. C.; He, Q.; Wang, H.; Chen, X.; Hakkani-Tür, D.; Tur, G.; and Ji, H. 2025. Toolrl: Reward is all tool learning needs. *arXiv preprint arXiv:2504.13958*.

Raileanu, R.; and Rocktäschel, T. 2020. Ride: Rewarding impact-driven exploration for procedurally-generated environments. *arXiv preprint arXiv:2002.12292*.

Saade, A.; Kapturowski, S.; Calandriello, D.; Blundell, C.; Sprechmann, P.; Sarra, L.; Groth, O.; Valko, M.; and Piot, B. 2023. Unlocking the power of representations in long-term novelty-based exploration. *arXiv preprint arXiv:2305.01521*.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551.

Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; et al. 2020. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839): 604–609.

Schulman, J.; Moritz, P.; Levine, S.; Jordan, M. I.; and Abbeel, P. 2015. High-Dimensional Continuous Control Using Generalized Advantage Estimation. *CoRR*, abs/1506.02438.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 8634–8652.

Shridhar, M.; Yuan, X.; Côté, M.-A.; Bisk, Y.; Trischler, A.; and Hausknecht, M. 2020. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33: 3008–3021.

Sun, H.; Qiao, Z.; Guo, J.; Fan, X.; Hou, Y.; Jiang, Y.; Xie, P.; Zhang, Y.; Huang, F.; and Zhou, J. 2025. Zerosearch: Incentivize the search capability of llms without searching. *arXiv preprint arXiv:2505.04588*.

Tan, W.; Zhang, W.; Liu, S.; Zheng, L.; Wang, X.; and An, B. 2024a. True knowledge comes from practice: Aligning llms with embodied environments via reinforcement learning. *arXiv preprint arXiv:2401.14151*.

Tan, W.; Zhang, W.; Xu, X.; Xia, H.; Ding, Z.; Li, B.; Zhou, B.; Yue, J.; Jiang, J.; Li, Y.; et al. 2024b. Cradle: Empowering foundation agents towards general computer control. *arXiv preprint arXiv:2403.03186*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wan, S.; Tang, Y.; Tian, Y.; and Kaneko, T. 2023. Deir: efficient and robust exploration through discriminative-model-based episodic intrinsic rewards. *arXiv preprint arXiv:2304.10770*.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Wang, H.; Leong, C. T.; Wang, J.; Wang, J.; and Li, W. 2025a. SPA-RL: Reinforcing LLM Agents via Stepwise Progress Attribution. *arXiv preprint arXiv:2505.20732*.

Wang, H.; Qian, C.; Zhong, W.; Chen, X.; Qiu, J.; Huang, S.; Jin, B.; Wang, M.; Wong, K.-F.; and Ji, H. 2025b. Otc: Optimal tool calls via reinforcement learning. *arXiv e-prints*, arXiv–2504.

Wang, J.; Xu, H.; Jia, H.; Zhang, X.; Yan, M.; Shen, W.; Zhang, J.; Huang, F.; and Sang, J. 2024a. Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration. *Advances in Neural Information Processing Systems*, 37: 2686–2710.

Wang, S.; Yu, L.; Gao, C.; Zheng, C.; Liu, S.; Lu, R.; Dang, K.; Chen, X.; Yang, J.; Zhang, Z.; et al. 2025c. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Wang, T.; Wu, Z.; Liu, J.; Hao, J.; Wang, J.; and Shao, K. 2024b. Distrl: An asynchronous distributed reinforcement learning framework for on-device control agents. *arXiv preprint arXiv:2410.14803*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wei, Y.; Duchenne, O.; Copet, J.; Carbonneaux, Q.; Zhang, L.; Fried, D.; Synnaeve, G.; Singh, R.; and Wang, S. I. 2025. Swe-rl: Advancing llm reasoning via reinforcement learning on open software evolution. *arXiv preprint arXiv:2502.18449*.

Wen, M.; Wan, Z.; Wang, J.; Zhang, W.; and Wen, Y. 2024. Reinforcing llm agents via policy optimization with action decomposition. *Advances in Neural Information Processing Systems*, 37: 103774–103805.

Xie, T.; Zhang, D.; Chen, J.; Li, X.; Zhao, S.; Cao, R.; Hua, T. J.; Cheng, Z.; Shin, D.; Lei, F.; et al. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37: 52040–52094.

Yao, S.; Chen, H.; Yang, J.; and Narasimhan, K. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35: 20744–20757.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Dai, W.; Fan, T.; Liu, G.; Liu, L.; et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Zhai, S.; Bai, H.; Lin, Z.; Pan, J.; Tong, P.; Zhou, Y.; Suhr, A.; Xie, S.; LeCun, Y.; Ma, Y.; et al. 2024. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. *Advances in neural information processing systems*, 37: 110935–110971.

Zhang, C.; Li, L.; He, S.; Zhang, X.; Qiao, B.; Qin, S.; Ma, M.; Kang, Y.; Lin, Q.; Rajmohan, S.; et al. 2024a. Ufo: A ui-focused agent for windows os interaction. *arXiv preprint arXiv:2402.07939*.

Zhang, K.; Li, J.; Li, G.; Shi, X.; and Jin, Z. 2024b. Codeagent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. *arXiv preprint arXiv:2401.07339*.

Zhang, Z.; Chen, Z.; Li, M.; Tu, Z.; and Li, X. 2025. RLVMR: Reinforcement Learning with Verifiable Meta-Reasoning Rewards for Robust Long-Horizon Agents. arXiv:2507.22844.

Zhang, Z.; and Zhang, A. 2023. You only look at screens: Multimodal chain-of-action agents. *arXiv preprint arXiv:2309.11436*.

Zheng, B.; Gou, B.; Kil, J.; Sun, H.; and Su, Y. 2024. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*.

Zheng, T.; Xing, T.; Gu, Q.; Liang, T.; Qu, X.; Zhou, X.; Li, Y.; Wen, Z.; Lin, C.; Huang, W.; et al. 2025. First Return, Entropy-Eliciting Explore. *arXiv preprint arXiv:2507.07017*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; Amodei, D.; Christiano, P.; and Irving, G. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Zitkovich, B.; Yu, T.; Xu, S.; Xu, P.; Xiao, T.; Xia, F.; Wu, J.; Wohlhart, P.; Welker, S.; Wahid, A.; et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2165–2183. PMLR.