SAME OBSERVATION, DIFFERENT FUTURES: CREDIT ASSIGNMENT VIA DYNAMICS MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents are moving from single-turn answers toward long-horizon workflows spanning dozens of decisions and tool calls. As the interaction horizon grows, a sparse terminal reward must be attributed across increasingly many decisions, progressively weakening the supervision available for each individual decision. Credit assignment therefore becomes a key bottleneck in agentic RL. A common approach is to estimate intermediate history values and use them to construct step-level advantages. Existing methods group histories using heuristic local matching and estimate values from within-group Monte Carlo (MC) returns. This can introduce bias by grouping histories with different continuation dynamics and increase variance by ignoring trajectories outside the group. In this work, we propose Credit Assignment via global DYnamics (CADY), a structured credit assignment method that models global trajectory dynamics. CADY maps histories with similar continuation dynamics to similar posterior distributions over latent states, mitigating the bias induced by heuristic grouping. It further exploits the global coupling among latent states for structured value estimation. This allows information from all rollouts to be more effectively leveraged through the shared transition dynamics. We further provide a theoretical analysis explaining why global dynamics-based value estimation can achieve lower first-order asymptotic variance than a simple MC estimator. Experiments on three challenging long-horizon agentic benchmarks show that CADY produces more accurate history-value estimates and substantially higher task success rates. Even in a tool-use environment with roughly 500 API operations, CADY can still separate locally indistinguishable histories with divergent continuations, demonstrating its scaling potential.

1 INTRODUCTION

Large language models (LLMs) are evolving from single-turn question-answering systems (Ouyang et al., 2022) into *agents capable of long-horizon, multi-turn interaction with external environments* (Xi et al., 2025). Such interaction capabilities substantially broaden the range of real-world applications of LLMs, including household manipulation in virtual embodied environments (Shridhar et al., 2020; Bai et al., 2024), multi-hop web search (Yao et al., 2022; Furuta et al., 2024; Jin et al., 2025; OpenAI, 2025), and computer-terminal operations (Marten et al., 2026; Huang & Jiang, 2026). These tasks typically involve long interaction horizons and sparse terminal rewards, making it difficult for reinforcement learning (RL) algorithms to accurately assess the value of intermediate decision histories (Yuan et al., 2026; Xu et al., 2025; Zhang, 2026; Fu et al., 2026). Poor credit assignment leads to noisy advantage estimates, substantially reducing the statistical efficiency of policy optimization. Therefore, credit assignment is a key challenge in long-horizon agentic RL.

Existing methods estimate intermediate history values in different ways. Actor–critic approaches (Schulman et al., 2017; Li et al., 2026a; Zhou et al., 2025; Yue et al., 2025; Hou et al., 2026) train a critic model end-to-end. Since this model is often comparable in size to the actor, it introduces substantial memory and computational overhead. Other methods (Tan et al., 2026; Kang et al., 2026; Wang et al., 2026; Tao et al., 2026; Li et al., 2026b) instead leverage internal model signals such as reflection or confidence, but depend on calibration and self-evaluation quality. A more common critic-free paradigm uses Monte Carlo estimates from locally matched trajectory groups. For example, GRPO (Shao et al., 2024) uses the mean reward of responses as a shared value baseline and subtracts it from the observed reward to form token-level advantages. Consequently, it cannot

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

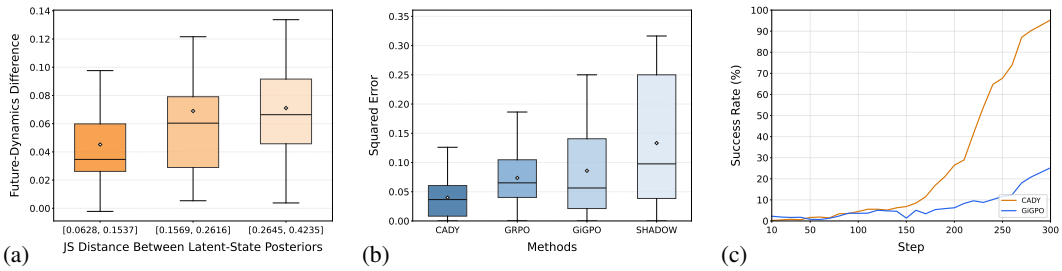


Figure 1: (a) Relationship between future-trajectory dynamics differences and latent-state posterior differences for history pairs sharing the same current action and observation. (b) Comparison of squared-error distributions for history-value estimation across CADY and three baselines. (c) Validation performance of CADY and GiGPO during training on CookingWorld.

distinguish the relative quality of decisions within the same trajectory. GiGPO (Feng et al., 2026) enables finer step-level credit by grouping histories with similar latest environmental feedback, while SHADOW (Liu et al., 2026) incorporates environmental-state consistency into the grouping process. However, these methods all rely on *local matching rules* for history grouping and value estimation. Specifically, heuristic criteria such as matching environmental feedback or local transitions may place histories with different continuation dynamics into the same group. Such histories can have different true values, yet are assigned the same group-level value estimate, introducing estimation bias. Moreover, the MC value of a target history is estimated only from trajectories containing histories in the corresponding group, discarding potentially useful information from other trajectories. Since history values are coupled through global trajectory dynamics, this local use of rollout data can also lead to higher estimation variance.

These observations suggest that the fundamental limitation of local matching lies in performing grouping and value estimation without accounting for the *global dynamics* on which history values depend, ultimately leading to larger value-estimation errors. To this end, we propose **CADY (Credit Assignment via global DYnamics)**, a structured credit assignment method that explicitly models global trajectory dynamics. CADY maps histories with similar continuation behavior to similar posterior distributions over latent states (as shown in Figure 1a). This representation enables more value-relevant grouping and mitigates the bias introduced by heuristic grouping. Moreover, CADY estimates continuation values by computing the probability of task success under the learned global trajectory dynamics. This allows trajectories that do not directly visit a target history to still contribute to its value estimate through shared transition structure. We further provide a theoretical analysis explaining why global-dynamics-based value estimation can achieve lower variance than simple Monte Carlo estimators. To instantiate this idea efficiently, CADY employs a lightweight Cloned Hidden Markov Model (CHMM) (Dedieu et al., 2019), whose structure explicitly represents both predictive latent states and their transition matrix. When optimizing the policy, we combine complementary episode-level and step-level advantages derived from CHMM-based value estimates, capturing both overall success and the contribution of individual actions.

We evaluate CADY with **Qwen2.5-1.5B-Instruct** and **Qwen2.5-7B-Instruct** on three challenging long-horizon agentic benchmarks: CookingWorld, TextWorld Commonsense, and multi-hop QA. Experiments show that CADY achieves the lowest mean squared error in value estimation (as shown in Figure 1b) and substantially improves downstream task accuracy compared with competitive baselines (as shown in Figure 1c). Fitting experiments on AutomationBench further demonstrate CADY's potential to scale to more complex and realistic settings (as shown in Figures 4a and 4b).

2 PRELIMINARIES

Problem Setup. We focus on training LLM-based agents for long-horizon tasks with sparse terminal rewards. At each time step t, given a task description $x \sim p(X)$ and the interactive history h_t, the agent generates a textual action a_t according to its policy $\pi_\theta(a_t \mid h_t, x)$. Typically, the action consists of an intermediate reasoning process ρ_t followed by an executable action u_t that interacts with the environment: $a_t = (\rho_t, u_t)$. After executing u_t, the environment transitions to the next

history state h_{t+1}. At the end of the episode, the environment returns a binary terminal reward $r_T \in \{0, 1\}$, indicating whether the task is successfully completed. A full episode thus forms a trajectory $\tau = (h_1, \rho_1, u_1, \ldots, h_T, \rho_T, u_T, r_T)$. Given a batch of N trajectories collected using the old policy $\pi_{\theta_{\text{old}}}$, the clipped policy optimization objective is

$$\mathcal{J}_{\text{RL}}(\theta) = \mathbb{E}_{\{\tau^{(i)}\}_{i=1}^{N} \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \min \left(\varrho_t^{(i)} \widehat{A}_t^{(i)}, \text{clip} \left(\varrho_t^{(i)}, 1-\epsilon, 1+\epsilon \right) \widehat{A}_t^{(i)} \right) \right] - \beta \, D_{\text{KL}} \left(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\text{ref}}(\cdot \mid x) \right), \tag{1}$$

where $\varrho_t^{(i)} = \frac{\pi_\theta\left(a_t^{(i)} | h_t^{(i)}, x\right)}{\pi_{\theta_{\text{old}}}\left(a_t^{(i)} | h_t^{(i)}, x\right)}$ is the importance-sampling ratio for the complete textual action $a_t^{(i)} = (\rho_t^{(i)}, u_t^{(i)})$. Here, ϵ is the clipping threshold, $\pi_{\theta_{\text{old}}}$ denotes the policy used to collect the rollout trajectories, π_{ref} is a fixed reference policy, β controls the strength of the KL regularization. To reduce the variance of the estimated advantage $\widehat{A}_t^{(i)}$, a common way is to estimate the value of the current interaction history $V(h_t^{(i)}) = \mathbb{E}_{\tau \sim \pi_{\theta_{\text{old}}}} \left[r_T \mid h_t^{(i)}, x \right]$, and use it as a baseline. Consequently, accurately estimating the value of an intermediate history is a central objective of credit assignment methods.

Local grouping methods such as GRPO and GiGPO estimate this value by partitioning interaction histories into groups according to heuristic local rules and estimate values from within-group Monte Carlo (MC) returns. Let $\mathcal{G}$ denote the group containing $h_t^{(i)}$. The resulting advantage estimate can be written as $\widehat{A}_t^{(i)} = \frac{r_T^{(i)} - \widehat{V}\left(h_t^{(i)}\right)}{\sigma(\mathcal{G})}$, where $r_T^{(i)}$ denotes the final return associated with $h_t^{(i)}$, and $\sigma(\mathcal{G})$ denotes the standard deviation of trajectory rewards within group $\mathcal{G}$. Commonly used symbols in the main text and their meanings are summarized in Appendix B.

Cloned Hidden Markov Model. A Cloned Hidden Markov Model (CHMM) is a structured HMM in which multiple cloned hidden states deterministically emit the same observation. Let $o_t \in \mathcal{O}$ denote the observation at step t and $z_t \in \mathcal{Z}$ denote the latent clone state. Each clone state emits exactly one observation through a fixed deterministic mapping

$$g : \mathcal{Z} \to \mathcal{O}, \qquad o_t = g(z_t), \tag{2}$$

while multiple clone states may share the same observation, i.e., $g(z) = g(z')$ for $z \neq z'$. The dynamics among clone states are parameterized by a transition matrix $P_{uv} = p(z_{t+1} = v \mid z_t = u)$, which is learned from sequential data.

CHMM introduces clone states to distinguish different temporal contexts that produce the same observation. Consequently, higher-order and long-range dependencies in the observed sequence can be encoded into the latent clone state, allowing the sequence dynamics to become first-order Markovian in the latent space.

3 METHOD

The core idea of CADY is to estimate value by exploiting the global trajectory structure on which value fundamentally depends, thereby reducing value estimation error. Section 3.1 presents an overview of CADY. By learning global trajectory dynamics, it represents interaction histories with latent predictive states that capture differences in their future behavior, enabling value-relevant statistical sharing beyond heuristic local matching. It then estimates history values through the learned dynamics, allowing credit information to propagate across states. Section 3.2 further provides a theoretical explanation for why this design can achieve lower variance than a Monte Carlo estimator.

3.1 CREDIT ASSIGNMENT VIA GLOBAL DYNAMICS

This subsection introduces the three main stages of CADY: fitting global trajectory dynamics, estimating continuation values for interaction histories, and constructing step-level advantages. Figure 2 illustrates the overall workflow of CADY.

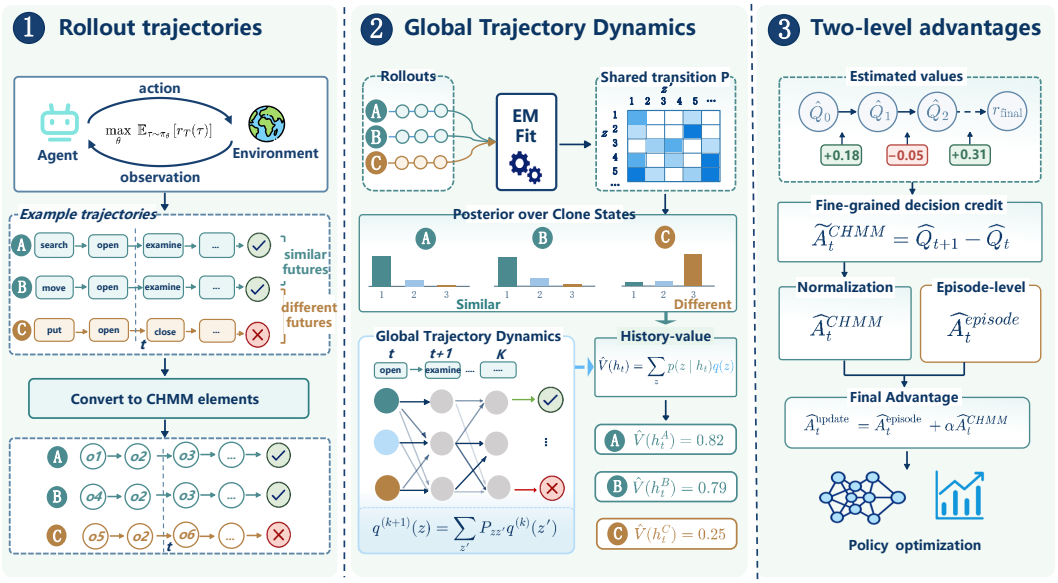


Figure 2: Overview of CADY. Trajectories generated through agent–environment interactions are first converted into sequences of elements. The CHMM is then trained on these sequences to model global trajectory dynamics, learn latent representations, and estimate history values. The resulting CHMM-based advantages are normalized and combined with episode-level advantages to form the final advantages used for policy updates.

Fitting Global Trajectory Dynamics. For each question, we fit a separate CHMM using its rollout trajectories. We first convert each agent–environment interaction trajectory into a sequence of discrete CHMM elements. The choice of elements depends on the environment; for exposition, we use executable actions u_t[1] as the running example. Let $o_{1:T} = (o_1, \ldots, o_T)$ denote the element sequence corresponding to a trajectory. We further append a terminal element indicating the binary task outcome, $o_{T+1} \in \{\texttt{CORRECT}, \texttt{WRONG}\}$. These elements constitute the observed trajectory sequence to be modeled by the CHMM. Each observed element o_t is deterministically associated with a set of clone states, while different clones of the same element capture different temporal contexts. The CHMM is then fitted to the rollout sequences of that question using the Expectation–Maximization (EM) algorithm. By fitting the global rollout dynamics, the CHMM maps histories with similar future trajectories to similar latent states. Since the value of a history is fundamentally determined by its expected future trajectory, such predictive grouping can therefore mitigate grouping-induced bias and improve the accuracy of value estimation.

History-Value Estimation. We next use the learned transition dynamics to estimate the value of each interaction history. Let T denote the maximum number of agent–environment interaction steps allowed by the task. Since each interaction step corresponds to exactly one CHMM element, the dynamic-programming computation below is performed directly at the interaction-step level. For an interaction history h_t, we also use h_t to denote the corresponding element prefix $(o_1, \ldots, o_{t-1})$, with a slight abuse of notation. Conditioned only on this prefix, causal CHMM inference yields a posterior distribution over the current clone state, $p(z_t = z \mid h_t)$. Let $\mathcal{Z}_{\mathrm{C}}$ and $\mathcal{Z}_{\mathrm{W}}$ denote the clone states associated with the terminal elements CORRECT and WRONG, respectively. Given the learned transition matrix P, we use dynamic programming to compute the probability of reaching a successful terminal state within a given number of future transitions. Specifically, let $q_t(z)$ denote the probability of reaching CORRECT from clone state z within at most $T + 1 - t$ future CHMM

[1]The choice of elements is flexible, as long as they sufficiently represent task progress. For example, in deterministic environments, executable actions may be sufficient to characterize trajectory progress, whereas in stochastic environments, environmental feedback can be additionally incorporated as elements. To keep the presentation concise, we use the former case as the running example throughout this paper.

transitions. Its boundary condition is

$$q_{T+1}(z) = \begin{cases} 1, & z \in \mathcal{Z}_{\mathrm{C}}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

and for $t \leq T$,

$$q_t(z) = \begin{cases} 1, & z \in \mathcal{Z}_{\mathrm{C}}, \\ 0, & z \in \mathcal{Z}_{\mathrm{W}}, \\ \sum_{z'} P_{zz'} q_{t+1}(z'), & \text{otherwise}. \end{cases} \tag{4}$$

Since the current prefix has consumed $t-1$ action elements, its remaining CHMM horizon is $T+1-t$, including the terminal outcome element. We therefore estimate the value of history h_t as *the probability of eventual task success within the remaining horizon* under the learned global trajectory dynamics.

$$\widehat{V}_{\mathrm{CHMM}}(h_t) = \widehat{Q}_{\mathrm{CHMM}}(h_{t-1}, a_{t-1}) = \sum_z p(z_t = z \mid h_t)\, q_t(z). \tag{5}$$

We denote $\widehat{Q}_t \triangleq \widehat{Q}_{\mathrm{CHMM}}(h_{t-1}, a_{t-1}) = \widehat{V}_{\mathrm{CHMM}}(h_t)$ for $t \geq 2$, and let $\widehat{Q}_1 = \widehat{V}_{\mathrm{CHMM}}(h_1)$ denote the value before the first environment action, estimated by the empirical success rate.

Importantly, a rollout need not directly visit h_t to improve its value estimate: it may still improve the shared transition dynamics P, which are then used to evaluate h_t through Eq. 5. Unlike approaches that estimate the value of h_t using only the average rewards of rollouts that pass through h_t, this global-dynamics-based approach makes use of cross-state information to reduce estimation variance.

Advantage Construction. CADY first constructs a fine-grained step-level credit signal from the CHMM-estimated history values using the one-step TD difference. We denote this raw CHMM credit by $\widetilde{A}_t^{\mathrm{CHMM}}$:

$$\widetilde{A}_t^{\mathrm{CHMM}} = \begin{cases} \widehat{Q}_{t+1} - \widehat{Q}_t, & t < T, \\ r_T - \widehat{Q}_T, & t = T. \end{cases} \tag{6}$$

For non-terminal steps, the credit reflects how much the corresponding action changes the predicted probability of eventual task success. For the final step, the bootstrapped value is replaced by the observed binary terminal reward r_T. The raw TD credits satisfy the telescoping identity

$$\sum_{t=1}^{T} \widetilde{A}_t^{\mathrm{CHMM}} = r_T - \widehat{Q}_1, \tag{7}$$

so that the CHMM primarily determines how the overall trajectory-level credit is distributed across individual decisions. To stabilize its scale, we normalize the raw CHMM credit by a robust scale factor $\widehat{s}_{\mathrm{CHMM}}$, defined using the median magnitude of the raw credit, and then apply a smooth asinh compression (see Appendix C.2 for details):

$$\widehat{A}_t^{\mathrm{CHMM}} = \mathrm{asinh}\left(\frac{\widetilde{A}_t^{\mathrm{CHMM}}}{\widehat{s}_{\mathrm{CHMM}}}\right). \tag{8}$$

The asinh retains fine-grained variations among small step-level credits while preventing occasional large values from dominating the policy update. Finally, we combine the fine-grained step-level CHMM advantage with the GRPO-style episode-level advantage to obtain the final advantage used by CADY:

$$\boxed{\widehat{A}_t^{\mathrm{update}} = \widehat{A}_t^{\mathrm{episode}} + \alpha \widehat{A}_t^{\mathrm{CHMM}}}, \tag{9}$$

where α is a hyperparameter that balances the two-level advantages.

3.2 Theoretical Analysis

We now provide a theoretical perspective on why estimating values through shared global trajectory structure can be statistically more efficient than simple Monte Carlo (MC) estimation. To isolate this effect from CADY's advantage in state abstraction, we consider the setting in which history

grouping is fixed and shared by both estimators. Specifically, let $C_t = \varphi(h_t) \in \mathcal{C}$ denote a fixed discrete abstraction of interaction histories, and assume that it is *continuation-sufficient*: histories mapped to the same state $c \in \mathcal{C}$ have the same policy-conditioned expected terminal reward. This assumption allows us to focus only on the statistical effect of exploiting shared trajectory structure.

Consider N independent rollout trajectories $\{\tau^{(i)}\}_{i=1}^{N}$ collected under a fixed policy, with terminal rewards $r_T^{(i)}$. For a target state c, let $K_{i,c}$ denote the number of times trajectory $\tau^{(i)}$ visits c. A simple MC estimator uses only trajectories that visit c and estimates its value by

$$\widehat{V}_{\mathrm{MC}}(c) = \frac{\sum_{i=1}^{N} K_{i,c} r_T^{(i)}}{\sum_{i=1}^{N} K_{i,c}}. \tag{10}$$

In contrast, a shared-dynamics estimator first fits a parametric trajectory model $\mathbb{P}_{\widehat{\theta}}$ using all rollout trajectories and then estimates the value of c by the expected terminal reward implied by $\mathbb{P}_{\widehat{\theta}}$:

$$\widehat{V}_{\mathrm{dyn}}(c) = \mathbb{E}_{\mathbb{P}_{\widehat{\theta}}}\left[r_T \mid C_t = c\right]. \tag{11}$$

Theorem 3.1 (**First-order efficiency of shared trajectory estimation**). *Assume that (i) the fixed abstraction φ is continuation-sufficient under $\mathbb{P}_\theta$ for θ in a neighborhood of the true parameter θ_0, (ii) the policy-induced trajectory distribution belongs to the correctly specified finite-dimensional regular parametric family $\{\mathbb{P}_\theta : \theta \in \Theta\}$, and (iii) $\widehat{\theta}$ is a regular first-order efficient estimator of its true parameter. Then both $\widehat{V}_{\mathrm{MC}}(c)$ and $\widehat{V}_{\mathrm{dyn}}(c)$ consistently estimate the same population quantity $V(c)$, and their first-order asymptotic variances satisfy*

$$C_{\mathrm{dyn}}(c) \le C_{\mathrm{MC}}(c). \tag{12}$$

Moreover, the inequality is strict if $\Pr(\psi_{\mathrm{dyn},c}(\tau) \neq 0 \mid \tau$ does not visit $c) > 0$, where $\psi_{\mathrm{dyn},c}(\tau)$ denotes the first-order contribution of trajectory τ to the asymptotic estimation error of $\widehat{V}_{\mathrm{dyn}}(c)$.

The complete proof in Appendix A.1 reveals how shared trajectory structure introduces additional statistical information. The first-order contribution of a rollout τ to the estimation error of $V(c)$ is quantified by the corresponding influence function.

$$\psi_{\mathrm{dyn},c}(\tau) = g_c^{\top} I^{-1} S(\tau), \tag{13}$$

where $S(\tau)$ is the score of the complete rollout with respect to the shared trajectory-model parameters, I is the corresponding Fisher information matrix, and g_c measures how sensitive the value $V(c)$ is to shared parameters. For the simple MC estimator, whenever a rollout τ does not visit the target state, it obviously contributes no information to the estimate of $V(c)$. However, for the shared-dynamics estimator, whenever the parameter information provided by τ, represented by $I^{-1}S(\tau)$, has a nonzero component along a direction to which $V(c)$ is sensitive, we have $\psi_{\mathrm{dyn},c}(\tau) \neq 0$. Such nonzero contributions yield a strict reduction in the estimator's first-order asymptotic variance. This provides a statistical interpretation of why cross-state information sharing is necessary, thereby supporting CADY's design of estimating values through learned global trajectory dynamics. More detailed discussion of Theorem 3.1 is provided in Appendix A.2.

4 Experiments

4.1 Setting

Models and Datasets We use Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct as the agent models. We evaluate our method on three challenging long-horizon agentic tasks: (1) CookingWorld (Madotto et al., 2020), where the agent operates in a simulated kitchen consisting of multiple rooms, reads a cookbook, searches for multiple ingredients, and completes the cooking procedure in the correct order; (2) TextWorld Commonsense (TWC) (Murugesan et al., 2021), where the agent must navigate across multiple rooms, locate various objects, and rearrange them into appropriate locations according to commonsense knowledge; and (3) multi-hop question answering (Jin et al., 2025), where the agent uses a retrieval tool to search over a large-scale Wikipedia corpus and obtain answers through multi-step information seeking. The long-horizon nature of these tasks, together with their spatial and logical complexity, makes them particularly challenging for agents to maintain

Table 1: Main results across embodied-agent and multi-hop QA benchmarks. All values are reported as percentages and averaged over three random seeds. Bold indicates the best performance in each category. Results across model scales are not comparable due to different environment settings.

Method	CookingWorld	TWC	Single-Hop QA			Multi-Hop QA		
			TriviaQA	NQ	PopQA	HotpotQA	MuSiQue	Bamboogle
Qwen2.5-1.5B-Instruct								
Turn-PPO	$10.2_{\pm 1.1}$	$56.8_{\pm 1.6}$	$51.7_{\pm 0.9}$	$31.5_{\pm 0.7}$	$40.4_{\pm 0.8}$	$28.5_{\pm 0.6}$	$21.0_{\pm 0.8}$	$31.1_{\pm 0.7}$
GRPO	$16.6_{\pm 1.3}$	$2.0_{\pm 0.5}$	$53.7_{\pm 0.8}$	$35.1_{\pm 0.6}$	$43.8_{\pm 0.7}$	$28.6_{\pm 0.7}$	$27.0_{\pm 0.7}$	$31.2_{\pm 0.8}$
GiGPO	$25.2_{\pm 1.5}$	$65.4_{\pm 1.2}$	$51.1_{\pm 1.0}$	$32.5_{\pm 0.8}$	$42.2_{\pm 0.7}$	$29.0_{\pm 0.6}$	$25.1_{\pm 0.7}$	$28.5_{\pm 0.9}$
SHADOW	$52.6_{\pm 1.4}$	$60.0_{\pm 1.4}$	$\mathbf{54.4}_{\pm 0.6}$	$36.1_{\pm 0.5}$	$43.7_{\pm 0.6}$	$30.6_{\pm 0.5}$	$28.1_{\pm 0.6}$	$30.0_{\pm 0.7}$
HCAPO	$64.3_{\pm 1.2}$	$65.0_{\pm 1.1}$	$49.7_{\pm 0.9}$	$32.7_{\pm 0.7}$	$40.3_{\pm 0.8}$	$30.4_{\pm 0.5}$	$25.3_{\pm 0.7}$	$30.2_{\pm 0.7}$
CADY (Ours)	$\mathbf{95.3}_{\pm 1.0}$	$\mathbf{75.3}_{\pm 0.7}$	$53.9_{\pm 0.5}$	$\mathbf{37.2}_{\pm 0.4}$	$\mathbf{46.5}_{\pm 0.5}$	$\mathbf{31.6}_{\pm 0.4}$	$\mathbf{32.2}_{\pm 0.5}$	$\mathbf{39.0}_{\pm 0.6}$
Qwen2.5-7B-Instruct								
Turn-PPO	$48.7_{\pm 1.5}$	$30.9_{\pm 1.4}$	$59.7_{\pm 0.8}$	$42.6_{\pm 0.7}$	$55.3_{\pm 0.8}$	$41.4_{\pm 0.7}$	$37.2_{\pm 0.9}$	$40.5_{\pm 0.8}$
GRPO	$55.2_{\pm 1.4}$	$5.3_{\pm 0.7}$	$62.1_{\pm 0.7}$	$45.1_{\pm 0.6}$	$53.0_{\pm 0.6}$	$41.8_{\pm 0.6}$	$40.0_{\pm 0.5}$	$41.2_{\pm 0.6}$
GiGPO	$65.9_{\pm 1.2}$	$36.2_{\pm 1.3}$	$61.5_{\pm 0.8}$	$44.3_{\pm 0.7}$	$57.0_{\pm 0.7}$	$42.4_{\pm 0.6}$	$40.2_{\pm 0.6}$	$38.6_{\pm 0.8}$
SHADOW	$83.2_{\pm 1.1}$	$32.6_{\pm 1.2}$	$62.9_{\pm 0.5}$	$45.2_{\pm 0.4}$	$57.6_{\pm 0.6}$	$43.8_{\pm 0.4}$	$41.4_{\pm 0.3}$	$39.7_{\pm 0.7}$
HCAPO	$76.8_{\pm 1.0}$	$34.8_{\pm 1.1}$	$59.9_{\pm 0.8}$	$46.0_{\pm 0.7}$	$54.8_{\pm 0.8}$	$41.5_{\pm 0.5}$	$40.5_{\pm 0.5}$	$39.2_{\pm 0.7}$
CADY (Ours)	$\mathbf{97.7}_{\pm 1.4}$	$\mathbf{68.5}_{\pm 0.8}$	$\mathbf{63.7}_{\pm 0.4}$	$\mathbf{46.6}_{\pm 0.4}$	$\mathbf{58.9}_{\pm 0.3}$	$\mathbf{44.4}_{\pm 0.3}$	$\mathbf{42.1}_{\pm 0.2}$	$\mathbf{43.8}_{\pm 0.6}$

consistent multi-step decision making. For CookingWorld and TWC, the 7B model is evaluated under more challenging environment configurations than the 1.5B model. Consequently, success rates should be compared across methods within the same model scale, rather than directly across model scales. See Appendix C.3 for more details of the task settings.

Baselines and Methods We compare CADY against three representative categories of credit assignment baselines: (1) methods based on within-group rollout statistics, including GRPO (Shao et al., 2024) and the more fine-grained grouping methods GiGPO (Feng et al., 2026) and SHADOW (Liu et al., 2026); (2) PPO-style methods that rely on a learned critic model, such as Turn-PPO (Li et al., 2026a); and (3) step-level scoring methods based on model self-reflection, such as HCAPO (Tan et al., 2026). Hyperparameter details are provided in Appendix C.4.

4.2 Main Results

Superior Performance on Benchmarks. As shown in Table 1, CADY achieves strong gains on long-horizon agentic tasks, particularly those requiring multi-step credit assignment. On the embodied benchmarks with up to 50 decision steps, the 1.5B model reaches **95.3%** success on CookingWorld, outperforming HCAPO (64.3%) by **31.0 points**, and **75.3%** on TWC, exceeding GiGPO (65.4%) by **9.9 points**. The gains persist at the 7B scale: CADY achieves **97.7%** on CookingWorld and **68.5%** on TWC, outperforming the strongest baselines by **14.5** and **32.3 points**, respectively. Figures 1c and 3a further show that CADY continues improving to substantially higher success rates, with the gap over GiGPO widening later in training, especially on CookingWorld. A similar trend appears in multi-hop QA (Figure 3b), where CADY maintains higher validation performance and delivers its largest gains on multi-hop datasets, improving Bamboogle by **+7.8 points** at 1.5B and **+2.6 points** at 7B, while remaining competitive on single-hop datasets.

More Accurate Credit Assignment. To examine whether CADY's accuracy gains stem from more accurate value estimation and, subsequently, more informative advantages, we directly evaluate the estimated history values. In Figure 1b, we randomly sample 1K history states during training and perform 16 additional rollouts from each state. The mean reward of these rollouts is used as an approximate ground-truth value. We then compute the squared error between this value and the estimates produced by CADY and the baselines. CADY shows a substantially lower error distribution, indicating more accurate value estimation. Figure 3c provides a real case study illustrating how this translates into better credit assignment. CADY assigns positive advantages to actions that make effective progress, such as taking and cooking the chicken wing, as well as useful preceding actions such as moving to the correct room and opening the fridge. In contrast, incorrect movement receives

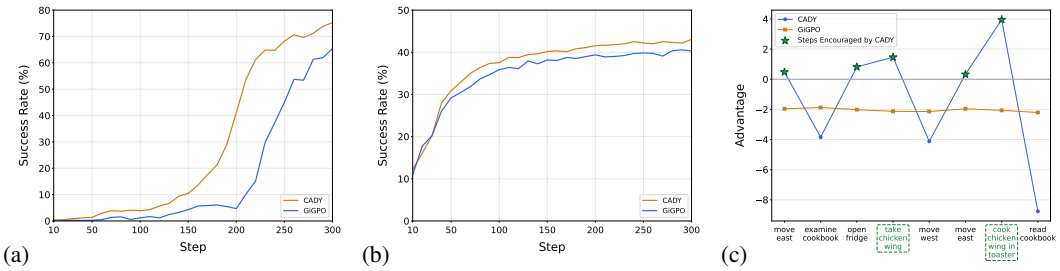


Figure 3: Validation performance of CADY and GiGPO on (a) TWC, (b) multi-hop QA. (c) Comparison of advantage estimates between CADY and GiGPO on a representative case.

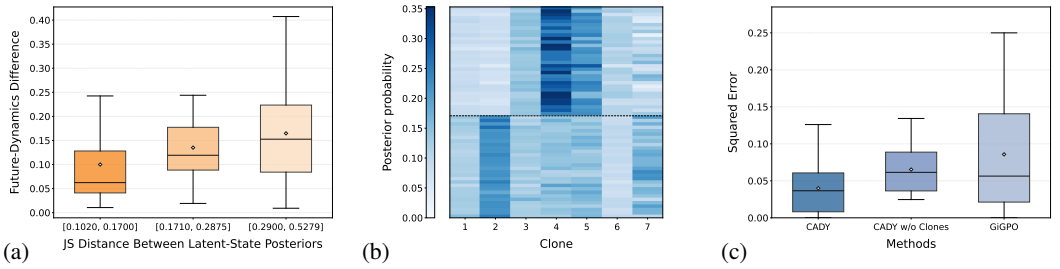


Figure 4: (a) Relationship on AutomationBench between future-trajectory dynamics differences and latent-state posterior differences for history pairs sharing the same current action and observation. (b) Clone-state posterior distributions on AutomationBench for held-out histories sharing the same current observation; the dashed line separates histories where the table-reading interface has been obtained (top) from those where it has not (bottom). (c) Comparison of squared errors in value estimation for the component ablation study.

negative advantages. By comparison, GiGPO produces nearly insensitive advantages across these steps. These results demonstrate the accuracy of CADY's value estimation and credit assignment.

Effective Separation of Future Trajectory Dynamics. We further examine whether CADY can distinguish histories with different future trajectory dynamics. We additionally sample 1K history states that are not used for training and perform 16 rollouts from each state. We then form pairs of histories sharing the same current action and environmental observation. For each pair, we measure the difference in future dynamics by the average mismatch between their continuation trajectories defined in Appendix C.1, and measure the representation difference by the JS distance between their latent-state posterior distributions. As shown in Figure 1a, larger future-dynamics differences correspond to larger posterior distances, indicating that CADY separates histories with distinct continuation behaviors. In contrast, local heuristic methods such as GiGPO map such pairs to the same state, which can introduce larger estimation bias.

Scalability of Latent-State Grouping A natural concern is whether the implicit grouping remains effective in environments with much larger observation and action spaces, where the CHMM element space grows and EM fitting may converge to poor local optima. In such cases, posterior distributions of locally aliased histories may collapse, limiting the CHMM's ability to distinguish them. We therefore isolate the fitting stage on AutomationBench (Shepard & Salimans, 2026), which contains 47 simulated applications and roughly 500 API operations (Appendix C.3). We fit a CHMM to 300 completed sales-task trajectories generated by DeepSeek-V4.1-Flash (DeepSeek-AI, 2026) and evaluate held-out trajectories. Figure 4a reproduces the future-dynamics/posterior-distance relationship observed in the embodied environments, suggesting that the CHMM latent states can encode distinct patterns of future dynamics. We further illustrate this behavior with a case where GiGPO-style local matching necessarily fails: two histories issue the same table-query API and receive identical metadata, but their continuations differ. An agent that has already obtained the table-reading interface proceeds to read the target content, whereas one that has not continues with

additional rounds of querying. Figure 4b shows clear posterior separation, with a cross-cluster JS distance of 0.289 versus 0.082 within clusters, comparable to that observed in embodied environments. These results suggest that per-question CHMM fitting remains effective in complex, realistic tool-use settings, demonstrating the scalability potential of CADY.

4.3 Ablation

Component Ablation. We disentangle the contributions of clone-based history representation and value estimation based on shared dynamics. As shown in Figure 4c, we construct a *CADY w/o Clones* variant that adopts the same state grouping as GiGPO but estimates values through shared global trajectory dynamics. This variant achieves lower value-estimation squared error than GiGPO, demonstrating the benefit of exploiting shared dynamics beyond local Monte Carlo estimation. However, it still performs worse than the full CADY model. This gap highlights the importance of clone latent states, which allow CADY to distinguish histories with different future dynamics.

Hyperparameter Ablation. We study the sensitivity of CADY to the advantage-weight coefficient α and the maximum number of clone states. As shown in Figure 5, performance remains stable across different settings for both hyperparameters. These results suggest that CADY is not sensitive to precise hyperparameter choices.

5 Related Work

Credit Assignment in Long-Horizon Agentic RL. Classical RL often uses critic models for value estimation (Schulman et al., 2017), while critic-based agentic RL methods (Xi et al., 2026; Li et al., 2026a; Yue et al., 2025; Hou et al., 2026) introduce substantial memory and computation costs and may generalize poorly (Chen et al., 2026). Other approaches (Tan et al., 2026; Kang et al., 2026; Wang et al., 2026; Tao et al., 2026; Li et al., 2026b) rely on internal signals such as confidence or self-reflection, but depend on reliable self-evaluation. A common critic-free alternative estimates values from Monte Carlo statistics of locally matched rollouts: GRPO (Shao et al., 2024) uses mean response rewards as a coarse trajectory-level baseline, GiGPO (Feng et al., 2026) and HGPO (He et al., 2026) form finer groups using environmental feedback, and SHADOW (Liu et al., 2026) further distinguishes environment states. However, these methods still rely on local heuristics. CADY instead performs structured credit assignment through global trajectory dynamics.

Latent State Models. Latent-state models, such as HMMs (Rabiner, 1989), are commonly used to capture statistical regularities in complex sequential data (Becker & Neumann, 2022). CHMMs (Dedieu et al., 2019) associate multiple clone states with the same observation, enabling different temporal contexts and long-range dependencies to be represented. George et al. (2021) demonstrate their ability to recover spatial structure and support planning. In the context of LLMs, Zhang et al. (2023) and Zhang et al. (2024) use HMMs to model language sequences and guide generation under logical constraints, while Kung et al. (2026) extend this idea to steer reasoning behaviors. CADY instead uses latent dynamics for credit assignment and can naturally extend to more expressive models such as HHMMs (Fine et al., 1998) and RSSMs (Hafner et al., 2019a;b).

6 Conclusion

In this work, we introduce **CADY**, a structured credit assignment method for long-horizon agentic reinforcement learning based on global trajectory dynamics. Instead of relying on heuristic local history matching, CADY learns predictive latent states from rollout trajectories and exploits shared transition dynamics to estimate history values and construct fine-grained step-level advantages. Our theoretical analysis shows that, under standard regularity and model-specification assumptions, shared-dynamics value estimation can achieve lower first-order asymptotic variance than local Monte Carlo estimation. Across challenging embodied-agent and multi-hop QA benchmarks, CADY improves both history-value estimation and downstream policy learning over strong baselines. Fitting experiments in complex tool-use settings further demonstrate the scalability potential of CADY. In practice, CADY can be naturally extended with more expressive latent-state models to capture increasingly complex trajectory dynamics.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, reference and literature search, and code generation. We did not use generative AI tools to develop the core methodology, design the experiments, or construct the mathematical proofs. All AI-assisted content was carefully reviewed by the authors. We verified the correctness of the generated code, checked the validity and reliability of the cited references, and ensured the accuracy of the translated text. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

REFERENCES

Hao Bai, Yifei Zhou, Mert Cemri, Jiayi Pan, Alane Suhr, Sergey Levine, and Aviral Kumar. Digirl: Training in-the-wild device-control agents with autonomous reinforcement learning. *Advances in Neural Information Processing Systems*, 37:12461–12495, 2024.

Philipp Becker and Gerhard Neumann. On uncertainty in deep state space models for model-based reinforcement learning. *arXiv preprint arXiv:2210.09256*, 2022.

Zizhe Chen, Jiqian Dong, Yizhou Tian, Garry Yang, Yongqiang Chen, Zhitang Chen, and James Cheng. Hista and numca: Estimate state value effectively for llm reinforcement learning. *arXiv preprint arXiv:2605.29782*, 2026.

Antoine Dedieu, Nishad Gothoskar, Scott Swingle, Wolfgang Lehrach, Miguel Lázaro-Gredilla, and Dileep George. Learning higher-order sequential structure with cloned hmms. *arXiv preprint arXiv:1905.00507*, 2019.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026.

Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32(1):41–62, 1998.

Preston Fu, Kevin Frans, Oleh Rybkin, Sergey Levine, and Aviral Kumar. Long-horizon language model reinforcement learning via progressive point matching. *arXiv preprint arXiv:2609.07303*, 2026.

Hiroki Furuta, Kuang-Huei Lee, Ofir Nachum, Yutaka Matsuo, Aleksandra Faust, Shixiang Gu, and Izzeddin Gur. Multimodal web navigation with instruction-finetuned foundation models. In *International Conference on Learning Representations*, volume 2024, pp. 29528–29558, 2024.

Dileep George, Rajeev V Rikhye, Nishad Gothoskar, J Swaroop Guntupalli, Antoine Dedieu, and Miguel Lázaro-Gredilla. Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. *Nature communications*, 12(1):2392, 2021.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019b.

Shuo He, Lang Feng, Xin Cheng, Lei Feng, Bo An, et al. Hierarchy-of-groups policy optimization for long-horizon agentic tasks. In *International Conference on Learning Representations*, volume 2026, pp. 27572–27593, 2026.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Zhenyu Hou, Yujiang Li, Jie Tang, and Yuxiao Dong. Single-rollout asynchronous optimization for agentic reinforcement learning. *arXiv preprint arXiv:2607.07508*, 2026.

Wenqi Huang and Peter Jiang. Deepswe v1.1: a cleaner, more reproducible benchmark for frontier coding agents, 2026. URL `https://github.com/datacurve-ai/deep-swe`.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Zhewei Kang, Aosong Feng, Sergey Levine, Dawn Song, and Xuandong Zhao. Vimpo: Value-implicit policy optimization for llms. *arXiv preprint arXiv:2606.20008*, 2026.

Po-Nien Kung, Zhen Yang, Jeffrey Luo, Cheng-Fu Yang, Haikang Deng, Zi-Yi Dou, Yinfei Yang, Nanyun Peng, Zhe Gan, and Kai-Wei Chang. Learning structured reasoning via tractable trajectory control. *arXiv preprint arXiv:2603.01641*, 2026.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Junbo Li, Peng Zhou, Rui Meng, Meet P Vadera, Lihong Li, and Yang Li. Turn-ppo: Turn-level advantage estimation with ppo for improved multi-turn rl in agentic llms. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6227–6243, 2026a.

Zhongyi Li, Wan Tian, Jinju Chen, Huiming Zhang, Yang Liu, Yikun Ban, and Fuzhen Zhuang. Counterfactual credit policy optimization for multi-agent collaboration. *arXiv preprint arXiv:2603.21563*, 2026b.

Yuze Liu, Chaochao Lu, and Chao Yang. Shadow: Dynamic-aware credit assignment against long-horizon tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23935–23944, 2026.

Andrea Madotto, Mahdi Namazifar, Joost Huizinga, Piero Molino, Adrien Ecoffet, Huaixiu Zheng, Alexandros Papangelis, Dian Yu, Chandra Khatri, and Gokhan Tur. Exploration based language learning for text-based games. *arXiv preprint arXiv:2001.08868*, 2020.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Ryan Marten, Alex Shaw, Ivan Bercovich, Benedikt Droste, Tommaso Cerruti, Steven Dillmann, Ruiyang Wang, Dariush Wahdany, Allen Hart, Karl Krauth, ScaleAI, Snorkel AI, Turing, gNucleus AI, Boolean AI, Nicholas Carlini, Shengrui Lyu, Anjiang Wei, Arpandeep Khatua, Björn Plüster, Chaitanya Dwivedi, Christine Sutcliffe, Yuming, David Tivris, Di Wang, Hui Wen Goh, Hanwen Xing, Haowei Lin, Irakli Salia, Jaejung Seol, Jiajun Bao, Jialin Ouyang, Junha Park, Liam Walsh, Luyang Kong, Maksim Ivanov, Malte Ubl, Mikhail Liamets, Orfeas Menis, Piotr Migdal, Qingquan Bao, Raj Movva, Roey Ben Chaim, Namburi Srinath, Sergey Bogdanik, Shubham Yadav, Stephen Benjamin, Tony Kung, Walker Hughes, Xin Lan, Himanshu Gupta, Swaroop Mishra, Chenguang Wang, Hao He, Jianhong Tu, Kyle Montgomery, Zengji Tu, Atharva Naik, David Mortensen, Ivan Zhang, Yash Mathur, Emmy Liu, Karanpartap Singh, Michael Yu, Steven Feng, Varun Gangal, Zhuofu Tao, Sherry Ruan, Jonas Mueller, Joan Cabezas, Justin Bauer, Kevin Xiang Li, Robert Zhang, Aaron Feller, Alec Madayan, Leon Chen, Ben Feuer, Xiangyi Li, Boxuan Li, Harsh Raj, Samuel Galler, Lin Shi, Ivgeni Segal, Kelly Buchanan, Shreyas P., Rishi Desai, Aaron Schneider, Chris Settles, Xiangning

Lin, Marianna Nezhurina, Andrew Wang, Michał Kowalczyk, Jin-Xiang Zhao, Sanyam Satia, Jessie Hu, Sherif Atef, Kobe Chen, Sam Vance, Gian Segato, Jenia Jitsev, Alex Dimakis, Mike Merrill, Andy Konwinski, and Ludwig Schmidt. Terminal-Bench, August 2026. URL https://github.com/harbor-framework/terminal-bench.

Keerthiram Murugesan, Mattia Atzeni, Pavan Kapanipathi, Pushkar Shukla, Sadhana Kumaravel, Gerald Tesauro, Kartik Talamadupula, Mrinmaya Sachan, and Murray Campbell. Text-based rl agents with commonsense knowledge: New challenges, environments and baselines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9018–9027, 2021.

OpenAI. Introducing deep research. Technical report, OpenAI, 2025. URL https://openai.com/index/introducing-deep-research/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Daniel Shepard and Robin Salimans. Automationbench. *arXiv preprint arXiv:2604.18934*, 2026.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Hui-Ze Tan, Xiao-Wen Yang, Hao Chen, Jie-Jing Shao, Yi Wen, Yuteng Shen, Weihong Luo, Xiku Du, Lan-Zhe Guo, and Yu-Feng Li. Hindsight credit assignment for long-horizon llm agents. *arXiv preprint arXiv:2603.08754*, 2026.

Leitian Tao, Baolin Peng, Wenlin Yao, Tao Ge, Hao Cheng, Mike Hang Wang, Jianfeng Gao, and Sharon Li. Trace: Turn-level reward assignment via credit estimation for long-horizon agents. *arXiv preprint arXiv:2607.13988*, 2026.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, et al. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. In *International Conference on Learning Representations*, volume 2026, pp. 104502–104529, 2026.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Honglin Guo, Jiaqi Liu, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, et al. Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning. *arXiv preprint arXiv:2509.08755*, 2025.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Zhihao Zhang, Wenxiang Chen, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, et al. Agentprm: Process reward models for llm agents via step-wise promise and progress. In *Proceedings of the ACM Web Conference 2026*, pp. 4184–4195, 2026.

Peiran Xu, Zhuohao Li, Xiaoying Xing, Guannan Zhang, Debiao Li, and Kunyu Shi. Hybrid reward normalization for process-supervised non-verifiable agentic tasks. *arXiv preprint arXiv:2509.25598*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.

Huining Yuan, Zelai Xu, Huaijie Wang, Xiangmin Yi, Jiaxuan Gao, Xiao-Ping Zhang, Yu Wang, Chao Yu, and Yi Wu. Verifiable process rewards for agentic reasoning. *arXiv preprint arXiv:2605.10325*, 2026.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Chenchen Zhang. From reasoning to agentic: Credit assignment in reinforcement learning for large language models. *arXiv preprint arXiv:2604.09459*, 2026.

Honghua Zhang, Meihua Dang, Nanyun Peng, and Guy Van den Broeck. Tractable control for autoregressive language generation. In *International Conference on Machine Learning*, pp. 40932–40945. PMLR, 2023.

Honghua Zhang, Po-Nien Kung, Masahiro Yoshida, Guy Van den Broeck, and Nanyun Peng. Adaptable logical control for large language models. *Advances in Neural Information Processing Systems*, 37:115563–115587, 2024.

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *arXiv preprint arXiv:2503.15478*, 2025.

A Theoretical Proofs

A.1 Proof of Theorem 3.1

Proof. Throughout this proof, we condition on a fixed task prompt x and suppress x from the notation. Thus, all trajectory distributions, probabilities, and expectations below are understood to be conditional on this fixed task prompt.

Recall that φ denotes the fixed discrete abstraction of interaction histories. Define

$$C_t := \varphi(h_t).$$

By continuation sufficiency, for every θ in a neighborhood of the true parameter θ_0, histories mapped to the same $c \in \mathcal{C}$ have the same policy-conditioned expected terminal reward.

Let $\mathbb{P}_\theta$ denote the parametric distribution of a complete rollout trajectory τ under the fixed policy, and let $\mathbb{E}_\theta$ denote expectation under $\mathbb{P}_\theta$. The model is correctly specified, so the true trajectory distribution is $\mathbb{P}_{\theta_0}$.

By regularity, let μ be a common dominating measure for the parametric family and write

$$p_\theta(\tau) := \frac{d\mathbb{P}_\theta}{d\mu}(\tau)$$

for the corresponding density or probability mass function.

Fix a target state $c \in \mathcal{C}$. For a generic rollout τ, define

$$K_c(\tau) := \sum_{t=1}^{T} \mathbf{1}\{C_t = c\}$$

as the number of times that τ visits c. Hence, for the sampled rollouts $\{\tau^{(i)}\}_{i=1}^{N}$ in Section 3.2,

$$K_{i,c} = K_c(\tau^{(i)}).$$

Define

$$\kappa_c := \mathbb{E}_{\theta_0}[K_c(\tau)].$$

We consider a target state satisfying $\kappa_c > 0$, as required for the state-wise Monte Carlo estimator to be asymptotically well defined.

For θ in a neighborhood of θ_0, continuation sufficiency implies that the conditional expected terminal reward depends on a history h_t only through $C_t = \varphi(h_t)$. We denote this common continuation value by

$$V_\theta(c).$$

Equivalently, whenever $C_t = c$,

$$\mathbb{E}_\theta[r_T \mid h_t] = V_\theta(c).$$

We write

$$V(c) := V_{\theta_0}(c).$$

The shared-dynamics estimator considered in Section 3.2 can therefore be written as

$$\widehat{V}_{\mathrm{dyn}}(c) = V_{\widehat{\theta}}(c).$$

We prove the result in five steps.

Step 1: Common population target. Using the definition of $K_c(\tau)$ and continuation sufficiency,

$$\begin{aligned}
\mathbb{E}_\theta[K_c(\tau) r_T] &= \sum_{t=1}^{T} \mathbb{E}_\theta\left[\mathbf{1}\{C_t = c\} r_T\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_\theta\left[\mathbf{1}\{C_t = c\} \mathbb{E}_\theta[r_T \mid h_t]\right] \\
&= V_\theta(c) \sum_{t=1}^{T} \mathbb{E}_\theta\left[\mathbf{1}\{C_t = c\}\right] \\
&= V_\theta(c)\, \mathbb{E}_\theta[K_c(\tau)]. \qquad (14)
\end{aligned}$$

Therefore,

$$V_\theta(c) = \frac{\mathbb{E}_\theta[K_c(\tau) r_T]}{\mathbb{E}_\theta[K_c(\tau)]}. \qquad (15)$$

At the true parameter,

$$V(c) = \frac{\mathbb{E}_{\theta_0}[K_c(\tau) r_T]}{\kappa_c}. \qquad (16)$$

By the law of large numbers,

$$\frac{1}{N} \sum_{i=1}^{N} K_{i,c} r_T^{(i)} \xrightarrow{p} \mathbb{E}_{\theta_0}[K_c(\tau) r_T],$$

and

$$\frac{1}{N} \sum_{i=1}^{N} K_{i,c} \xrightarrow{p} \kappa_c > 0.$$

Hence

$$\widehat{V}_{\mathrm{MC}}(c) \xrightarrow{p} V(c).$$

The event on which no sampled rollout visits c satisfies

$$\mathbb{P}_{\theta_0}\left(\sum_{i=1}^{N} K_{i,c} = 0\right) = \mathbb{P}_{\theta_0}\left(K_c(\tau) = 0\right)^N \longrightarrow 0,$$

so any bounded convention used on this event does not affect the first-order asymptotics.

Since $\widehat{\theta}$ is a regular first-order efficient estimator, it is consistent:

$$\widehat{\theta} \xrightarrow{p} \theta_0.$$

By continuity of $\theta \mapsto V_\theta(c)$,

$$\widehat{V}_{\mathrm{dyn}}(c) = V_{\widehat{\theta}}(c) \xrightarrow{p} V_{\theta_0}(c) = V(c).$$

Thus $\widehat{V}_{\mathrm{MC}}(c)$ and $\widehat{V}_{\mathrm{dyn}}(c)$ consistently estimate the same population quantity.

Step 2: First-order expansion of the Monte Carlo estimator. Because T is finite and r_T is bounded, $K_c(\tau)$ and $K_c(\tau)r_T$ have finite second moments. Therefore, the multivariate central limit theorem applies to

$$\left(K_c(\tau)r_T,\, K_c(\tau)\right).$$

Applying the delta method to the ratio defining $\widehat{V}_{\mathrm{MC}}(c)$ yields

$$\sqrt{N}\left(\widehat{V}_{\mathrm{MC}}(c) - V(c)\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \psi_{\mathrm{MC},c}(\tau^{(i)}) + o_p(1), \tag{17}$$

where

$$\psi_{\mathrm{MC},c}(\tau) := \frac{K_c(\tau)}{\kappa_c}\big(r_T - V(c)\big). \tag{18}$$

By equation 16,

$$\mathbb{E}_{\theta_0}\left[\psi_{\mathrm{MC},c}(\tau)\right] = 0.$$

It follows that

$$C_{\mathrm{MC}}(c) = \mathbb{E}_{\theta_0}\left[\psi_{\mathrm{MC},c}(\tau)^2\right] = \frac{\mathbb{E}_{\theta_0}\left[K_c(\tau)^2\big(r_T - V(c)\big)^2\right]}{\kappa_c^2}. \tag{19}$$

Step 3: First-order expansion of the shared-dynamics estimator. Define the score of a complete rollout at the true parameter by

$$S(\tau) := \nabla_\theta \log p_\theta(\tau)|_{\theta=\theta_0}, \tag{20}$$

and define the Fisher information matrix by

$$I := \mathbb{E}_{\theta_0}\left[S(\tau)S(\tau)^\top\right]. \tag{21}$$

Under the regularity assumptions,

$$\mathbb{E}_{\theta_0}[S(\tau)] = 0,$$

and I is finite and nonsingular.

Define

$$g_c := \nabla_\theta V_\theta(c)|_{\theta=\theta_0}. \tag{22}$$

Thus g_c measures the sensitivity of the continuation value of state c to the shared trajectory-model parameters.

By the assumed regular first-order efficiency of $\widehat{\theta}$,

$$\sqrt{N}(\widehat{\theta} - \theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} I^{-1}S(\tau^{(i)}) + o_p(1). \tag{23}$$

Since $\theta \mapsto V_\theta(c)$ is differentiable at θ_0, the delta method gives

$$\sqrt{N}\left(\widehat{V}_{\mathrm{dyn}}(c) - V(c)\right) = g_c^\top \sqrt{N}(\widehat{\theta} - \theta_0) + o_p(1)$$
$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi_{\mathrm{dyn},c}(\tau^{(i)}) + o_p(1), \tag{24}$$

where

$$\psi_{\mathrm{dyn},c}(\tau) := g_c^\top I^{-1} S(\tau). \tag{25}$$

This agrees with the influence-function expression stated in Section 3.2.

Because $\mathbb{E}_{\theta_0}[S(\tau)] = 0$,

$$\mathbb{E}_{\theta_0}\left[\psi_{\mathrm{dyn},c}(\tau)\right] = 0.$$

Moreover,

$$C_{\mathrm{dyn}}(c) = \mathbb{E}_{\theta_0}\left[\psi_{\mathrm{dyn},c}(\tau)^2\right]$$
$$= g_c^\top I^{-1} \mathbb{E}_{\theta_0}\left[S(\tau)S(\tau)^\top\right] I^{-1} g_c$$
$$= g_c^\top I^{-1} g_c. \tag{26}$$

Step 4: Score-space projection. For any square-integrable function $f(\tau)$ that does not depend on θ, regularity of the dominated parametric model gives the score-differentiation identity

$$\nabla_\theta \mathbb{E}_\theta[f(\tau)]|_{\theta=\theta_0} = \mathbb{E}_{\theta_0}\left[S(\tau) f(\tau)\right]. \tag{27}$$

Since the abstraction φ is fixed, $K_c(\tau)$ is such a function of the observed trajectory.

Differentiating equation 15 at θ_0 and using equation 27 gives

$$g_c = \frac{1}{\kappa_c} \mathbb{E}_{\theta_0}\left[S(\tau) K_c(\tau)\big(r_T - V(c)\big)\right]$$
$$= \mathbb{E}_{\theta_0}\left[S(\tau)\psi_{\mathrm{MC},c}(\tau)\right]. \tag{28}$$

On the other hand, using equation 25,

$$\mathbb{E}_{\theta_0}\left[S(\tau)\psi_{\mathrm{dyn},c}(\tau)\right] = \mathbb{E}_{\theta_0}\left[S(\tau)S(\tau)^\top\right] I^{-1} g_c$$
$$= g_c. \tag{29}$$

Therefore,

$$\mathbb{E}_{\theta_0}\left[S(\tau)\left(\psi_{\mathrm{MC},c}(\tau) - \psi_{\mathrm{dyn},c}(\tau)\right)\right] = 0. \tag{30}$$

Hence $\psi_{\mathrm{MC},c} - \psi_{\mathrm{dyn},c}$ is orthogonal in $L^2(\mathbb{P}_{\theta_0})$ to every linear combination of the score components. Meanwhile,

$$\psi_{\mathrm{dyn},c}(\tau) = g_c^\top I^{-1} S(\tau)$$

belongs to the finite-dimensional span of those score components. Thus $\psi_{\mathrm{dyn},c}$ is exactly the $L^2(\mathbb{P}_{\theta_0})$ orthogonal projection of $\psi_{\mathrm{MC},c}$ onto the parametric score space.

By the Pythagorean identity,

$$C_{\mathrm{MC}}(c) = C_{\mathrm{dyn}}(c) + \mathbb{E}_{\theta_0}\left[\left(\psi_{\mathrm{MC},c}(\tau) - \psi_{\mathrm{dyn},c}(\tau)\right)^2\right]. \tag{31}$$

Therefore,

$$C_{\mathrm{dyn}}(c) \le C_{\mathrm{MC}}(c),$$

which proves the claimed first-order asymptotic variance ordering.

Step 5: Strict improvement from off-state information. If a rollout τ does not visit c, then

$$K_c(\tau) = 0,$$

and consequently

$$\psi_{\mathrm{MC},c}(\tau) = 0.$$

The condition in Theorem 3.1,

$$\Pr(\psi_{\mathrm{dyn},c}(\tau) \neq 0 \,|\, \tau \text{ does not visit } c) > 0,$$

therefore implies

$$\mathbb{P}_{\theta_0}\left(K_c(\tau) = 0,\, \psi_{\mathrm{dyn},c}(\tau) \neq 0\right) > 0.$$

Using equation 31,

$$\begin{aligned} C_{\mathrm{MC}}(c) - C_{\mathrm{dyn}}(c) &= \mathbb{E}_{\theta_0}\left[\left(\psi_{\mathrm{MC},c}(\tau) - \psi_{\mathrm{dyn},c}(\tau)\right)^2\right] \\ &\geq \mathbb{E}_{\theta_0}\left[\psi_{\mathrm{dyn},c}(\tau)^2 \mathbf{1}\{K_c(\tau) = 0\}\right] \\ &> 0. \end{aligned} \tag{32}$$

Hence

$$C_{\mathrm{dyn}}(c) < C_{\mathrm{MC}}(c).$$

Thus, a rollout need not directly visit the target state c in order to contain first-order information about its continuation value. Through the parameters shared by the global trajectory model, a trajectory that does not visit c may still contribute to estimating $V(c)$. □

A.2 Interpretation and Connection to CADY

Theorem 3.1 isolates the statistical benefit of estimating values through shared trajectory structure. To avoid confounding this effect with CADY's advantage in state abstraction, the comparison conditions on the same fixed continuation-sufficient abstraction for both estimators.

The theorem shows that, once a sufficient representation is fixed, and under idealized assumptions on model specification and parameter estimation, exploiting global trajectory structure can be statistically more efficient than purely Monte Carlo estimation. This directly supports CADY's principle of estimating values by fitting shared global trajectory dynamics. The result therefore provides a theoretical justification for global information sharing in value estimation, rather than a universal finite-sample guarantee for a practically fitted CHMM.

B Summary of Notation

Table 2 summarizes the meanings of the commonly used symbols in the main text.

C Experimental Details

C.1 Technical Details

CHMM Training Details For CookingWorld and TWC, both environments are deterministic and have finite action spaces. We therefore directly use executable actions as the observed CHMM elements. For multi-hop QA, we instead treat the full content returned by the search engine at each step as a single observed element. That is, identical retrieval results are mapped to the same element. In our experiments, the roughly 400 retrievals associated with each question typically correspond to about 80 distinct elements, making the resulting sequences suitable for CHMM learning. For more general open-ended text observations, semantic-similarity clustering or related discretization methods can be used to construct the element space and avoid having a large number of singleton elements. To improve the stability of EM fitting, we apply lightweight smoothing to the transition estimation using an EM pseudocount and a transition-memory coefficient, which regularize low-frequency transitions and reduce abrupt changes in the estimated transition dynamics across updates. Training is conducted on 8 NVIDIA A100 GPUs. For the baselines, a full training run

Table 2: Summary of notation.

Symbol	Meaning
t	Interaction step index.
T	Maximum number of interaction steps.
x	Task prompt.
h_t	Interaction history at step t.
a_t	All response content generated by the agent at interaction step t.
ρ_t	Reasoning process generated by the agent at interaction step t.
u_t	Executable action generated by the agent at interaction step t.
r_T	Terminal reward of an episode.
N	Number of rollout trajectories in a training batch.
o_t	Observation variable in the CHMM at interaction step t.
z_t	Latent state in the CHMM at interaction step t.
P	Transition matrix of the CHMM.
$q_t(z)$	Probability of reaching a `CORRECT` terminal state from CHMM state z at step t.
$\widetilde{A}_t^{\mathrm{CHMM}}$	Unnormalized CHMM-based advantage estimate.
$\widehat{A}_t^{\mathrm{CHMM}}$	Normalized CHMM-based advantage estimate.

takes approximately 10 hours for the 1.5B model and 30 hours for the 7B model. CADY incurs an additional wall-clock overhead of about 20% due to CHMM fitting, which could potentially be reduced through improved computational parallelism. Since each CHMM contains only a few hundred element types with roughly 10 clone states per element and can be discarded immediately after value estimation, it introduces negligible memory overhead relative to the LLM policy model.

Measuring Future-Dynamics Differences. For each pair of histories h_i and h_j, we compare sampled continuation trajectories containing L executable-action elements. For two such sequences $\tau = (u_1, \ldots, u_L)$ and $\tau' = (u'_1, \ldots, u'_L)$, the discounted mismatch is

$$d_{\gamma_f}(\tau, \tau') = \frac{\sum_{\ell=1}^{L} \gamma_f^{\ell-1} \mathbf{1}\{u_\ell \neq u'_\ell\}}{\sum_{\ell=1}^{L} \gamma_f^{\ell-1}}, \qquad 0 < \gamma_f < 1. \tag{33}$$

Here, each u_ℓ is a complete executable-action element rather than a language-model token, and earlier future steps receive larger weights. To account for variation among continuations from the same history, we report the baseline-corrected average mismatch

$$D_{\mathrm{future}}(h_i, h_j) = \mathbb{E}[d_{\gamma_f}(X, Y)] - \tfrac{1}{2}\mathbb{E}[d_{\gamma_f}(X, X')] - \tfrac{1}{2}\mathbb{E}[d_{\gamma_f}(Y, Y')], \tag{34}$$

where X, X' are independent continuations from h_i and Y, Y' are independent continuations from h_j. We estimate the first expectation using all cross-history rollout pairs, and the latter two using all distinct rollout pairs within each history. Thus, the correction removes the average within-history mismatch from the cross-history mismatch. We use $\gamma_f = 0.9$. Continuations with fewer than L elements are excluded rather than padded.

C.2 Raw CHMM Credit Scaling Mechanism

This subsection details the normalization used in Eq. 8. The goal is to control the absolute scale of the raw CHMM credits while preserving their relative magnitudes across interaction steps. This prevents the CHMM credit from becoming overly sensitive to batch-dependent scale variations and stabilizes its combination with the base advantage.

Let $\mathcal{I}$ denote the interaction steps with valid CHMM credits in the current training batch. For each $j \in \mathcal{I}$, let d_j denote the raw scalar CHMM credit and n_j the number of unmasked response tokens to which this credit is applied. We pool the credits across all eligible steps in the batch and compute a robust scale as

$$\widehat{s}_{\mathrm{CHMM}} = \max\left\{\varepsilon,\ \mathrm{wmed}_{j\in\mathcal{I}}\big(|d_j|; n_j\big)\right\}, \tag{35}$$

where $\varepsilon = 10^{-12}$ is used only for numerical stability. Here, $\mathrm{wmed}(|d_j|; n_j)$ denotes the weighted median of the credit magnitudes, where each interaction step is weighted by its number of active

Table 3: Main training configurations for CookingWorld, TWC, and Multi-hop QA.

Parameter	Value
Environment Configuration	
CookingWorld configuration	2 locations, 2 ingredients, 2 distractor items, no doors, and a limited inventory for the 1.5B model; 2 locations, 3 ingredients, 2 distractor items, no doors, and a limited inventory for the 7B model
TWC configuration	2 locations, 3 items to put away, no doors, and a limited inventory for the 1.5B model; 2 locations, 4 items to put away, no doors, and a limited inventory for the 7B model
Maximum interaction steps	CookingWorld and TWC: 50; Multi-hop QA: 6
RL Configuration	
Questions per training batch	CookingWorld and TWC: 8; Multi-hop QA: 16
Training group size	CookingWorld and TWC: 32; Multi-hop QA: 64
PPO mini-batch size	CookingWorld and TWC: 64; Multi-hop QA: 128
Maximum prompt length	CookingWorld and TWC: 2048; Multi-hop QA: 4096
Maximum response length per interaction	CookingWorld and TWC: 128; Multi-hop QA: 512
Actor KL-loss coefficient	CookingWorld and TWC: 0.01; Multi-hop QA: 0.001
Invalid-action penalty	CookingWorld and TWC: 0.1; Multi-hop QA: 0.001
ϵ	0.2
Actor learning rate	10^{-6}
Validation group size	8
Total training steps	300
CHMM Configuration	
Clone-count rule	Ceiling of the mean per-trajectory element frequency plus 3, clipped to 1–6 clones per element; for AutomationBench, we instead add 5 and clip to 1–12 clones per element
EM updates per training batch	5
EM pseudocount	10^{-6}
Transition-memory coefficient	0.6
α	1.0

response tokens n_j. Intuitively, a step whose credit is applied to twice as many tokens contributes twice as much weight when determining the batch-level scale. This weighting is consistent with the token-level policy objective, while the median makes the scale estimate robust to a small number of unusually large credits.

The normalization then proceeds in two stages. First, we divide each raw credit by the robust batch scale:

$$\bar{d}_j = \frac{d_j}{\widehat{s}_{\mathrm{CHMM}}}. \tag{36}$$

This step places the typical credit magnitude on a comparable numerical scale across training batches. Since all credits are divided by the same positive factor, their signs and relative magnitudes are preserved.

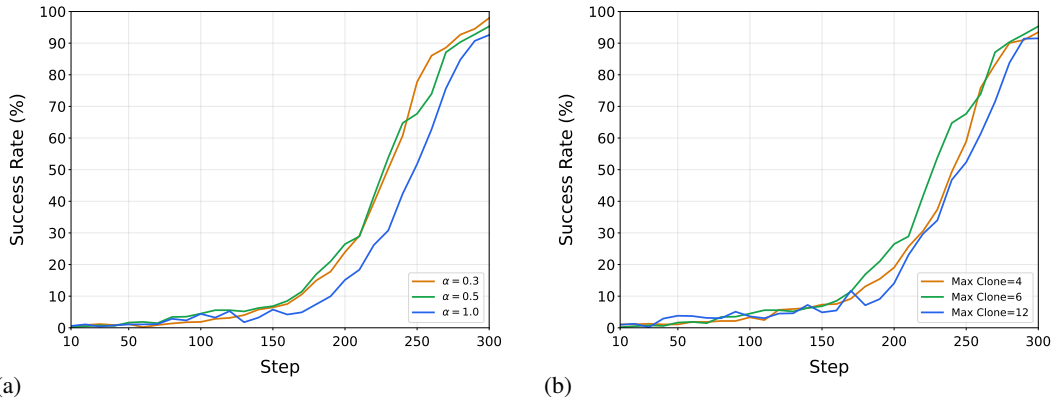
Second, we apply a smooth asinh transformation:

$$\widehat{A}_j^{\mathrm{CHMM}} = \mathrm{asinh}\big(\bar{d}_j\big) = \mathrm{asinh}\left(\frac{d_j}{\widehat{s}_{\mathrm{CHMM}}}\right). \tag{37}$$

The asinh function is approximately linear around zero, so differences among small and moderate credits are largely retained. For large magnitudes, it grows only logarithmically, preventing a few extreme credits from dominating the policy update. Together, the robust rescaling and smooth compression preserve informative step-level differences while keeping the CHMM credit at a stable scale throughout training.

C.3 Benchmark Introduction

CookingWorld. CookingWorld (Madotto et al., 2020) is a text-based embodied environment that requires an agent to complete a multi-step cooking task by following a given recipe. Specifically,



(a) (b)

Figure 5: Validation performance of CADY under different (a) α values and (b) maximum numbers of clone states.

the agent must read the recipe, navigate among different locations, collect the required ingredients, and execute the specified preparation and cooking operations in the correct order. The environment provides several parameters for controlling task difficulty. `numLocations` determines the number of accessible locations and hence the navigation complexity; `numIngredients` controls the number of ingredients required by the recipe; and `numDistractorItems` specifies the number of irrelevant ingredients that may distract the agent during exploration. `includeDoors` determines whether the agent must additionally open doors when moving between rooms, while `limitInventorySize` controls whether the number of objects carried at once is restricted. Increasing these factors generally requires longer and more carefully coordinated action sequences. In our setting, an episode can contain up to 50 interaction steps, making CookingWorld particularly demanding for long-horizon and temporally consistent decision making.

TextWorld Commonsense (TWC). TextWorld Commonsense (Murugesan et al., 2021) is a text-based embodied benchmark that requires an agent to complete household organization tasks using commonsense knowledge. Specifically, the agent must explore multiple rooms, locate relevant objects, infer their appropriate destinations, and place them correctly through a sequence of navigation and pick-and-place actions. Task complexity can be adjusted through several environment parameters. `numLocations` controls the number of locations that must be explored, while `numItemsToPutAway` determines how many objects need to be correctly relocated. `includeDoors` specifies whether doors must be opened before entering connected rooms, and `limitInventorySize` determines whether the agent can carry only a limited number of objects at a time. Larger environments and more objects substantially increase the number of possible interaction paths and the need for long-term planning. We allow up to 50 interaction steps, making TWC a challenging test bed for maintaining coherent decisions over long trajectories.

Multi-Hop Question Answering. We further consider a multi-turn retrieval-augmented question answering environment following Search-R1 (Jin et al., 2025), where the agent repeatedly searches a large-scale Wikipedia corpus, reasons over retrieved evidence, and produces the final answer. Multi-hop questions are particularly challenging because the evidence required for the final answer is distributed across multiple retrieval steps: information obtained from an earlier search often determines what should be queried next. The agent must therefore maintain a consistent reasoning state across successive search–reasoning interactions and correctly attribute the final outcome to intermediate retrieval decisions. We construct the training set by combining 800 examples from MuSiQue (Trivedi et al., 2022) with 200 examples from 2Wiki (Ho et al., 2020), and allow up to six interaction steps per episode. For evaluation, we use three single-hop question answering datasets—TriviaQA (Joshi et al., 2017), NQ (Kwiatkowski et al., 2019), and PopQA (Mallen et al., 2023)—and three multi-hop question answering datasets—HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023). Although shorter than the embodied tasks, each step involves an open-ended retrieval and reasoning decision, making the benchmark a complementary test of long-range credit assignment in tool-augmented agents.

AutomationBench. AutomationBench (Shepard & Salimans, 2026) evaluates how well AI models complete realistic business workflows—the kind of tasks that happen daily in sales, marketing, operations, support, finance, and HR. Each task places the agent in a simulated enterprise environment and requires it to autonomously discover relevant REST APIs, follow business policies, filter irrelevant or misleading information, and coordinate actions across applications such as CRM, email, calendars, spreadsheets, and messaging systems. The benchmark spans six business domains and 47 simulated applications, exposing roughly 500 API endpoints. In our setting, a trajectory allows up to 50 interaction steps and involves roughly 100 API calls. These long and highly interdependent tool-use sequences closely resemble practical business automation workflows, making AutomationBench a challenging and practically relevant testbed for evaluating whether CHMM can capture distinct future-dynamics patterns in complex agent trajectories.

C.4 Main Parameters

Table 3 summarizes the main training and CHMM configurations for CookingWorld, TWC, and Multi-hop QA.

C.5 Additional Results

Figures 5a and 5b show validation performance during training under different values of the hyperparameter α and different maximum numbers of clone states, respectively. CADY is relatively insensitive to both hyperparameters, with similar final validation accuracy across settings.