LEARNING FROM FUTURE INTERACTIONS IN ON-POLICY DISTILLATION FOR MULTI-TURN AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) provides dense teacher supervision for training large language model (LLM) agents on student-generated trajectories. However, each action shapes future observations and contexts, making next-token optimization alone *myopic* in multi-turn settings: actions with similar local supervision can lead to substantially different future interactions. We show that the trajectory-level reverse-KL objective intrinsically connects an action's supervision to its influence on future interactions. Teacher-student mismatch thus provides not only a local matching signal, but also hindsight credit for earlier actions. Based on this insight, we propose **FarOPD** (On-Policy Distillation with Future Action Rewards), which combines local-token supervision with a future-action reward derived from subsequent interactions. This makes action updates sensitive to future teacher-student mismatch, extending local imitation toward trajectory-level matching. Extensive experiments on three multi-turn agent benchmarks demonstrate consistent improvements over OPD baselines, with success rate gains of up to 8.4% relative to Vanilla OPD. These gains extend to multi-teacher distillation and are especially pronounced on tasks requiring longer interactions. Our code is available at https://anonymous.4open.science/r/FarOPD/.

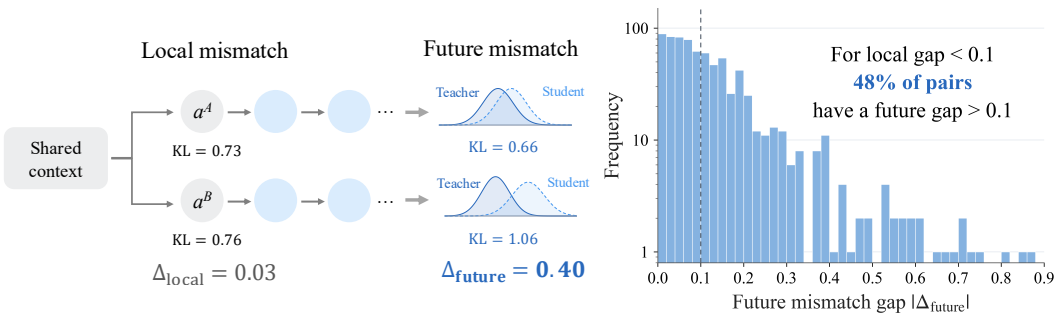
1 INTRODUCTION

Large language models (LLMs) are increasingly being deployed as multi-turn agents that solve tasks by interacting with external environments, such as tool use, web navigation, and embodied planning (Yao et al., 2022b; Xi et al., 2025a). Developing these capabilities through post-training has consequently become an important research direction (Zhang et al., 2025). Recently, on-policy distillation (OPD) has emerged as an effective post-training paradigm for LLMs (Xu et al., 2026a; Xiao et al., 2026). OPD is motivated by reverse-KL minimization on student-generated responses (Gu et al., 2024) and is commonly implemented with token-level log-ratio in current LLM pipelines (Lu & Lab, 2025; Team et al., 2026). These methods have achieved strong performance on single-turn reasoning, enabling both capability transfer from larger teachers to smaller students and integration from multiple domain teachers (Yang et al., 2026; Ma et al., 2026), and have also emerged as attractive approaches for training multi-turn agents (Wang et al., 2026; Zhong et al., 2026).

However, each action in a multi-turn interaction changes subsequent observations and thereby shapes the contexts for future decisions. Its influence can extend across the remaining interaction. The common practice of next-token optimization alone is therefore *myopic*, as it focuses on matching the teacher under the current context rather than accounting for how actions shape future interactions. As illustrated in Figure 1(a), two actions generated from the same context receive similar local mismatch yet lead to future interactions with substantially different teacher-student mismatch. Figure 1(b) further shows that such differences occur frequently among action branches with similar local mimatch. Consequently, OPD mainly teaches the student how to behave once it reaches a problematic context, rather than how to avoid earlier actions that make such a context likely in later turns. These observations raise a central question: *how can OPD supervise actions according to both their local teacher-student matching and their influence on the course of future interactions?*

In this work, we identify that the answer lies in the trajectory-level reverse-KL objective itself, where future teacher-student mismatch induces a hindsight advantage for earlier actions. Specifically, the gradient of reverse-KL could be decomposed into a local component that matches the teacher and

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.



(a) Similar local mismatch, different futures

(b) Distribution of future mismatch gaps

Figure 1: Myopic Credit Assignment of OPD in Multi-Turn Interaction. (a) Two candidate actions from the same context have similar teacher-student mismatch, yet their sampled continuations exhibit substantially different future mismatch. OPD provides supervision through token-level teacher–student matching. To compare candidate actions, we use local mismatch as an action-level summary, computed by averaging KL between studen and teacher over tokens in a response. Future mismatch averages these per-turn values over subsequent turns. (b) Among 764 continuation pairs from 191 action pairs with $|\Delta_{\text{local}}| < 0.1$, 48% have $|\Delta_{\text{future}}| > 0.1$. Results are from ALFWorld with a Qwen2.5-0.5B-Instruct student and an RL-trained Qwen2.5-7B-Instruct teacher.

student under the current context and a hindsight component that assigns mismatch from future interactions to earlier actions. In the latter component, an action's hindsight advantage measures how much it reduces expected future mismatch relative to the average over actions sampled from the student in the same context. This advantage establishes a direct connection between an action's influence on future contexts and the supervision it receives. Teacher-student mismatch therefore provides supervision for the action at which it is observed and intrinsic hindsight credit for the earlier actions whose influence on subsequent observations made that context likely to arise.

Building on this analysis, we propose FarOPD (On-Policy Distillation with Future-Action Reward), which enables teacher-student mismatch revealed in future interactions to supervise earlier actions. FarOPD preserves the local-token reward in standard OPD and introduces a future-action reward shared by all tokens in the same action. The local-token reward supervises current generation, while the future-action reward incorporates teacher-student mismatch from subsequent actions, translating the hindsight component identified by the gradient analysis into a practical training signal. To construct the future-action reward, we treat future actions as the basic units of supervision and first compute the mean token reward for each future action before averaging these action-level values, which also controls scale variation across action lengths and remaining horizons. Together, the local-token reward and future-action reward allow actions with similar local mismatch to receive different updates according to the teacher-student mismatch revealed in future interactions. FarOPD thereby extends local imitation of standard OPD toward trajectory-level matching.

We evaluate FarOPD on ALFWorld (Shridhar et al., 2020), WebShop (Yao et al., 2022a), and ScienceWorld (Wang et al., 2022) across single-teacher and multi-teacher distillation settings, observing consistent outperformance over other OPD methods, improving success rates by up to 8.4% over Vanilla OPD. Further analysis shows that the benefits are especially pronounced for tasks requiring longer interaction, with a 22.3% relative improvement in success rate for the 1.5B student on the longest task group in ScienceWorld. These results demonstrate the effectiveness of using teacher-student mismatch revealed in future interactions to supervise earlier actions.

In summary, our contributions are as follows: (1) We identify myopic credit assignment as a limitation of OPD in multi-turn settings and reveal the corresponding missing supervision structure:an intrinsic hindsight component through which future teacher-student mismatch provides credit for earlier actions. (2) We propose FarOPD, which equips OPD with a future-action reward that makes actions sensitive to teacher-student mismatch in future interaction. Furthermore, FarOPD employs action averaging to control scale variation across action lengths and remaining horizons. (3) We conduct extensive experiments across three datasets in single-teacher and multi-teacher distillation settings, demonstrating that FarOPD consistently improves OPD performance and showing that incorporating the future-action reward is especially beneficial for tasks requiring longer interactions.

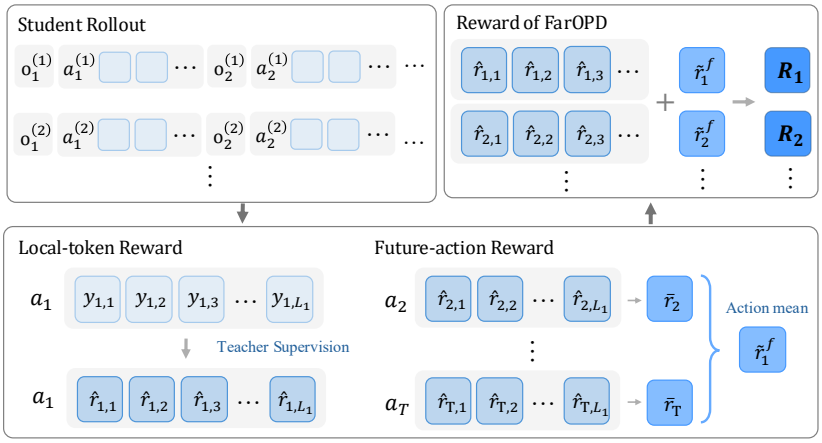


Figure 2: Overview of FarOPD. $\mathbf{R}_t$ collects rewards $\{\widehat{R}_{t,n}\}_{n=1}^{L_t}$.

2 Preliminaries

Multi-Turn Agent Interaction. Let $\mathcal{D}$ denote the task distribution and π_θ an LLM policy. Given a task $x \sim \mathcal{D}$, the agent interacts with an environment over T turns. At turn t, it receives an observation o_t and generates an action $a_t \sim \pi_\theta(\cdot \mid c_t)$, where $c_t = (x, o_1, a_1, \ldots, a_{t-1}, o_t)$ is the interaction context. Each action $a_t = (y_{t,1}, \ldots, y_{t,L_t})$ consists of L_t tokens. These tokens are generated autoregressively, with $\pi_\theta(a_t \mid c_t) = \prod_{n=1}^{L_t} \pi_\theta(y_{t,n} \mid h_{t,n})$, where $h_{t,n} = (c_t, y_{t,<n})$ denotes the prefix for predicting token $y_{t,n}$. The environment then returns the next observation o_{t+1}. This process produces a trajectory $\tau = (o_1, a_1, \ldots, o_T, a_T, o_{T+1})$, whose distribution under π_θ factorizes as

$$P_\theta(\tau \mid x) = \rho_0(o_1 \mid x) \prod_{t=1}^{T} \pi_\theta(a_t \mid c_t)\, \mathcal{T}(o_{t+1} \mid c_t, a_t), \tag{1}$$

where ρ_0 is the initial-observation distribution and $\mathcal{T}$ is the environment transition kernel. We write $\tau \sim P_\theta(\cdot \mid x)$ to indicate the trajectory τ is rolling out by π_θ in the environment according to Eq. (1).

On-Policy Distillation. Let π_T denote a fixed teacher policy and π_θ the student policy. On-policy distillation (OPD) samples trajectories from the current student policy and minimizes the reverse KL divergence between the student and teacher (Gu et al., 2024). In the multi-turn setting, this objective can be expressed as a sum of divergences evaluated on student-generated trajectories:

$$\mathcal{J}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{\substack{x \sim \mathcal{D} \\ \tau \sim P_\theta(\cdot \mid x)}} \left[\sum_{t=1}^{T} \sum_{n=1}^{L_t} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid h_{t,n}) \,\|\, \pi_T(\cdot \mid h_{t,n})) \right]. \tag{2}$$

Here, $h_{t,n}$ and $y_{t,n}$ are the prefixes and sampled tokens contained in the student-generated trajectory τ. In practice, the common implementations use a discount factor of zero, assigning each token only its own teacher–student log-ratio (Lu & Lab, 2025; Xiao et al., 2026; Team et al., 2026). With the sampled-prefix distribution treated as fixed, the gradient can be approximated as:

$$\nabla_\theta \mathcal{J}_{\mathrm{OPD}}(\theta) \approx -\mathbb{E}_{\substack{x \sim \mathcal{D} \\ \tau \sim P_\theta(\cdot \mid x)}} \left[\sum_{t=1}^{T} \sum_{n=1}^{L_t} \log \frac{\pi_T(y_{t,n} \mid h_{t,n})}{\pi_\theta(y_{t,n} \mid h_{t,n})} \nabla_\theta \log \pi_\theta(y_{t,n} \mid h_{t,n}) \right], \tag{3}$$

where $\log \pi_T(y_{t,n} \mid h_{t,n}) - \log \pi_\theta(y_{t,n} \mid h_{t,n})$ serves as a token-level advantage. This sampled log-ratio expresses teacher–student mismatch at the current token by comparing the probabilities assigned to that token under the same prefix, thereby providing dense token-level supervision.

3 Method

In this section, we present FarOPD, an on-policy distillation method that uses teacher–student mismatch from future interactions to supervise earlier actions. We first establish the hindsight advantage inherent in trajectory-level OPD, then construct a future-action reward through action averaging. Finally, we combine this reward with local-token supervision and characterize the resulting gradient estimator. Figure 2 provides an overview of FarOPD.

3.1 HINDSIGHT ADVANTAGE IN TRAJECTORY-LEVEL OPD

The update in Eq. (3) treats the prefix distribution as fixed, although earlier actions influence the contexts in which future teacher–student mismatch is evaluated. Let P_θ and P_{T} denote the trajectory distributions induced by the student and teacher policies, respectively, in the same environment. The trajectory-level reverse-KL objective, $\mathcal{J}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}[D_{\mathrm{KL}}(P_\theta(\cdot \mid x)\|P_{\mathrm{T}}(\cdot \mid x))]$, is equivalent to the OPD objective in Eq. (2) and can be written as

$$\mathcal{J}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\tau\sim P_\theta(\cdot|x)}\left[\sum_{t=1}^{T} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid c_t)\,\|\,\pi_{\mathrm{T}}(\cdot \mid c_t))\right]. \tag{4}$$

Each divergence compares the action distributions at a context reached by the student. Earlier actions influence which subsequent contexts are reached and thus where future mismatch is evaluated. This dependence is retained in the expectation over student trajectories. To quantify this influence, define the expected future mismatch as

$$\mathcal{M}_t(c,a) = \mathbb{E}\left[\sum_{j=t+1}^{T} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid c_j)\,\|\,\pi_{\mathrm{T}}(\cdot \mid c_j))\,\middle|\, c_t = c,\ a_t = a\right], \tag{5}$$

where the expectation averages over environment transitions and subsequent student actions after taking action a at context c. We define the **hindsight advantage** as

$$A_t^{\mathrm{H}}(c,a) := \mathbb{E}_{a'\sim\pi_\theta(\cdot|c)}[\mathcal{M}_t(c,a')] - \mathcal{M}_t(c,a), \tag{6}$$

where the first term is the average expected future mismatch over all potential actions at the same context. Thus, a positive advantage indicates that the action leads to lower expected future mismatch.

Theorem 1 (Intrinsic Hindsight Credit in OPD)**.** *Assume $T \le H < \infty$, a fixed teacher and environment, and a differentiable student policy π_θ with parameter-independent support. If the KL and gradient terms are integrable and differentiation commutes with expectation, then*

$$\nabla_\theta \mathcal{J}(\theta) = \underbrace{-\,\mathbb{E}\left[\sum_{t=1}^{T} \log\frac{\pi_{\mathrm{T}}(a_t \mid c_t)}{\pi_\theta(a_t \mid c_t)}\nabla_\theta \log \pi_\theta(a_t \mid c_t)\right]}_{\textit{local component}} \underbrace{-\,\mathbb{E}\left[\sum_{t=1}^{T} A_t^{\mathrm{H}}(c_t,a_t)\nabla_\theta \log \pi_\theta(a_t \mid c_t)\right]}_{\textit{hindsight component}}.$$

where $\mathbb{E}$ averages over $x \sim \mathcal{D}$ and $\tau \sim P_\theta(\cdot \mid x)$.

Theorem 1 establishes two roles for teacher–student mismatch within the same objective. The **local component** matches the student and teacher action distributions under the current context. For a token being updated, we call the token itself the *local token* and the later tokens in the same response the *within-action tokens*. The local component uses both the local token's own teacher–student log-ratio and those of the within-action tokens. The **hindsight component** connects an earlier action's update to the mismatch at the future contexts it induces, relative to the average over student actions at the same context. Crucially, this connection arises from the exact trajectory gradient. Teacher–student mismatch provides not only supervision at the context where it is observed, but also intrinsic hindsight credit for the earlier actions that shaped that context. This connection is absent from updates using only each token's own log-ratio. Detailed derivations are provided in Appendix A.

3.2 FUTURE-ACTION REWARD

To realize the hindsight component in Theorem 1, we construct a future-action reward from subsequent interactions in the sampled trajectory. We use the teacher–student log-ratio in Eq. (3) as the reward for each sampled token. In standard OPD, this reward directly serves as the token-level advantage, since no subsequent rewards enter the update. For the token being updated, its own reward is the local-token reward, whereas rewards from later tokens in the same response provide within-action supervision. The raw future-action reward instead sums rewards from subsequent actions:

$$\widehat{r}_t^{\mathrm{f}} = \sum_{j=t+1}^{T}\sum_{m=1}^{L_j} \widehat{r}_{j,m}, \qquad \widehat{r}_{j,m} = \log\frac{\pi_T(y_{j,m} \mid h_{j,m})}{\pi_\theta(y_{j,m} \mid h_{j,m})}. \tag{7}$$

Under on-policy continuation, $\mathbb{E}[\widehat{r}_t^{\mathrm{f}} \mid c_t, a_t] = -\mathcal{M}_t(c_t, a_t)$. Since the baseline depends only on c_t, its expected product with $\nabla_\theta \log \pi_\theta(a_t \mid c_t)$ is zero. Thus, the raw future-action reward recovers the expected advantage update without explicitly estimating the baseline. Derivation is in Appendix A.3.

Action averaging. The raw future-action reward sums over all future tokens, making its scale sensitive to response length and horizons. FarOPD instead treats future actions as the units of supervision. We first average token rewards within each future action and then average across future actions:

$$\widetilde{r}_t^{\mathrm{f}} = \frac{1}{T-t} \sum_{j=t+1}^{T} \frac{1}{L_j} \sum_{m=1}^{L_j} \widehat{r}_{j,m}, \qquad t = 1, \ldots, T-1, \qquad \widetilde{r}_T^{\mathrm{f}} = 0. \tag{8}$$

The inner average prevents longer responses from receiving more weight simply because they contain more tokens, while the outer average prevents the reward from accumulating with the number of remaining actions. Together, they control the reward scale and reduce its sensitivity to response length and the remaining horizon. Action averaging replaces cumulative token weighting with equal action weighting while preserving the hindsight attribution structure. The averaged reward yields the same expected action update as its corresponding hindsight advantage. Details are in Appendix A.4.

3.3 FarOPD Optimization

FarOPD combines the future-action reward with local-token supervision. The action-level hindsight update has a direct token-level implementation. Autoregressive factorization gives $\nabla_\theta \log \pi_\theta(a_t \mid c_t) = \sum_{n=1}^{L_t} \nabla_\theta \log \pi_\theta(y_{t,n} \mid h_{t,n})$. Consequently, multiplying the action score by $\widetilde{r}_t^{\mathrm{f}}$ is equivalent to assigning the same reward to every token in that action. The final token reward is

$$\widehat{R}_{t,n} = \widehat{r}_{t,n} + \widetilde{r}_t^{\mathrm{f}}. \tag{9}$$

The local-token reward $\widehat{r}_{t,n}$ supervises current generation, while the future-action reward $\widetilde{r}_t^{\mathrm{f}}$ incorporates evidence from subsequent interactions. We do not add the within-action reward-to-go: it assigns later-token credit within the same autoregressive response, whereas the additional supervision in FarOPD concerns the consequences of completed actions across environment transitions.

Training objective. At each iteration, the rollout policy π_{old} generates trajectories and provides the student probabilities used to compute the rewards. These rewards remain fixed during optimization. With the token-level importance ratio $\omega_{t,n}(\theta) = \pi_\theta(y_{t,n} \mid h_{t,n}) / \pi_{\mathrm{old}}(y_{t,n} \mid h_{t,n})$, we optimize

$$\widehat{\mathcal{L}}_{\mathrm{FAR}}(\theta) = -\mathbb{E}_{\tau \sim P_{\mathrm{old}}} \left[\sum_{t,n} \mathrm{sg}\left[\omega_{t,n}(\theta) \widehat{R}_{t,n} \right] \log \pi_\theta(y_{t,n} \mid h_{t,n}) \right], \tag{10}$$

where sg denotes stop-gradient and the expectation also includes $x \sim \mathcal{D}$. At the rollout policy, the negative gradient of this surrogate equals the expected local-plus-future update.

Proposition 2 (Bias–Variance Trade-off of Action Averaging)**.** *Let $\widehat{g}_{\mathrm{raw}}$ and $\widehat{g}_{\mathrm{FAR}}$ denote the on-policy estimators combining the same local-token reward with raw and averaged future-action rewards, respectively. Assume finite second moments. If the raw bias and the change in mean are bounded in Euclidean norm by b_W and ϵ, respectively, then*

$$\mathrm{MSE}(\widehat{g}_{\mathrm{FAR}}) - \mathrm{MSE}(\widehat{g}_{\mathrm{raw}}) \le \epsilon(2b_W + \epsilon) - \Delta\mathcal{V}, \tag{11}$$

where mean-squared error and bias are measured relative to $g^\star = -\nabla_\theta \mathcal{J}(\theta)$, and $\Delta\mathcal{V}$ is the raw estimator's total variance minus that of the averaged estimator.

Remark. Action averaging reweights future rewards and may increase bias, while controlling variation in reward scale caused by response length and the remaining horizon. The mean-squared error improves when the reduction in total variance exceeds the possible increase in squared bias, bounded above by $\epsilon(2b_W + \epsilon)$. The separate term b_W bounds the bias from omitting within-action supervision, which transfers rewards from later tokens to earlier tokens in the same response. Future-action supervision instead credits an action using teacher–student mismatch observed in subsequent interactions. Ablation in Table 3 further supports adding future-action reward to local-token reward over adding within-action reward or both. More analysis are available in Appendices A.5 and A.6.

Table 1: Single-teacher distillation results on WebShop, ALFWorld, and ScienceWorld. The best and second-best results are denoted in **bold** and underlined fonts, respectively.

Params	Method	WebShop		ALFWorld							ScienceWorld	
		Score	Succ.	Pick	Look	Clean	Heat	Cool	Pick2	Avg.	Score	Succ.
Teacher Models: RL-Tuned Qwen2.5-7B-Instruct												
7B	GiGPO	90.46	80.21	93.33	84.62	96.30	93.75	81.33	80.56	88.81	70.96	66.15
Student Models: Qwen2.5-Instruct Series												
0.5B	Zero-Shot	7.98	0.78	2.86	0.00	0.00	0.00	0.00	0.00	0.71	3.24	1.30
	Vanilla OPD	83.86	68.49	91.43	74.36	91.36	89.58	69.33	79.17	84.52	52.80	48.44
	TCOD-F2B	82.52	68.23	92.38	76.92	95.06	91.67	66.67	81.94	85.00	53.80	47.92
	TCOD-B2F	68.38	61.72	90.48	69.23	93.83	87.50	77.33	79.17	84.52	51.60	46.09
	SOD	84.43	71.35	**93.33**	74.36	88.89	91.67	76.00	76.39	84.76	53.88	47.14
	FarOPD	**84.85**	**74.22**	91.43	**82.05**	**97.53**	**93.75**	**82.67**	**88.89**	**90.00**	**57.45**	**52.08**
1.5B	Zero-Shot	21.68	3.13	8.57	7.69	7.41	0.00	0.00	4.17	5.00	6.01	2.86
	Vanilla OPD	84.02	73.44	96.19	74.36	91.36	91.67	74.67	80.56	86.43	60.90	56.25
	TCOD-F2B	85.57	74.22	96.19	76.92	96.30	89.58	**80.00**	83.33	88.57	59.95	55.21
	TCOD-B2F	79.40	71.61	96.19	82.05	92.59	87.50	77.33	88.89	88.81	59.97	55.47
	SOD	85.63	71.88	**100.00**	74.36	88.89	91.67	74.67	81.94	86.90	61.22	57.55
	FarOPD	**88.08**	**76.82**	97.14	**84.62**	**100.00**	**93.75**	76.00	**91.67**	**91.43**	**63.67**	**60.16**

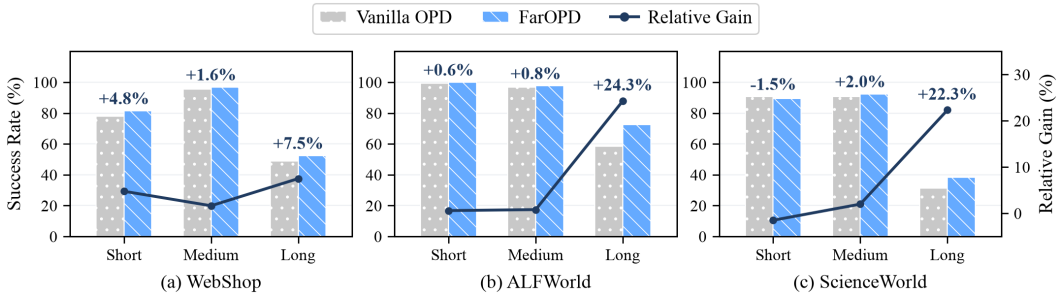


Figure 3: Performance across task horizons. Tasks are grouped using the 33rd and 67th percentiles of teacher trajectory length as cutoffs. Lines show FarOPD's relative gains over Vanilla OPD.

4 Experiments

We evaluate FarOPD on three benchmarks that require multi-turn interaction across different domains. Our experiments are designed to address the following key questions:

- **Q1**: How well does FarOPD improve agent performance in single-teacher and multi-teacher distillation? (Sections 4.2 & Section 4.3)
- **Q2**: Which reward components and averaging schemes contribute to its effectiveness? (Section 4.4)
- **Q3**: How does future-action reward evolve during training, and how sensitive is performance to its coefficient? (Sections 4.5 & Section 4.6)

4.1 Experimental Setup

Benchmarks and Metrics. We evaluate FarOPD on three benchmarks: **ALFWorld** (Shridhar et al., 2020) evaluates embodied instruction following in text-based household environments. *Avg.* denotes the success rate across all ALFWorld evaluation tasks; *Succ.* in the ablation tables denotes the same metric. **WebShop** (Yao et al., 2022a) evaluates goal-conditioned web navigation in an online shopping environment. *Score* denotes the task score scaled to 0–100, while *Succ.* denotes the percentage of episodes that achieve a perfect task score. **ScienceWorld** (Wang et al., 2022) evaluates scientific reasoning in a text-based environment. *Score* is the normalized environment score scaled to 0–100, and *Succ.* is the percentage of successful episodes. More details can be found in Appendix B.1.

Baselines. We compare against Vanilla OPD and two multi-turn OPD methods. Vanilla OPD uses the sampled-token log ratio as its distillation signal (Lu & Lab, 2025). TCOD progressively expand

Table 2: Multi-teacher distillation results on ALFWorld and WebShop.

Params	Method	WebShop		ALFWorld						
		Score	Succ.	Pick	Look	Clean	Heat	Cool	Pick2	Avg.
Teacher Models: RL-Tuned Qwen2.5-7B-Instruct										
7B	GiGPO (ALFWorld)	22.71	2.34	93.33	84.62	96.30	93.75	81.33	80.56	88.81
	GiGPO (WebShop)	90.46	80.21	25.71	46.15	14.81	12.50	12.00	8.33	18.57
Student Models: Qwen2.5-Instruct Series										
0.5B	Zero-Shot	7.98	0.78	2.86	0.00	0.00	0.00	0.00	0.00	0.71
	Vanilla OPD	82.03	66.67	93.33	71.79	88.89	89.58	70.67	80.56	83.81
	FarOPD	**87.50**	**71.09**	**94.29**	**76.92**	**90.12**	**91.67**	**76.00**	**87.50**	**87.14**
1.5B	Zero-Shot	21.68	3.13	8.57	7.69	7.41	0.00	0.00	4.17	5.00
	Vanilla OPD	85.54	74.22	96.19	64.10	95.06	**91.67**	74.67	86.11	86.90
	FarOPD	**87.68**	**76.30**	**97.14**	**76.92**	**96.30**	89.58	**78.67**	**87.50**	**89.29**

the rollout horizon, with F2B extending student rollouts from the initial state and B2F extending them backward from teacher-reached near-terminal states (Wang et al., 2026). SOD reweights each step's distillation signal based on the changes in teacher-student divergence (Zhong et al., 2026).

Implementation Details. We conduct experiments using the Qwen2.5 series models (Yang et al., 2024). The student models are initialized as Qwen2.5-0.5B-Instruct and Qwen2.5-1.5B-Instruct, while the teacher models are obtained by applying GiGPO (Feng et al., 2026) to Qwen2.5-7B-Instruct. We evaluate two settings: single-teacher and multi-teacher distillation. In the single-teacher setting, we train a separate student for each benchmark using its corresponding domain teacher. In the multi-teacher setting, one student is distilled on ALFWorld and WebShop using their respective domain teachers. For fair comparison, all OPD methods use the same training configurations, and all results are averaged over three random seeds. Full configuration details are in Appendix B.3.

4.2 Results of Single-Teacher Distillation

For **Q1**, Table 1 presents the single-teacher distillation results on ALFWorld, WebShop, and ScienceWorld. FarOPD consistently outperforms Vanilla OPD, TCOD, and SOD at both student sizes. Compared with Vanilla OPD, FarOPD improves ALFWorld success rates by 5.48% and 5.00% and WebShop success rates by 5.73% and 3.38% for the 0.5B and 1.5B students, respectively. On ScienceWorld, the corresponding success-rate gains are 3.64% and 3.91%, accompanied by task-score increases of 4.65 and 2.77. WebShop task scores also improve at both sizes. These results demonstrate the effectiveness of incorporating teacher-student mismatch from future actions along the interaction trajectory into the supervision of earlier actions when distilling multi-turn LLM agents.

To examine how these benefits vary with interaction horizon, we further examine how FarOPD performs across tasks with different interaction horizons. Figure 3 presents the success rates and relative gains for the 1.5B student. The largest relative gains occur in the Long group across all three benchmarks, reaching 24.3% on ALFWorld, 7.5% on WebShop, and 22.3% on ScienceWorld. Changes in the Short and Medium groups are comparatively small, including a slight decline in the Short group on ScienceWorld. Longer interactions allow earlier actions to influence more subsequent contexts, making the mismatch revealed in those contexts particularly relevant to their supervision. The stronger gains on long tasks support the value of future-action reward for capturing this influence.

4.3 Results of Multi-Teacher Distillation

For **Q1**, Table 2 shows that the benefit of future-action supervision extends to multi-teacher distillation. In this setting, one shared student is distilled from two domain-specific teachers, with each teacher supervising its corresponding domain. The teachers exhibit strong domain specialization. The ALFWorld teacher obtains 88.81% success on ALFWorld but only 2.34% on WebShop, whereas the WebShop teacher obtains 80.21% success on WebShop but only 18.57% on ALFWorld. Compared with Vanilla OPD, FarOPD improves ALFWorld success rates by 3.33% and 2.39% and WebShop success rates by 4.42% and 2.0% for the 0.5B and 1.5B students, respectively. WebShop task scores also increase by 5.47 and 2.14, respectively. These consistent gains across both domains show that future-action reward remains effective when the teacher varies across domains. Together with the single-teacher results, this supports the generality of FarOPD across teacher configurations.

Table 3: Ablation Study with Qwen2.5-0.5B-Instruct.

Dataset	Metric	Reward				Averaging		
		r	$r+\widetilde{r}^{\mathrm{w}}$	$r+\widetilde{r}^{\mathrm{w}}+\widetilde{r}^{\mathrm{f}}$	$r+\widetilde{r}^{\mathrm{f}}$	w/o	Token.	Action.
ALFWorld	Succ.	84.52	85.24	86.67	**90.00**	0.00	86.43	**90.00**
WebShop	Score	83.86	82.43	83.25	**84.85**	0.00	82.01	**84.85**
	Succ.	68.49	66.41	68.23	**74.22**	0.00	65.89	**74.22**

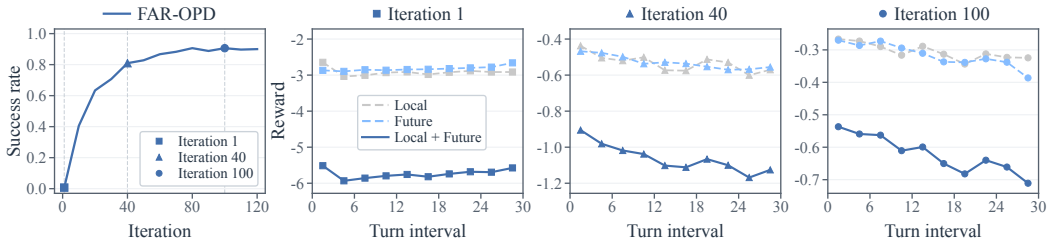


Figure 4: Dynamics of rewards during training on ALFWorld. Left: Success rate over training iterations. Squares, triangles, and circles denote the selected checkpoints (Iterations 1, 40, and 100). Right: Turn-wise profiles of local-token reward, future-action reward, and their sum (ours).

4.4 Ablation Study

For **Q2**, we assess the contributions of different reward components through ablation studies on the 0.5B student, as shown in Table 3. Here, $\widetilde{r}^{\mathrm{w}}_{t,n} = L_t^{-1} \sum_{m=n+1}^{L_t} \widehat{r}_{t,m}$ denotes the length-normalized within-action reward. Adding within-action reward (i.e., $r + \widetilde{r}^{\mathrm{w}}$) yields no consistent improvement, with a small gain on ALFWorld and decreases in both WebShop metrics. Combining local-token reward with future-action reward (i.e., $r + \widetilde{r}^{\mathrm{f}}$) achieves the best performance, exceeding the within-action variant by 4.76% and 7.81% in success rate on ALFWorld and WebShop, respectively. Adding future-action reward to the within-action variant (i.e., $r + \widetilde{r}^{\mathrm{w}} + \widetilde{r}^{\mathrm{f}}$) also improves all metrics, although using all three components remains less effective than FarOPD. Future-action reward improves performance with or without within-action reward. This pattern is consistent with their different scopes: future-action reward attributes mismatch in subsequent contexts to the action that influenced them, whereas within-action reward only propagates mismatch within the same response. Since within-action reward adds no further benefit, we retain local-token and future-action rewards in Eq. (9) to supervise both current generation and its influence on future interactions.

For **Q2**, we next examine how averaging affects the future-action reward for the 0.5B student by comparing three variants: (1) *W/o Averaging.* We directly sum all future token rewards, $\sum_{j>t} \sum_m r_{j,m}$, causing training to collapse completely, since the reward scale varies with response lengths and the remaining horizons. (2) *Token Averaging.* We average over all future tokens, $\frac{\sum_{j>t} \sum_m r_{j,m}}{\sum_{j>t} L_j}$, controlling the overall reward scale and restoring effective training. However, because each token is weighted equally, longer responses still contribute more to the supervision. (3) *Action Averaging.* We first average within each future action and then across actions, $\frac{1}{T-t} \sum_{j>t} \frac{1}{L_j} \sum_m r_{j,m}$, controlling the effects of both action length and remaining horizons. Action averaging achieves the best performance on both tasks in Table 3. By treating future actions as the units of supervision, action averaging makes such cross-action signal stable and effective.

4.5 Dynamics of Future-Action Reward

For **Q3**, we examine how future-action reward contributes during training by analyzing the 0.5B student on ALFWorld, as shown in Figure 4. Success rate increases to 90% as both reward components move toward zero. At Iteration 100, local-token rewards remain relatively similar across the later turn intervals, while future-action rewards become more negative toward the end of the interaction. The combined reward therefore distinguishes these turns more clearly, assigning lower rewards to actions followed by greater teacher-student mismatch. This distinction makes the updates sensitive to differences in future interactions even when local-token reward is similar. These dynamics demonstrate the effectiveness of assigning future teacher-student mismatch back to earlier actions, encouraging actions inducing lower future mismatch and improving task performance.

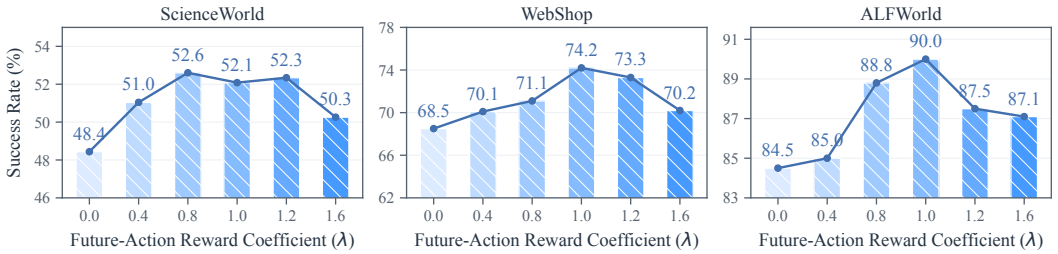


Figure 5: Sensitivity of future-action reward coefficients λ on ScienceWorld (left), WebShop (middle), and ALFWorld (right). $\lambda = 0$ recovers Vanilla OPD, $\lambda = 1$ is the default setting of FarOPD.

4.6 Sensitivity to Future-Action Reward Coefficient

For **Q3**, we examine the relative contribution of future-action supervision by introducing the future-action reward coefficient λ and formulating the token reward as $\widehat{R}_{t,n}(\lambda) = \widehat{r}_{t,n} + \lambda \widetilde{r}^{\mathrm{f}}_{t}$. Here, $\lambda = 0$ recovers Vanilla OPD, and $\lambda = 1$ is the default setting of FarOPD. Figure 5 shows that performance is highest near $\lambda = 1$ across three benchemarks. Small coefficients yield limited gains, while the decline at large coefficients suggests that excessive weighting of future-action reward disturbs its balance with local-token reward. All tested nonzero coefficients outperform Vanilla OPD, demonstrating the benefit of future-action reward is robust. Thus, $\lambda = 1$ is a broadly effective default, allowing FarOPD to incorporate future-action reward without additional hyperparameter tuning.

5 Related Work

LLM-based Multi-turn Agents. LLM-based multi-turn agents solve tasks by interleaving reasoning and actions with observations from environments (Yao et al., 2022b; Xi et al., 2025b). For instance, embodied tasks require these agents to adapt their decisions to the observations returned after each action (Shridhar et al., 2020; Wang et al., 2022). Recent reinforcement learning methods train such behavior from environment rewards and extend policy optimization to long interactions (Guo et al., 2025; Jin et al., 2025; Feng et al., 2026). However, long rollouts are costly, while sparse and delayed rewards make it difficult to identify which actions contribute to task outcomes (Zhang et al., 2025). These challenges motivate denser training signals on model-generated trajectories.

On-Policy Distillation. On-policy distillation (OPD) trains students on their own rollouts with dense teacher supervision (Agarwal et al., 2024; Gu et al., 2024). Subsequent work improves this supervision through token selection, signal weighting, and objective refinements (Xu et al., 2026b; Zheng et al., 2026; Jin et al., 2026; Ko et al., 2026). For multi-turn agents, existing methods improve teacher supervision through horizon curricula, step-wise reweighting, and turn-level guidance improve supervision reliability as errors accumulate across interactions (Wang et al., 2026; Zhong et al., 2026; Li et al., 2026). However, more reliable supervision doesn't address myopia: an action's influence on future contexts remains outside its learning signal. More recently, FTB incorporates such future influence by using subsequent student continuations to validate teacher interventions (Chen et al., 2026). FAR-OPD instead attributes future mismatch back to the earlier actions, thereby making action updates sensitive to their influence on future interactions rather than local imitation.

6 Conclusion

In this work, we presented FarOPD, an on-policy distillation method that accounts for how actions shape future interactions. Guided by our trajectory-level reverse-KL analysis, we developed future-action reward that uses teacher-student mismatch as hindsight credit alongside token-level supervision. Across three multi-turn agent benchmarks (ALFWorld, WebShop, ScienceWorld), FarOPD consistently outperforms OPD baselines at both student sizes. These gains extend to multi-teacher distillation and are particularly strong on longer-horizon tasks. The theoretical and empirical results highlight that teacher supervision can guide not only behavior within the current context, but also the actions that shape future interactions. This perspective motivates extending FarOPD to longer-horizon tasks and exploring more selective credit assignment across future actions in the future.

AI USE STATEMENT

We used generative AI tools to assist with manuscript polishing and code development. The authors reviewed and revised the AI-assisted text and checked and tested the AI-assisted code. We take full responsibility for the final manuscript, implementation, and reported results.

ETHICS STATEMENT

This work studies on-policy distillation for multi-turn language model agents. Our experiments use the publicly available ALFWorld, WebShop, and ScienceWorld benchmarks and do not involve new human-subject studies or the collection of personal information. The agents operate in benchmark environments. As distilled agents may inherit biases and unsafe behaviors from their teachers, deployment beyond these environments requires additional safety evaluation.

REPRODUCIBILITY STATEMENT

We describe the models, benchmarks, and evaluation metrics in Section 4. Appendix A provides mathematical derivations and the training algorithm, while Appendix B.3 documents the training framework and common experimental settings. These details support reproduction and verification of our method and experiments.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, et al. Look ahead before you distill: Future trajectory validation of teacher guidance for agentic on-policy distillation. *arXiv preprint arXiv:2608.01953*, 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.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The twelfth international conference on learning representations*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

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.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling reasoning efficiently via relaxed on-policy distillation. *arXiv preprint arXiv:2603.11137*, 2026.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, and Jinqiao Wang. On-policy distillation with curriculum turn-level guidance for multi-turn agents. *arXiv preprint arXiv:2606.15912*, 2026.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training. *arXiv preprint arXiv:2606.30406*, 2026. doi: 10.48550/arXiv.2606.30406. URL https://arxiv.org/abs/2606.30406.

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.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Jiaqi Wang, Wenhao Zhang, Weijie Shi, Yaliang Li, and James Cheng. TCOD: Exploring temporal curriculum in on-policy distillation for multi-turn autonomous agents. *arXiv preprint arXiv:2604.24005*, 2026.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022. URL https://aclanthology.org/2022.emnlp-main.775/.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, Qi Zhang, and Tao Gui. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025a. doi: 10.1007/s11432-024-4222-0. URL https://doi.org/10.1007/s11432-024-4222-0.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China information sciences*, 68(2):121101, 2025b.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026b.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/arXiv.2412.15115. URL https://arxiv.org/abs/2412.15115.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022b.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, et al. The landscape of agentic reinforcement learning for llms: A survey. *arXiv preprint arXiv:2509.02547*, 2025.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. SCOPE: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting. *arXiv preprint arXiv:2604.10688*, 2026.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. SOD: Step-wise on-policy distillation for small language model agents. *arXiv preprint arXiv:2605.07725*, 2026.

A Additional Method Details

This appendix proves Theorem 1 and Proposition 2, and connects their action-level quantities to the sampled token updates. Unless a rollout distribution is specified, expectations use the on-policy distribution in Section 3. The teacher subscript T in π_T denotes the teacher; the upper limit T denotes the realized number of interaction turns.

A.1 Trajectory Objective and Autoregressive Factorization

We use the normalized complete-action distributions of Section 2. An action includes its termination decision under the autoregressive factorization, so its length is part of the sampled outcome, not a constant when differentiating. A fixed truncation rule, if used, must be part of the same action convention for both policies. The trajectory likelihood in Eq. (1) uses fixed, shared initial-observation and transition laws. On student-supported trajectories, their factors therefore cancel:

$$\log \frac{P_\theta(\tau \mid x)}{P_T(\tau \mid x)} = \sum_{t=1}^{T} \log \frac{\pi_\theta(a_t \mid c_t)}{\pi_T(a_t \mid c_t)}. \tag{12}$$

Finiteness of the KL terms under the integrability assumptions of Theorem 1 implies absolute continuity, so the teacher denominator is positive almost surely under the student wherever the likelihood ratio is evaluated. Conditioning each action log-ratio on its reached context gives its action KL, $D_t(c_t)$. Taking expectations in Eq. (12) gives Eqs. (4) and (13).

$$\mathcal{J}(\theta) = \mathbb{E}\left[\sum_{t=1}^{T} D_t(c_t)\right]. \tag{13}$$

For the token connection, define the fixed-prefix divergence

$$d_{t,n}(h) \coloneqq D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid h) \,\|\, \pi_T(\cdot \mid h)\big). \tag{14}$$

The KL chain rule for the normalized autoregressive action law yields

$$D_t(c) = \mathbb{E}\left[\sum_{n=1}^{L_t} d_{t,n}(h_{t,n}) \,\middle|\, c_t = c\right]. \tag{15}$$

Here the expectation includes the random action length. One can first apply the chain rule to a finite token prefix, including the termination symbol, and then take the limit; the conditional KL contributions are nonnegative. Thus the chain-rule identity does not replace L_t by a deterministic length. Substituting Eq. (15) into Eq. (13) gives the token-level objective in Eq. (2).

A.2 Proof of the Hindsight Advantage Decomposition

For reference, the future mismatch defined in Section 3.1 is

$$\mathcal{M}_t(c, a) := \mathbb{E}\left[\sum_{j=t+1}^{T} D_j(c_j) \,\middle|\, c_t = c, a_t = a\right]. \tag{16}$$

Proof of Theorem 1. Use an absorbing terminal context to pad a trajectory to the deterministic bound H. After termination, set its divergence and score contributions to zero. The added action is deterministic for both policies and adds no likelihood-ratio term. This is a bookkeeping device: the distribution of the actual stopping time T is retained in the student-induced trajectory law. Write $S_t = \nabla_\theta \log \pi_\theta(a_t \mid c_t)$ on active turns and $S_t = 0$ otherwise. Define

$$C(\tau) \coloneqq \sum_{j=1}^{H} D_j(c_j), \qquad U(\tau) \coloneqq \sum_{t=1}^{H} S_t. \tag{17}$$

Because the initial and environment laws are fixed, U is the trajectory score. Differentiating the expectation of C under the theorem's interchange and integrability assumptions gives

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}\left[\sum_{j=1}^{H} \partial_\theta D_j(c_j)\right] + \mathbb{E}[C(\tau)U(\tau)]. \tag{18}$$

The partial derivative fixes the realized context, but differentiates the entire conditional action distribution. Parameter-independent support and differentiation of its normalization imply

$$\mathbb{E}[S_t \mid c_t] = \int \nabla_\theta \pi_\theta(a \mid c_t)\, \mathrm{d}a = 0. \tag{19}$$

The integral also denotes summation for discrete actions. On an inactive turn the identity holds because $S_t = 0$. For $j \le t$, the history in c_t determines $D_j(c_j)$, including whether that contribution is active. The conditional-expectation identity therefore gives $\mathbb{E}[D_j(c_j)S_t] = 0$. Expanding the finite product in Eq. (18) leaves only the future terms:

$$\mathbb{E}[CU] = \mathbb{E}\left[\sum_{t=1}^{H} S_t \sum_{j=t+1}^{H} D_j(c_j)\right]. \tag{20}$$

Conditioning on c_t, a_t and using Eq. (16) replaces the inner sum by $\mathcal{M}_t(c_t, a_t)$. Define $B_t(c)$ as the mean of $\mathcal{M}_t(c, a)$ over $a \sim \pi_\theta(\cdot \mid c)$. Equation (19) removes the context-only coefficient $B_t(c_t)$, and Eq. (6) then yields

$$\mathbb{E}[\mathcal{M}_t(c_t, a_t)S_t] = -\mathbb{E}[A_t^{\mathrm{H}}(c_t, a_t)S_t]. \tag{21}$$

Substitute this identity into Eq. (18) and remove the zero padded terms. The fixed-context KL derivative calculation in Section A.5 expresses $\partial_\theta D_t(c_t)$ as the negative expected product of the complete-action teacher–student log-ratio and S_t. Substituting that identity gives the gradient decomposition in Theorem 1, with its stated minus signs.

No token-local approximation was used in this argument. □

Zero conditional variance. For every reached context c, the same definitions imply

$$\mathrm{Var}_{a\sim\pi_\theta(\cdot|c)}[\mathcal{M}_t(c, a)] = 0 \quad \Longrightarrow \quad A_t^{\mathrm{H}}(c, a) = 0 \quad \pi_\theta(\cdot \mid c)\text{-a.s.} \tag{22}$$

To verify this consequence, zero action-conditional variance means that $\mathcal{M}_t(c, a)$ equals its action-conditional mean $B_t(c)$ almost surely under $\pi_\theta(\cdot \mid c)$. Their difference in Eq. (6) is therefore zero.

A.3 Score Identity and Sampled Future Rewards

The raw future-action reward introduced in Section 3.2 is

$$\widehat{r}_t^{\mathrm{f}} := \sum_{j=t+1}^{T} \sum_{m=1}^{L_j} \widehat{r}_{j,m}. \tag{23}$$

For any context-only $b(c_t)$ with an integrable score product, conditioning on c_t and applying Eq. (19) gives Eq. (24).

$$\mathbb{E}[b(c_t)S_t] = 0. \tag{24}$$

Even if b depends on the policy parameter, it is a coefficient in this score expectation; the identity does not take a further total derivative of that coefficient.

For clarity about existence, the sampled-score identities below are used when their displayed expectations and score products are integrable, and the indicated derivatives can be exchanged with conditional expectations. These conditions make the expected gradient contributions defined in Section A.5 well-defined and justify the sampled-score representations used here. A bounded turn horizon alone does not bound token lengths or establish these token-level moment conditions. We do not infer them from a bound on T, nor infer component second moments from second moments of a sum.

At a reached prefix, the definition of the sampled local reward gives

$$\mathbb{E}[\widehat{r}_{t,n} \mid h_{t,n}] = -d_{t,n}(h_{t,n}). \tag{25}$$

Indeed, summing the teacher-minus-student log probability against the student token probabilities is the negative token KL. Define the complete-action sampled reward by $q_j = \sum_{m=1}^{L_j} \widehat{r}_{j,m}$. Autoregressive factorization makes q_j the teacher-minus-student complete-action log probability, so

$\mathbb{E}[q_j \mid c_j] = -D_j(c_j)$. Apply the tower property to each active future turn, using zero padding to H, to obtain

$$\mathbb{E}\left[\sum_{j=t+1}^{T} q_j \,\middle|\, c_t, a_t\right] = -\mathcal{M}_t(c_t, a_t). \tag{26}$$

This establishes the conditional expectation identity stated in Section 3.2. Since S_t is determined by c_t, a_t, Eqs. (26) and (21) imply

$$\mathbb{E}[\widehat{r}_t^{\mathrm{f}} S_t] = \mathbb{E}[A_t^{\mathrm{H}}(c_t, a_t) S_t]. \tag{27}$$

This is an equality of score-weighted expectations. Neither a sampled raw reward nor its conditional mean is asserted to equal the centered hindsight advantage pointwise.

A.4 Action Averaging and Centering

Fix a complete realized continuation with $t < T$. Its realized lengths and number of remaining turns define the two-stage distribution in Eq. (28). For its allowed positions let $w_{t,j,m} = 1/((T-t)L_j)$.

$$\Pr(J = j, M = m \mid \tau) = \frac{1}{(T-t)L_j}, \qquad t < j \le T, \quad 1 \le m \le L_j. \tag{28}$$

The weight sum is

$$\sum_{j=t+1}^{T} \sum_{m=1}^{L_j} w_{t,j,m} = 1. \tag{29}$$

The action convention supplies at least one token per sampled action, including its termination token when appropriate, as required by the denominators in Eq. (8). Conditional on this fixed trajectory, averaging only over the auxiliary choices J, M gives

$$\mathbb{E}_{J,M}[\widehat{r}_{J,M} \mid \tau] = \sum_{j=t+1}^{T} \sum_{m=1}^{L_j} w_{t,j,m} \widehat{r}_{j,m}. \tag{30}$$

The right-hand side is exactly Eq. (8). Nonnegative weights with unit sum imply $|\widetilde{r}_t^{\mathrm{f}}| \le B$ whenever every contributing $|\widehat{r}_{j,m}| \le B$. At $t = T$ there is no future position to sample, and the reward is defined to be zero. The bound also holds in that case.

Let $\mu_t(c, a) = \mathbb{E}[\widetilde{r}_t^{\mathrm{f}} \mid c_t = c, a_t = a]$ and $\beta_t(c) = \mathbb{E}[\widetilde{r}_t^{\mathrm{f}} \mid c_t = c]$. The hindsight advantage corresponding to the averaged reward in Section 3.2 is denoted by $\widetilde{A}_t^{\mathrm{H}}$ and defined as their difference:

$$\widetilde{A}_t^{\mathrm{H}}(c, a) := \mu_t(c, a) - \beta_t(c). \tag{31}$$

Conditioning on c_t, a_t first and then removing the context-only coefficient $\beta_t(c_t)$ gives

$$\mathbb{E}[\widetilde{r}_t^{\mathrm{f}} S_t] = \mathbb{E}[\widetilde{A}_t^{\mathrm{H}}(c_t, a_t) S_t]. \tag{32}$$

The denominators $T - t$ and L_j are random before conditioning on τ. They have not been taken outside a trajectory expectation. This construction preserves centered action attribution, but it changes the weights of the raw cumulative return. Accordingly, $\widetilde{A}_t^{\mathrm{H}}$ generally differs from the raw-return advantage A_t^{H} in Eq. (6); Eq. (32) is the corresponding expected-update identity stated in Section 3.2.

A.5 Token Decomposition and Estimator Family

Define the reached-token score $s_{t,n} = \nabla_\theta \log \pi_\theta(y_{t,n} \mid h_{t,n})$. Differentiating the log of the autoregressive action probability gives Eq. (33); in this notation $S_t = \sum_{n=1}^{L_t} s_{t,n}$. The derivative is evaluated at the realized action, including its realized length.

$$\nabla_\theta \log \pi_\theta(a_t \mid c_t) = \sum_{n=1}^{L_t} \nabla_\theta \log \pi_\theta(y_{t,n} \mid h_{t,n}). \tag{33}$$

At a fixed prefix h, write $p_y = \pi_\theta(y \mid h)$ and $q_y = \pi_T(y \mid h)$. The product rule and fixed teacher give

$$\partial_\theta D_{\mathrm{KL}}(p \,\|\, q) = \sum_y \nabla_\theta p_y \log(p_y/q_y) + \sum_y \nabla_\theta p_y. \tag{34}$$

The last sum is zero by normalization. Using the sampled log-ratio as the coefficient consequently gives

$$-\partial_\theta d_{t,n}(h_{t,n}) = \mathbb{E}[\widehat{r}_{t,n} s_{t,n} \mid h_{t,n}]. \tag{35}$$

In contrast, multiplying the scalar negative KL by the same current token score gives zero conditional expectation, because that KL is already fixed by $h_{t,n}$. It cannot replace the sampled reward in Eq. (35).

Use the following within-action return notation solely for this appendix:

$$W_{t,n} \coloneqq \sum_{m=n+1}^{L_t} \widehat{r}_{t,m}. \tag{36}$$

The same fixed-context log-ratio calculation as in Eq. (34), now for a complete action, gives $-\partial_\theta D_t(c) = \mathbb{E}[q_t S_t \mid c_t = c]$. Expanding this product over realized tokens, a term with reward index $m < n$ vanishes: the earlier reward is determined by $h_{t,n}$, while $\mathbb{E}[s_{t,n} \mid h_{t,n}] = 0$. Reaching token n is determined before it is sampled, so this argument keeps the random action termination rule rather than treating L_t as fixed. Under the integrability and interchange conditions stated above, the remaining terms give

$$-\partial_\theta D_t(c) = \mathbb{E}\left[\sum_{n=1}^{L_t} (\widehat{r}_{t,n} + W_{t,n}) s_{t,n} \,\middle|\, c_t = c\right]. \tag{37}$$

For potentially unbounded token lengths, the expansion is understood by finite-prefix limits with the stated interchange conditions; the turn bound H is not used as a token bound.

Define the trajectory-sum estimators

$$\widehat{G}_{\mathrm{L}} \coloneqq \sum_{t=1}^{T} \sum_{n=1}^{L_t} \widehat{r}_{t,n} s_{t,n}, \qquad \widehat{G}_{\mathrm{W}} \coloneqq \sum_{t=1}^{T} \sum_{n=1}^{L_t} W_{t,n} s_{t,n}. \tag{38}$$

The two future estimators are

$$\widehat{G}_{\mathrm{F}} \coloneqq \sum_{t=1}^{T} \widehat{r}^{\,\mathrm{f}}_t S_t, \qquad \widehat{G}_{\bar{\mathrm{F}}} \coloneqq \sum_{t=1}^{T} \widetilde{r}^{\,\mathrm{f}}_t S_t. \tag{39}$$

Write $g_X = \mathbb{E}[\widehat{G}_X]$ for each component and $g^\star = -\nabla_\theta \mathcal{J}(\theta)$ for the exact descent direction. Equation (37) identifies $g_{\mathrm{L}} + g_{\mathrm{W}}$ with the negative of current-context matching in Theorem 1. Equation (27) identifies g_{F} with its hindsight descent contribution. Taking the negative of the gradient decomposition in Theorem 1 therefore proves Eq. (40).

$$g^\star = g_{\mathrm{L}} + g_{\mathrm{W}} + g_{\mathrm{F}}. \tag{40}$$

Equivalently, the complete token update has expectation

$$g^\star = \mathbb{E}\left[\sum_{t=1}^{T} \sum_{n=1}^{L_t} (\widehat{r}_{t,n} + W_{t,n} + \widehat{r}^{\,\mathrm{f}}_t) s_{t,n}\right]. \tag{41}$$

This supplies all three token contributions without identifying the fixed-context action derivative with only its local-token part.

To compare the local-plus-one-component choices, set $\widehat{G}_0 = 0$ and $g_0 = 0$. For $X \in \{0, \mathrm{W}, \mathrm{F}, \bar{\mathrm{F}}\}$ define

$$\widehat{g}^{(X)} \coloneqq \widehat{G}_{\mathrm{L}} + \widehat{G}_X, \qquad b^{(X)} \coloneqq g_X - g_{\mathrm{W}} - g_{\mathrm{F}}. \tag{42}$$

Here $b^{(X)} = \mathbb{E}[\widehat{g}^{(X)}] - g^\star$ is the bias vector. Thus the local-only choice omits both g_{W} and g_{F}, the local-plus-within choice omits g_{F}, and the raw local-plus-future choice omits g_{W}. The averaged

choice has the additional reweighting bias specified below. These vector identities do not imply that adding a component always reduces the norm of the bias; different omitted vectors can cancel.

For the following componentwise covariance expansion only, assume the involved $\widehat{G}_X$ have finite second moments. Define $V_X = \mathcal{V}(\widehat{G}_X)$ and

$$C_{XY} \coloneqq \mathbb{E}\langle \widehat{G}_X - g_X, \widehat{G}_Y - g_Y \rangle. \tag{43}$$

This is the trace of the cross-covariance matrix. Expanding the squared Euclidean norm of the centered sum gives, for every family member,

$$\mathcal{V}(\widehat{g}^{(X)}) = V_{\mathrm{L}} + V_X + 2C_{\mathrm{L}X}. \tag{44}$$

In particular, $\widehat{g}_{\mathrm{raw}} = \widehat{g}^{(\mathrm{F})}$ and $\widehat{g}_{\mathrm{FAR}} = \widehat{g}^{(\bar{\mathrm{F}})}$. Their variance difference is

$$\mathcal{V}(\widehat{g}_{\mathrm{FAR}}) - \mathcal{V}(\widehat{g}_{\mathrm{raw}}) = V_{\bar{\mathrm{F}}} - V_{\mathrm{F}} + 2(C_{\mathrm{L}\bar{\mathrm{F}}} - C_{\mathrm{LF}}). \tag{45}$$

For completeness, the full estimator $\widehat{g}^{\mathrm{full}} = \widehat{G}_{\mathrm{L}} + \widehat{G}_{\mathrm{W}} + \widehat{G}_{\mathrm{F}}$ has zero bias and total variance

$$\mathcal{V}(\widehat{g}^{\mathrm{full}}) = V_{\mathrm{L}} + V_{\mathrm{W}} + V_{\mathrm{F}} + 2(C_{\mathrm{LW}} + C_{\mathrm{LF}} + C_{\mathrm{WF}}). \tag{46}$$

No independence assumption was used. The covariance expansion makes no unconditional prediction that averaging reduces the actual variance. Its component moment assumption is not needed in the proof of Proposition 2, which compares the two complete estimators directly.

A.6 Proof of the Bias–Variance Bound

Proof of Proposition 2. Let $\delta_{\mathrm{avg}} = g_{\mathrm{F}} - g_{\bar{\mathrm{F}}}$, using the expected contributions from Section A.5. This difference measures the reweighting of the raw future contribution. The estimator means and Eq. (40) give

$$\mathbb{E}[\widehat{g}_{\mathrm{raw}}] - g^\star = -g_{\mathrm{W}}. \tag{47}$$

For the averaged estimator the same subtraction yields

$$\mathbb{E}[\widehat{g}_{\mathrm{FAR}}] - g^\star = -g_{\mathrm{W}} - \delta_{\mathrm{avg}}. \tag{48}$$

Subtracting Eq. (47) from Eq. (48) gives

$$\mathbb{E}[\widehat{g}_{\mathrm{FAR}}] - \mathbb{E}[\widehat{g}_{\mathrm{raw}}] = -\delta_{\mathrm{avg}}. \tag{49}$$

The two conditions in Proposition 2 are therefore equivalent to $\|g_{\mathrm{W}}\|_2 \le b_W$ and $\|\delta_{\mathrm{avg}}\|_2 \le \epsilon$, respectively. The triangle inequality and those bounds give

$$\|\mathbb{E}[\widehat{g}_{\mathrm{FAR}}] - g^\star\|_2 \le b_W + \epsilon. \tag{50}$$

For any estimator Z with finite second moment, write $Z - g^\star = (Z - \mathbb{E}Z) + (\mathbb{E}Z - g^\star)$. The expectation of its cross term vanishes since $\mathbb{E}[Z - \mathbb{E}Z] = 0$. Consequently,

$$\mathrm{MSE}(Z) = \mathcal{V}(Z) + \|\mathbb{E}Z - g^\star\|_2^2. \tag{51}$$

The assumptions of the proposition apply this identity separately to the complete raw and FAR estimators, without requiring finite second moments for their individual components. Let Δ_b denote the FAR squared bias minus the raw squared bias. Using their bias vectors gives

$$\Delta_b = 2\langle g_{\mathrm{W}}, \delta_{\mathrm{avg}} \rangle + \|\delta_{\mathrm{avg}}\|_2^2. \tag{52}$$

Cauchy–Schwarz, $\|g_{\mathrm{W}}\|_2 \le b_W$ and $\|\delta_{\mathrm{avg}}\|_2 \le \epsilon$ imply

$$\Delta_b \le \epsilon(2b_W + \epsilon). \tag{53}$$

Subtract Eq. (51) for the two estimators and substitute Eq. (53). With $\Delta\mathcal{V} = \mathcal{V}(\widehat{g}_{\mathrm{raw}}) - \mathcal{V}(\widehat{g}_{\mathrm{FAR}})$, this gives Eq. (11). Its right-hand side is strictly negative under the variance-gap sufficient condition stated after Proposition 2 in Section 3.3, proving the stated sufficient condition for smaller MSE. Both variances are total variances and retain every covariance term. The conclusion does not assert unconditional variance reduction, bias reduction, or improved task performance. □

A.7 Surrogate Objective and Training Algorithm

At each iteration let θ_{old} be the current parameters and freeze $\pi_{\text{old}} = \pi_{\theta_{\text{old}}}$. Tasks and trajectories are sampled from $\mathcal{D}$ and P_{old}, respectively. During reward construction the local reward is

$$\widehat{r}_{t,n} = \log \pi_T(y_{t,n} \mid h_{t,n}) - \log \pi_{\text{old}}(y_{t,n} \mid h_{t,n}). \tag{54}$$

This is the on-policy local reward evaluated at the rollout policy. It is stored and held fixed during optimization. Equations (8) and (9) then specify the double averaging and final token reward, including $\widetilde{r}^{\text{f}}_T = 0$. Equation (33) shows why multiplying the complete-action score by the future reward assigns that same reward to every token in the action.

The loss is exactly $\widehat{\mathcal{L}}_{\text{FAR}}$ in Eq. (10): its expectation is over the fixed rollout distribution, and its stop-gradient encloses the product $\omega_{t,n}(\theta)\widehat{R}_{t,n}$. Differentiation therefore acts only on the log student probability. At the rollout parameters the importance ratio equals one, giving

$$-\nabla_\theta \widehat{\mathcal{L}}_{\text{FAR}}(\theta_{\text{old}}) = \mathbb{E}_{\tau \sim P_{\text{old}}} \left[\sum_{t,n} \widehat{R}_{t,n} s_{t,n}(\theta_{\text{old}}) \right]. \tag{55}$$

As in the main text, the expectation also samples $x \sim \mathcal{D}$ and then $\tau \sim P_{\text{old}}(\cdot \mid x)$. Thus the surrogate's negative gradient equals the expected local-plus-averaged-future update at the rollout policy. Away from those parameters, a single-token importance ratio does not correct the distributions of earlier prefixes or entire trajectories; no exact trajectory-gradient recovery is claimed.

Algorithm 1 FarOPD Training

Input: Student π_θ, fixed teacher π_T, environment, task distribution $\mathcal{D}$, iterations N, learning rate α
for $k = 1, \ldots, N$ **do**
  Freeze $\theta_{\text{old}} \leftarrow \theta$ and $\pi_{\text{old}} \leftarrow \pi_{\theta_{\text{old}}}$
  Sample tasks $x \sim \mathcal{D}$ and trajectories $\tau \sim P_{\text{old}}(\cdot \mid x)$
  for each sampled trajectory **do**
    Store local rewards from teacher and rollout log probabilities using Eq. (54)
    For each action j, average its stored token rewards: $u_j \leftarrow L_j^{-1} \sum_{m=1}^{L_j} \widehat{r}_{j,m}$
    Set $\widetilde{r}^{\text{f}}_T \leftarrow 0$
    for $t = 1, \ldots, T-1$ **do**
      Average future action means: $\widetilde{r}^{\text{f}}_t \leftarrow (T-t)^{-1} \sum_{j=t+1}^{T} u_j$
    end for
    For every token set $\widehat{R}_{t,n} \leftarrow \widehat{r}_{t,n} + \widetilde{r}^{\text{f}}_t$
  end for
  Keep stored rewards and the rollout policy fixed during optimization
  Evaluate $\omega_{t,n}(\theta)$ against π_{old} and the loss in Eq. (10), with $\text{sg}[\omega_{t,n}(\theta)\widehat{R}_{t,n}]$
  Update $\theta \leftarrow \theta - \alpha \nabla_\theta \widehat{\mathcal{L}}_{\text{FAR}}(\theta)$
end for
return π_θ

Within-action ablation. The within-action reward used in Section 4.4 normalizes the raw return $W_{t,n}$ from Eq. (36) by the complete action length:

$$\widetilde{r}^{\text{w}}_{t,n} = \frac{W_{t,n}}{L_t}, \qquad \widetilde{r}^{\text{w}}_{t,L_t} = 0. \tag{56}$$

The two ablation variants use $\widehat{r}_{t,n} + \widetilde{r}^{\text{w}}_{t,n}$ and $\widehat{r}_{t,n} + \widetilde{r}^{\text{w}}_{t,n} + \widetilde{r}^{\text{f}}_t$, respectively. The exact gradient decomposition in Section A.5 uses the unnormalized $W_{t,n}$; the ablations use its length-normalized form above.

B Experimental Details

B.1 Benchmarks

We consider three interactive-agent benchmarks.

- **ALFWorld** (Shridhar et al., 2020) aligns the abstract text interface of TextWorld with household goals derived from ALFRED, allowing agents to act from textual observations through high-level language commands. It contains six task families: Pick & Place, Examine in Light, Clean & Place, Heat & Place, Cool & Place, and Pick Two & Place. These tasks require agents to locate objects and receptacles, satisfy object-state constraints, and complete multi-step household goals.
- **WebShop** (Yao et al., 2022a) is a simulated e-commerce website built from 1.18 million real-world products and 12,087 crowdsourced shopping instructions. Each instruction specifies a product requirement through combinations of attributes, options, and price constraints. Agents interact with search, result, item, and item-detail pages to formulate queries, compare products, choose options, and purchase an item that satisfies the instruction.
- **ScienceWorld** (Wang et al., 2022) is an interactive text environment for evaluating scientific reasoning through experimentation. It contains 30 tasks across 10 elementary-science topics, including changes of state, temperature measurement, electrical circuits, chemical mixtures, and biological processes. Agents navigate rooms, gather and manipulate objects, operate instruments, and observe the effects of their actions. Completing a task requires translating a scientific goal into a sequence of experiments and using the resulting observations to guide later decisions.

B.2 Baselines

We compare the initial student with three on-policy distillation methods that differ in how they apply teacher supervision during interaction.

- **Initial (Zero-Shot).** We directly evaluate the instruction-tuned Qwen2.5 student before task-specific training. This measures the starting agent capability and corresponds to the Zero-Shot rows in Table 1.
- **Vanilla OPD.** On-policy distillation trains on trajectories sampled by the student, with a frozen teacher scoring the generated tokens (Lu & Lab, 2025). Our sampled-token implementation uses the teacher–student log-probability difference as a local supervision signal. It applies guidance at student-visited contexts without step-wise reweighting or a temporal curriculum.
- **SOD.** Step-wise On-policy Distillation adjusts the strength of teacher supervision across interaction steps (Zhong et al., 2026). Changes in teacher–student divergence determine step-wise weights, allowing different turns to contribute differently to the distillation update. We use the reward-free variant without the GRPO objective.
- **TCOD.** Temporal Curriculum On-Policy Distillation progressively increases the interaction horizon used for training (Wang et al., 2026). We report both curriculum directions: TCOD-F2B extends student rollouts forward from the initial state, whereas TCOD-B2F starts from teacher-reached states near the end of a trajectory and gradually moves the starting point backward. The curriculum growth interval is $\eta = 2$.

B.3 Experiment Setups

We build the training pipeline on veRL-Agent (Feng et al., 2026). The students are Qwen2.5-0.5B-Instruct and Qwen2.5-1.5B-Instruct. Table 4 summarizes the common distillation and evaluation settings.

The benchmark-specific teachers are initialized from Qwen2.5-7B-Instruct and trained locally with GiGPO (Feng et al., 2026). Once trained, the teachers remain frozen throughout distillation.

All distillation methods use the same student initialization and benchmark-specific teacher, with common settings summarized in Table 4. Method-specific differences are specified in Appendix B.2.

Table 4: Distillation and evaluation settings across the three benchmarks.

Parameter	ALFWorld	WebShop	ScienceWorld
Models			
Student model	Qwen2.5-0.5B/1.5B-Instruct	Qwen2.5-0.5B/1.5B-Instruct	Qwen2.5-0.5B/1.5B-Instruct
Teacher model	GiGPO Qwen2.5-7B-Instruct	GiGPO Qwen2.5-7B-Instruct	GiGPO Qwen2.5-7B-Instruct
Environment			
Max prompt length	2,048	4,096	10,240
Max response length	512 per turn	512 per turn	1,024 per turn
Max turns	30	15	30
Rollout			
Rollouts per task	1	1	1
Top-p	1.0	1.0	1.0
Temperature	1.0	1.0	1.0
Evaluation			
Top-p	1.0	1.0	1.0
Temperature	0.4	0.4	0.4
Optimization			
Training steps	120	120	120
Optimizer	AdamW	AdamW	AdamW
Learning rate	2×10^{-6}	2×10^{-6}	2×10^{-6}
Warmup steps	0	0	0
Weight decay	0.01	0.01	0.01
Batch size	32	32	32
Mini-batch size	32	32	32
Policy update epochs	1	1	1
PPO clip range	0.2	0.2	0.2
Maximum gradient norm	1.0	1.0	1.0

C Environment Prompts

This section provides the detailed prompts used for ALFWorld, WebShop, and ScienceWorld during training and evaluation. All three environments use a consistent structure: the current task state, recent observation–action history when available, available actions (and objects in ScienceWorld), and the required thinking/action format. Braced fields are populated online from the environment; model-specific chat control tokens added by the tokenizer are omitted.

C.1 ALFWorld Prompts

ALFWorld requires navigation and object manipulation in household environments. The prompts ask the model to reason within `<think> </think>` tags and then return one admissible action within `<action> </action>` tags.

ALFWorld Prompt

You are an expert agent operating in the ALFRED Embodied Environment. Your task is to: {task_description}
Prior to this step, you have already taken {step_count} step(s). Below are the most recent {history_length} observations and the corresponding actions you took: {action_history}
You are now at step {current_step} and your current observation is: {current_observation}
Your admissible actions of the current situation are: [{admissible_actions}].
Now it's your turn to take an action.
You should first reason step-by-step about the current situation. This reasoning process MUST be enclosed within <think> </think> tags.

ALFWorld Prompt (continued)

Once you've finished your reasoning, you should choose an admissible action for current step and present it within `<action> </action>` tags.

C.2 WebShop Prompts

WebShop requires the agent to search and select products that satisfy a natural-language shopping instruction. The prompts expose the actions available on the current page and require the same thinking/action response format as in ALFWorld.

WebShop Prompt

You are an expert autonomous agent operating in the WebShop e-commerce environment.
Your task is to: {task_description}.
Prior to this step, you have already taken {step_count} step(s). Below are the most recent {history_length} observations and the corresponding actions you took: {action_history}
You are now at step {current_step} and your current observation is: {current_observation}.
Your admissible actions of the current situation are:
[{available_actions}].
Now it's your turn to take one action for the current step.
You should first reason step-by-step about the current situation, then think carefully which admissible action best advances the shopping goal. This reasoning process MUST be enclosed within `<think> </think>` tags.
Once you've finished your reasoning, you should choose an admissible action for current step and present it within `<action> </action>` tags.

Action Format for WebShop. The available-action list includes `search[<query>]` when a search bar is present and `click[<text>]` for each clickable element.

C.3 ScienceWorld Prompts

ScienceWorld requires the agent to complete scientific tasks in a text-based environment. The prompts provide available action commands and objects and ask the model to combine them into a valid action, using the same thinking/action response format as in ALFWorld and WebShop.

ScienceWorld Prompt

Your ScienceWorld task is: {task_description}
Prior to this step, you have already taken {step_count} step(s). Below are the most recent {history_length} observations and the corresponding actions you took: {action_history}
You are now at step {current_step} and your current observation is: {current_observation}
Your available actions and objects are:
{available_actions}
Now it's your turn to take an action. Combine an action command with appropriate object(s) to form a valid action.
You should first reason step-by-step about the current situation. This reasoning process MUST be enclosed within `<think> </think>` tags.

ScienceWorld Prompt (continued)

Once you've finished your reasoning, you should choose a valid action for the current step and present it within `<action>` `</action>` tags.