TIDE: TEACHER-STUDENT TRANSITION VIA INFORMATIVE DISTILLATION AND EXPLORATION FOR AGENTIC RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective multi-turn agents require interaction strategies that coordinate information gathering, actions, and feedback over long horizons. GRPO is a reinforcement learning algorithm used to train these agents, but sparse trajectory-level rewards limit early exploration in small models. Recent methods augment RL with on-policy distillation (OPD) from a stronger teacher. However, a fixed mixture assumes that teacher guidance and reward optimization should retain a constant relative role throughout training and across interaction turns. This assumption can fail at two scales. Globally, as training progresses, maintaining strong distillation pressure can constrain the model from moving beyond the teacher's capabilities. Locally, teacher–student disagreement identifies where the student departs from the teacher, but cannot tell whether that departure is exploration supported by better outcomes or low-quality policy drift. Our methodological insight is that teacher guidance and reward optimization should be dynamically rebalanced over training and jointly allocated across turns. We instantiate this insight in TIDE. Globally, TIDE uses the measured disagreement trend as a practical schedule signal, advancing an OPD-to-RL handoff when discrepancy reduction becomes slow but remains positive and progressively increasing the relative weight of RL. Locally, TIDE jointly modulates teacher-guided and reward-driven updates: relative action value and disagreement prioritize the OPD signal, whereas relative action value supplies the RL advantage and normalized disagreement reweights it across turns. Coupled with the global handoff, TIDE allocates stronger teacher guidance early and gives reward-driven updates greater relative weight later in training. Experiments across multiple benchmarks, student scales, and controlled ablations support the effectiveness of TIDE's adaptive OPD–RL coordination.

1 INTRODUCTION

Language-model agents have emerged as a paradigm for long-horizon tasks such as web navigation and embodied control (Yao et al., 2022; Shridhar et al., 2020). In these tasks, actions alter subsequent states but may reveal consequences only after many turns, making effective policy learning from sparse outcome rewards difficult and motivating richer supervision.

GRPO (Shao et al., 2024) and on-policy distillation (OPD) (Agarwal et al., 2024) provide complementary supervision for learning long-horizon interaction policies. GRPO optimizes the policy through group-relative comparisons of trajectory-level task rewards, whereas OPD provides token-level guidance by distilling a stronger teacher's distribution on student-generated trajectories.

Yet combining OPD with RL raises two allocation questions: how should teacher supervision be scheduled across training, and how should learning signals be allocated within each trajectory? At the global scale, a fixed-weight mixture cannot accommodate changing supervision utility. As training progresses, the marginal utility of continued OPD may diminish (Tan et al., 2026; Yu et al., 2026); retaining a large OPD coefficient in this latter phase can constrain reward-driven departures from the teacher (Yu et al., 2026). At the local scale, student rollouts can enter low-quality regions that depart from the teacher distribution, where teacher-provided distillation signals may become

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.

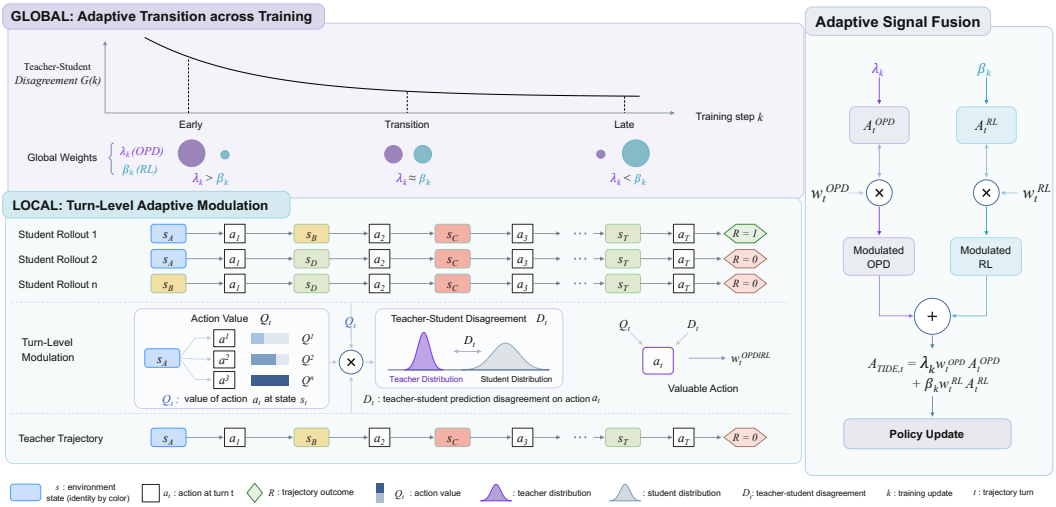


Figure 1: Overview of TIDE. The global handoff shifts relative weight from OPD to RL during training, while local modulation allocates their signals across turns.

less reliable (Zheng et al., 2026a; Lin et al., 2026). Conversely, some departures from the teacher distribution can be beneficial and yield higher returns (Liu et al., 2026; Wang et al., 2026a).

We address this allocation problem with *Teacher–Student Transition via Informative Distillation and Exploration* (TIDE), illustrated in Figure 1. Globally, TIDE uses batch-level teacher–student disagreement as a practical event signal. When its positive reduction becomes slow, TIDE gradually and monotonically shifts optimization weight from OPD to RL, adapting transition timing to each task's observed training dynamics without manually specifying a task-specific handoff step. Each transition increment reduces the OPD coefficient and increases the RL coefficient. Locally, TIDE modulates learning signals across turns. Building on GiGPO, it derives a relative action value for each turn. Its trajectory-relative priority is combined with disagreement to prioritize teacher guidance, whereas its value is scaled by disagreement to form reward-driven updates. Combined with the global controller, these local signals allocate teacher guidance and reward-driven updates to turns according to their relative value and teacher–student disagreement.

Experiments on WebShop and ALFWorld assess TIDE's two-scale design. At 1.5B, TIDE achieves 77.2% WebShop success and 86.0% ALFWorld average success; at 3B and 7B, it attains the highest WebShop success among the evaluated baselines. Controlled ablations and analyses further support the effectiveness of the global handoff and local modulation.

Overall, our contributions are as follows:

- We formulate hybrid OPD–RL coordination for multi-turn agents as a two-scale allocation problem: scheduling teacher influence across training stages and allocating update priority across turns within each trajectory.
- We propose TIDE, which uses a disagreement-triggered schedule to set a phase-wise OPD–RL balance and combines relative action value with disagreement as a turn-priority signal.
- We evaluate TIDE on WebShop and ALFWorld across student scales and controlled ablations, supporting its two-scale OPD–RL coordination for interactive-agent training.

2 Method

2.1 Problem Formulation

Multi-turn agent interaction. Given a task $x \sim \mathcal{D}$ and an initial observation o_1, the agent repeatedly reasons, acts, and receives environment feedback. At turn t, its interaction history $h_t = (x, o_1, y_1, a_1, \ldots, o_t)$ contains the task and all preceding responses, executed actions, and

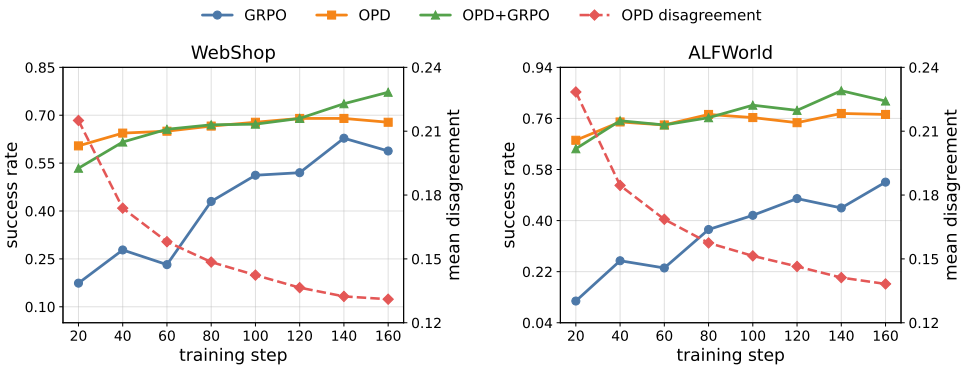


Figure 2: Training trajectories on WebShop and ALFWorld. Solid curves show task success for GRPO, OPD, and OPD+GRPO; dashed curves show teacher–student disagreement for OPD.

Table 1: Turn-level diagnostics of WebShop trajectories generated with OPD, grouped by teacher–student disagreement and GPT-5.5 process quality.

Group	Trajectory success (%)	GPT-5.5 process quality	Teacher–student disagreement	Action Distribution (%)				
				Search Items	Open Item	Select Option	Go Back	Buy Item
All	44.9	2.0	0.131	18.5	24.9	39.0	10.6	6.8
G_{low}	56.7	2.2	0.069	55.7	6.5	26.7	1.9	9.2
G_{high}	36.8	1.9	0.189	3.3	37.6	43.2	11.4	4.5
Q_{low}	0.0	1.0	0.134	19.9	26.5	36.0	15.0	2.3
Q_{high}	100.0	3.0	0.123	17.2	21.0	44.3	5.3	12.2
$G_{\text{high}} \cap Q_{\text{low}}$	0.0	1.0	0.189	3.7	39.6	40.0	14.2	2.5
$G_{\text{high}} \cap Q_{\text{high}}$	100.0	3.0	0.188	2.2	28.0	53.8	8.8	7.1

observations. The policy π_θ generates a textual response $y_t \sim \pi_\theta(\cdot \mid h_t)$, consisting of reasoning followed by an environment-specific command. A task-specific parser extracts the executable action a_t from y_t, and $\mathcal{E}$ returns the next observation o_{t+1}. We write the resulting turn as $u_t = (y_t, a_t, o_{t+1})$ and the complete T-turn rollout as $\tau = (x, o_1, u_1, u_2, \ldots, u_T)$. The interaction terminates when the agent submits a final answer, the environment signals completion, or the turn budget is exhausted.

2.2 Preliminary Studies on OPD–RL Coordination

Training-stage variation in the OPD–RL balance. To examine how the relative utility of teacher supervision and reward optimization changes over training, we compare the task success of GRPO, OPD with a task-trained 7B teacher, and a fixed 1:1 weighting of OPD and RL on WebShop and ALFWorld, while tracking the average teacher–student log-probability disagreement.

In these controlled runs, OPD yields higher initial task success than GRPO on both benchmarks, but its success later plateaus or fluctuates even as teacher–student disagreement continues to fall. Combining the two objectives achieves better performance, indicating that their relative utility changes over training: stronger teacher guidance can stabilize early learning, whereas reward-driven optimization becomes more useful for later refinement. Because this transition occurs at different rates across tasks, a fixed mixing ratio cannot adapt to the changing balance.

Turn-level disagreement and externally assessed process quality. Teacher–student disagreement alone identifies response-level policy mismatch but does not indicate the quality of a mismatch. We analyze WebShop trajectories generated with OPD along two axes: teacher–student disagreement, computed from log-probability differences over student-sampled response tokens, and an independent GPT-5.5 process-quality score assigned to each turn. In Table 1, G_{low} and G_{high} denote the bottom and top 20% of turns by disagreement, and Q_{low} and Q_{high} the bottom and top 20% by GPT-5.5 process quality. Details of the quality-scoring rubric are provided in Appendix A.3.

In these trajectories, response-level disagreement is distributed differently across parsed action types: low-disagreement turns are dominated by routine search actions (55.7%), whereas high-

disagreement turns are concentrated on opening items and selecting options (37.6% and 43.2%). Within the high-disagreement subset, however, the bottom- and top-quality turns have virtually identical disagreement (0.189 and 0.188) but opposite GPT-5.5 scores and outcomes. This external diagnostic shows that response-level disagreement identifies policy mismatch but cannot determine whether a departure is exploration supported by better outcomes or low-quality policy drift.

2.3 Teacher–Student Transition via Informative Distillation and Exploration

Guided by these findings, TIDE combines OPD and RL at two granularities. For token j in the student-sampled response y_t at interaction turn t, we first measure the teacher–behavior log-probability difference:

$$\delta_{t,j} = \log \pi_T(y_{t,j} \mid h_t, y_{t,<j}) - \log \pi_{\text{old}}(y_{t,j} \mid h_t, y_{t,<j}). \tag{1}$$

We then aggregate its absolute value over the complete student-sampled response and across the update batch to obtain turn- and batch-level discrepancy:

$$D_t = \frac{1}{|y_t|} \sum_{j=1}^{|y_t|} |\delta_{t,j}|, \qquad G_k = \frac{1}{|\mathcal{U}_k|} \sum_{t \in \mathcal{U}_k} D_t. \tag{2}$$

Thus, D_t is a response-level token discrepancy, including reasoning and executable-command tokens, rather than a divergence between full action distributions.

Global adaptation: discrepancy-triggered handoff. The global controller uses the observed discrepancy trend as an event signal for a monotone transition from OPD to RL. It determines transition timing online from the observed training dynamics of each task, without manually specifying a task-specific handoff step. To attenuate batch-level noise, we smooth G_k with an exponential moving average m_k and measure its relative reduction I_k over a window of W updates:

$$m_k = \mu m_{k-1} + (1-\mu) G_k, \qquad I_k = \frac{m_{k-W+1} - m_k}{\max(m_{k-W+1}, \epsilon)}, \tag{3}$$

The EMA makes the event rule depend on a persistent trend rather than a single noisy batch. When I_k is large, the controller leaves the handoff state unchanged; when $0 < I_k \le \zeta$, the discrepancy is decreasing slowly but remains positive, and the controller advances the RL-favored phase:

$$r_k = \begin{cases} \min(r_{k-1} + \eta, 1), & 0 < I_k \le \zeta, \\ r_{k-1}, & \text{otherwise.} \end{cases} \tag{4}$$

Each trigger advances the handoff state by the bounded increment η, yielding a gradual transition to RL. The handoff state determines the weights of the two branches:

$$\lambda_k = \lambda_{\min} + (\lambda_{\max} - \lambda_{\min})(1 - r_k), \qquad \beta_k = \beta_{\min} + (\beta_{\max} - \beta_{\min}) r_k. \tag{5}$$

As training progresses, the observed discrepancy trend gradually decreases the OPD coefficient and increases the RL coefficient.

Local adaptation: relative-value–disagreement modulation. We use a relative process-reward signal, Q_t, to indicate whether an action leads to a better or worse outcome than matched alternatives. Within each matched-history group, its magnitude reflects the extent of this difference. Its detailed computation, following GiGPO (Feng et al., 2026), is provided in Appendix A.4.

Within each trajectory τ, we map relative process value Q_t and turn-level disagreement D_t to comparable priorities, and combine them into an OPD turn-priority score. Let $\mathcal{V}_\tau = \{t \in \tau : \mathcal{H}_t \neq \emptyset\}$ denote turns with at least one value-discriminative matched-history group, i.e., a group with another member and nonzero return variation:

$$\tilde{Q}_t = \mathcal{N}_{\mathcal{V}_\tau}(Q_t), \qquad \tilde{D}_t = \mathcal{N}_\tau(D_t), \qquad z_t = \tilde{Q}_t \times \tilde{D}_t. \tag{6}$$

For the value signal, normalization is computed only over valid turns; let $\Delta_{\mathcal{V}_\tau}(Q)$ denote the range of Q_t over $\mathcal{V}_\tau$:

$$\mathcal{N}_{\mathcal{V}_\tau}(Q_t) = \begin{cases} \dfrac{Q_t - \min_{u \in \mathcal{V}_\tau} Q_u}{\max_{u \in \mathcal{V}_\tau} Q_u - \min_{u \in \mathcal{V}_\tau} Q_u + \epsilon}, & t \in \mathcal{V}_\tau,\ \Delta_{\mathcal{V}_\tau}(Q) > \epsilon, \\ 1, & \text{otherwise,} \end{cases} \tag{7}$$

For disagreement, we use the standard min–max normalization over all turns,

$$\mathcal{N}_\tau(D_t) = \begin{cases} \dfrac{D_t - \min_{u\in\tau} D_u}{\max_{u\in\tau} D_u - \min_{u\in\tau} D_u + \epsilon}, & \Delta_\tau(D) > \epsilon, \\ 1, & \Delta_\tau(D) \le \epsilon. \end{cases}$$

A turn without a value-discriminative matched group retains $Q_t = 0$ and uses $\tilde{Q}_t = 1$ as a disagreement-only OPD fallback. The same fallback applies when valid turns have no discriminative action values. Otherwise, normalization maps the signals to non-negative priorities while preserving their ordering. We use this score for OPD, with $w_t^{\mathrm{OPD}} = z_t$. For the RL branch, normalized disagreement is the turn weight. Thus, relative action value determines the direction of reward-driven updating, while disagreement determines its magnitude:

$$w_t^{\mathrm{OPD}} = z_t, \qquad w_t^{\mathrm{RL}} = \tilde{D}_t. \tag{8}$$

Consequently, high-value, high-disagreement turns receive strong positive RL updates, whereas low-value, high-disagreement turns receive strong negative RL updates. The OPD and combined advantages are defined as

$$\begin{aligned} A_{t,j}^{\mathrm{RL}} &= Q_t, \qquad A_{t,j}^{\mathrm{OPD}} = \delta_{t,j}, \\ A_{t,j}^{\mathrm{TIDE}} &= \beta_k w_t^{\mathrm{RL}} A_{t,j}^{\mathrm{RL}} + \lambda_k w_t^{\mathrm{OPD}} A_{t,j}^{\mathrm{OPD}}. \end{aligned} \tag{9}$$

We optimize the advantage with the clipped surrogate objective:

$$\mathcal{J}(\theta) = \mathbb{E}\Big[\sum_{t,j} \min\big(\rho_{t,j}\, A_{t,j}^{\mathrm{TIDE}},\ \mathrm{clip}(\rho_{t,j}, 1-c, 1+c)\, A_{t,j}^{\mathrm{TIDE}}\big)\Big], \tag{10}$$

Here, Q_t, w_t^{RL}, and w_t^{OPD} are turn-level quantities shared by all tokens in y_t, whereas $\delta_{t,j}$ is token-level. The importance ratio $\rho_{t,j}$ is between the current and behavior policies:

$$\rho_{t,j} = \frac{\pi_\theta(y_{t,j} \mid h_t, y_{t,<j})}{\pi_{\mathrm{old}}(y_{t,j} \mid h_t, y_{t,<j})} \tag{11}$$

The local signals allocate teacher and reward updates across turns, while the global schedule determines their phase-specific OPD–RL balance. Coupled with the global handoff, TIDE provides more reliable distillation signals during the OPD-weighted early phase and promotes high-value exploration during the RL-weighted later phase.

3 Experiments

3.1 Experimental Setup

Benchmarks and metrics. We evaluate TIDE on WebShop (Yao et al., 2022) and ALFWorld (Shridhar et al., 2020). On the full 500-task WebShop evaluation set, we report exact success rate and normalized task score, using up to 30 interaction steps and temperature 1.0 with top-p 1.0 decoding. Main ALFWorld results use the official valid-seen split (140 tasks), while controlled ablations and sensitivity analyses additionally report valid-unseen results; both splits use identical task prompts, a two-step interaction history, at most 50 environment steps, temperature 0.4, and top-p 1.0 decoding. For every method evaluated by us and for every controlled ablation, each "$\pm$" denotes the sample standard deviation across n=3 independently trained random seeds. Appendix B.1 also reports a SearchQA experiment.

Models and baselines. We train Qwen2.5-Instruct students at the 1.5B, 3B, and 7B scales (Qwen et al., 2025). Table 2 groups the compared methods by training signal: Vanilla is a prompting baseline; GRPO and GiGPO (Feng et al., 2026) are reward-optimization baselines; OPD and OPSD are distillation references; and ATOD (Tan et al., 2026) and SDAR are hybrid OPD–RL baselines. At every scale, OPD, ATOD, and TIDE use the same frozen task-trained Qwen2.5-7B teacher and matched training configuration. Full baseline descriptions are in Appendix A.

Table 2: Main results across model scales on ALFWorld and WebShop (%). Prior-work values are marked †. Bold indicates the highest displayed value in each column at each scale.

Method	Type	ALFWorld							WebShop	
		Pick	Look	Clean	Heat	Cool	Pick2	Avg.	Score	SR
Qwen2.5-1.5B-Instruct										
Vanilla†	Prompt	11.1	0.0	6.2	0.0	0.0	4.2	5.5	17.8	5.5
GRPO	RL	80.0 ±2.9	53.8 ±7.7	50.6 ±4.3	37.5 ±6.3	68.0 ±4.0	44.4 ±4.8	58.8 ±0.8	82.4 ±1.2	62.8 ±2.0
GiGPO†	RL	94.4	67.5	94.8	**94.4**	79.8	**76.4**	**86.7**	83.1	65.0
OPD	Distill	88.6 ±5.0	66.7 ±4.4	84.0 ±4.0	72.9 ±5.1	61.3 ±4.7	56.9 ±5.5	73.6 ±2.1	79.4 ±2.3	67.8 ±3.1
OPSD†	Distill	26.3	16.7	9.1	6.7	9.1	5.3	14.1	22.3	10.2
ATOD	Hybrid	88.6 ±2.9	**76.9** ±7.7	**96.3** ±3.7	93.8 ±6.3	80.0 ±4.0	62.5 ±4.2	83.6 ±0.7	85.5 ±1.1	73.2 ±2.4
TIDE	Hybrid	**95.2** ±3.3	71.8 ±4.4	88.9 ±3.7	93.8 ±6.3	**85.3** ±2.3	72.2 ±4.8	86.0 ±0.8	**89.8** ±1.2	**77.2** ±1.8
Qwen2.5-3B-Instruct										
Vanilla†	Prompt	44.4	11.1	6.2	15.4	28.6	12.5	21.9	6.7	0.8
GRPO	RL	91.4 ±2.9	61.5 ±7.7	96.3 ±3.7	62.5 ±6.3	65.3 ±4.6	47.2 ±4.8	74.0 ±0.8	79.8 ±1.1	63.3 ±2.0
OPD	Distill	90.5 ±3.4	74.4 ±4.5	90.1 ±3.3	77.1 ±4.1	61.3 ±4.4	52.8 ±5.0	75.7 ±2.0	80.1 ±2.1	69.3 ±2.9
OPSD†	Distill	48.8	41.7	16.7	0.0	15.8	16.7	28.1	11.3	3.1
SDAR†	Hybrid	**97.1**	62.5	**100.0**	61.9	75.0	**84.2**	84.4	85.0	68.0
ATOD	Hybrid	94.3 ±2.9	**76.9** ±7.7	96.3 ±3.7	87.5 ±6.3	88.0 ±4.0	66.7 ±4.2	86.4 ±0.7	86.6 ±0.8	74.1 ±2.0
TIDE	Hybrid	94.3 ±2.9	**76.9** ±7.7	92.6 ±3.7	**93.8** ±6.3	**96.0** ±4.0	75.0 ±4.2	**89.3** ±0.7	**90.2** ±0.9	**79.0** ±1.5
Qwen2.5-7B-Instruct										
Vanilla†	Prompt	36.1	22.2	3.1	0.0	0.0	0.0	12.5	5.9	1.6
GRPO	RL	91.4 ±2.9	87.2 ±4.4	96.3 ±3.7	81.3 ±6.3	65.3 ±4.6	58.3 ±4.2	80.5 ±0.8	80.9 ±1.0	72.6 ±1.7
GiGPO†	RL	**97.7**	82.7	98.8	83.7	89.3	**79.2**	90.8	84.4	72.8
OPD	Distill	91.4 ±2.8	84.6 ±4.0	95.1 ±3.0	79.2 ±3.8	64.0 ±4.1	56.9 ±4.6	79.3 ±1.9	80.5 ±1.8	71.7 ±2.6
OPSD†	Distill	50.0	60.0	22.7	21.4	17.6	9.5	32.8	4.5	2.3
SDAR†	Hybrid	94.7	75.0	**100.0**	86.7	68.2	78.9	85.9	89.4	82.8
ATOD	Hybrid	97.1 ±2.9	84.6 ±7.7	96.3 ±3.7	93.8 ±6.3	88.0 ±4.0	70.8 ±4.2	89.3 ±0.7	89.1 ±0.7	79.0 ±1.5
TIDE	Hybrid	95.2 ±1.6	**92.3** ±7.7	95.1 ±2.1	**97.9** ±3.6	**94.7** ±2.3	77.8 ±2.4	**92.1** ±0.7	**92.1** ±0.7	**83.2** ±1.3

Training details. WebShop and ALFWorld use 16 prompts with 8 rollouts per prompt and train for 160 updates. We optimize students with AdamW at a learning rate of 10^{-6} and weight decay 0.01. PPO performs one epoch per update with an importance-ratio clipping radius of c=0.2 and entropy coefficient 0.001; minibatch sizes are 64 on WebShop and 256 on ALFWorld. Disagreement is smoothed with an exponential moving average of μ=0.9 over a W=10-update window, and the handoff rates are η=ζ=0.02. The OPD and RL coefficients are bounded by $\lambda \in [0.1, 1.0]$ and $\beta \in [0.1, 1.0]$, respectively; training therefore starts with $\lambda_0 = 1.0$ and $\beta_0 = 0.1$. Detailed relative-action-value configurations are provided in Appendix A.4.

3.2 Main Results

Table 2 compares TIDE with prompting, reward-optimization, distillation, and hybrid baselines across Qwen2.5-1.5B, 3B, and 7B scales. Values marked with † are reported from prior work; every unmarked row was evaluated by us under the matched protocol. Prompting baselines remain weak on both benchmarks. Among reward-optimization methods, GiGPO reports the highest displayed ALFWorld average at 1.5B, while TIDE achieves the highest displayed WebShop success rate at all three scales. Among distillation methods, task-trained OPD consistently outperforms the OPSD reference. For hybrid methods, the controlled ATOD–TIDE comparison holds the teacher and training setting fixed: TIDE exceeds ATOD on both benchmarks at every scale. At 7B, TIDE also exceeds the displayed SDAR reference on WebShop success (83.2% vs. 82.8%) and ALFWorld average success (92.1% vs. 85.9%), and achieves the highest displayed ALFWorld average (92.1%).

3.3 Ablation Studies

3.3.1 Discrepancy-triggered handoff against alternative schedules

Table 3(a) tests whether using the measured discrepancy trend as a schedule signal improves on fixed, preset, and success-guided alternatives in our evaluated setting. All hybrid variants use the same Qwen2.5-1.5B-Instruct student and frozen GRPO-trained Qwen2.5-7B-Instruct teacher; they

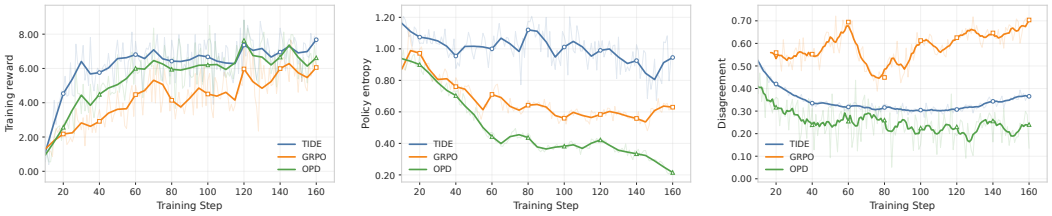
Table 3: TIDE ablations with Qwen2.5-1.5B-Instruct students (%). Bold indicates the highest displayed value.

(a) Global handoff strategies

Method	WebShop		ALFWorld	
	SR ↑	Score ↑	Seen ↑	Unseen ↑
GRPO	62.8 ±2.0	82.4 ±1.2	58.8 ±0.8	47.8 ±2.6
Success Handoff	71.4 ±2.5	82.2 ±1.6	73.6 ±2.5	73.1 ±2.7
Fixed 1:1 Mixture	72.0 ±1.7	84.4 ±0.8	77.1 ±1.4	79.9 ±1.5
Linear Handoff	72.2 ±2.2	86.3 ±1.3	85.7 ±1.9	81.3 ±2.0
Cosine Handoff	73.2 ±1.4	87.3 ±0.7	85.0 ±1.4	81.3 ±1.5
TIDE	**77.2** ±1.8	**89.8** ±1.2	**86.0** ±0.8	**85.1** ±1.5

(b) Local signal allocation

Method	WebShop		ALFWorld	
	SR ↑	Score ↑	Seen ↑	Unseen ↑
w/o OPD Mod.	74.4 ±1.5	85.5 ±0.8	83.6 ±1.2	80.6 ±1.5
w/o RL Mod.	72.6 ±2.4	84.7 ±1.4	85.0 ±1.9	79.1 ±2.0
w/o Process Reward	72.2 ±2.6	84.4 ±1.7	80.0 ±2.1	79.1 ±2.2
w/o Disagreement	74.4 ±1.6	85.0 ±1.0	83.6 ±1.4	79.9 ±1.5
Additive Fusion	67.4 ±3.8	74.7 ±2.7	70.7 ±2.5	70.1 ±2.2
TIDE	**77.2** ±1.8	**89.8** ±1.2	**86.0** ±0.8	**85.1** ±1.5



(a) Training reward. (b) Policy entropy. (c) Teacher–student disagreement.

Figure 3: Training dynamics of TIDE and baselines on WebShop. Faint lines show per-step values; bold curves show smoothed trends.

differ only in the evolution of the global coefficients. We compare TIDE with GRPO, Success Handoff, Fixed 1:1 Mixture, and preset linear and cosine handoffs. Definitions of the time schedules, together with realized handoff trajectories and parameter sensitivity, are provided in Appendix B.2.

TIDE's discrepancy-triggered handoff outperforms all fixed, preset, and success-guided alternatives. Success Handoff reaches 71.4% WebShop success and 73.6%/73.1% ALFWorld seen/unseen success, below the fixed and preset schedules. Fixed 1:1 Mixture improves over GRPO on both benchmarks, confirming the value of combining the two signals. Among preset schedules, Cosine Handoff is strongest on WebShop at 73.2% success; TIDE exceeds it by 4.0 points on WebShop success and 3.8 points on ALFWorld unseen success.

3.3.2 Relative-value–disagreement ablation

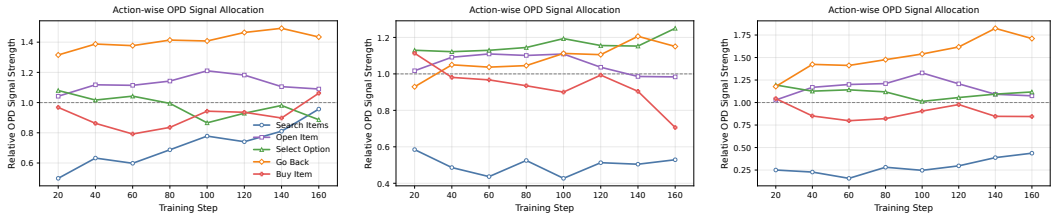
Following the setup in Section 3.3, Table 3(b) evaluates the design of TIDE's local allocation. Every variant retains TIDE's global handoff, teacher, training budget, and OPD–RL objective, and differs only in its local allocation. We compare removing the process-reward or disagreement signal, replacing their product with Additive Fusion, and disabling local modulation in the OPD or RL branch. Detailed definitions of these variants are provided in Appendix A.4.

The full method reaches 77.2% WebShop SR and outperforms every local-allocation variant. Removing the process-reward or disagreement signal reduces SR to 72.2% or 74.4%, while Additive Fusion reduces it to 67.4%, showing that both signals and their multiplicative fusion are necessary. Disabling local modulation in the RL or OPD branch reduces SR to 72.6% or 74.4%. Thus, the strongest performance requires joint local modulation of both OPD and RL branches.

3.4 Analysis

3.4.1 Global handoff dynamics

To characterize how the discrepancy-triggered handoff changes the optimization, we track training reward, policy entropy, and teacher–student disagreement throughout training. All methods use the 1.5B setup in Section 3.1; TIDE and OPD share the frozen GRPO-trained 7B teacher. Appendix B.2 additionally reports the realized benchmark-level handoff trajectories and parameter sensitivity.



(a) Relative-value diagnostic. (b) Disagreement diagnostic. (c) TIDE OPD allocation.

Figure 4: Turn-level learning-signal allocation by parsed action type. Relative-value and disagreement diagnostics, followed by TIDE's joint OPD allocation.

Figure 3 shows that TIDE reaches higher training reward earlier than GRPO and OPD on this WebShop setting while maintaining higher policy entropy. Its teacher–student disagreement decreases rapidly in the early updates and later stabilizes at an intermediate level between GRPO and OPD. Together, these trajectories illustrate the intended progression: teacher-weighted updates stabilize learning early, while the later increase in reward optimization preserves policy diversity and supports refinement toward higher-value behavior. This plot characterizes the resulting training dynamics; the disagreement statistic is a practical event signal for setting the schedule, not a direct measurement of teacher utility or RL reliability.

3.4.2 Turn-level learning-signal allocation

To understand how local modulation shapes behavior, we examine turn-level OPD signal strength across parsed agent-action types under relative-value-only, disagreement-only, and joint modulation. We group turns from the same TIDE rollouts by action type and normalize the mean local weight of each group by the mean over all turns. Thus, a value above or below one indicates that turns of the corresponding type are emphasized or suppressed. The underlying disagreement remains the response-level statistic in Equation 2. We use the same 1.5B WebShop configuration as Section 3.1, including the rollout budget and optimization settings.

Figure 4 highlights Go Back, an action that returns to a previous page after an unproductive search or option-selection branch. Such corrective actions can redirect the subsequent trajectory and therefore affect task completion. Disagreement-only modulation also emphasizes other teacher–student differences, whereas joint modulation assigns Go Back a higher relative OPD weight by combining disagreement with relative process value. The resulting allocation focuses teacher guidance on corrective decisions associated with favorable outcome evidence.

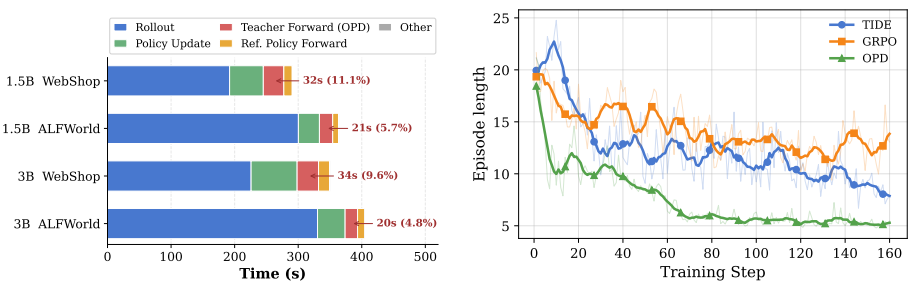
3.4.3 Computational cost

Given a frozen task-trained teacher, we measure student update time for 1.5B and 3B models on WebShop and ALFWorld, and compare WebShop episode lengths of TIDE, GRPO, and OPD.

Figure 5a shows that rollout generation dominates update time; the teacher forward pass shared with OPD accounts for 5–11%, and TIDE-specific computation adds negligible overhead. This excludes the one-time GRPO training used to construct a teacher, which may be reused across students on the same task. Figure 5b shows that, after initially longer OPD-dominated episodes, TIDE uses fewer interactions than GRPO later in training while remaining above OPD. It therefore improves WebShop success while retaining efficient interaction.

4 Related Work

On-policy distillation. On-policy distillation transfers teacher knowledge using samples drawn from the current student policy. Conventional knowledge distillation trains a compact student on teacher predictions over a fixed dataset (Hinton et al., 2015); Generalized Knowledge Distillation instead scores student-generated sequences with the teacher to reduce distribution mismatch (Agarwal et al., 2024). Subsequent work constructs reasoning supervision from verified on-policy solutions (Zhao et al., 2026) or incorporates self-generated distillation targets into reward-



(a) Student-training time per update.

(b) Interactions per episode.

Figure 5: Student-training and interaction cost of TIDE, given a frozen task-trained teacher. (a) Time per student update. (b) Environment interactions per episode.

based training (Hübotter et al., 2026; Yang et al., 2026a). For interactive agents, recent methods use extracted skills (Wang et al., 2026b), on-policy experience (Yang et al., 2026b), or iterative distillation (Wu et al., 2026) to support agent self-evolution. On-policy samples are particularly important in interactive settings, where each action changes the subsequent context on which teacher guidance is evaluated. These approaches establish teacher guidance on the student's evolving trajectory distribution; TIDE further adapts its allocation across both training stages and interaction turns.

Credit assignment for interactive agents. Fine-grained credit assignment uses intermediate evidence to identify decisions that contribute to an eventual outcome, especially when supervision is available only at the end of a long interaction. In mathematical reasoning, step-level labels (Wang et al., 2024), divide-and-conquer search (Luo et al., 2024), implicit process rewards (Cui et al., 2025), and rollout-based value estimation (Kazemnejad et al., 2024) provide process supervision. For interactive agents, GiGPO uses repeated environment states to construct relative process advantages (Feng et al., 2026), HGPO uses matched observation histories to estimate hierarchical process credit (He et al., 2026), and GraphGPO propagates credit through graph-structured interaction histories (Cheng et al., 2026). These methods provide outcome-relevant evidence at a finer granularity than trajectory-level rewards. Building on this line, TIDE uses a history-based relative process-value signal as outcome evidence for local OPD–RL allocation and combines it with teacher–student disagreement to prioritize turns for both branches.

Hybrid distillation and reinforcement learning. Recent concurrent work studies different mechanisms for coordinating dense distillation with outcome-based policy optimization. ATOD uses a prescribed linear annealing schedule from OPD toward RL (Tan et al., 2026); RetireOPD retires teacher supervision once discrepancy plateaus and the student reaches a teacher-relative performance threshold (Yu et al., 2026). Other methods use signal-calibrated weighting (Zheng et al., 2026a), sample routing (Li et al., 2026b), verifiable rewards (Lin et al., 2026), teacher distributions (Liu et al., 2026; Wang et al., 2026a), trajectory-relative normalization (Zheng et al., 2026b), objective-influence estimates (Lan et al., 2026), or lookahead agreement (Qu et al., 2026); SDAR uses per-token gating (Lu et al., 2026), and two-stage approaches sequence distillation and RL (Ye et al., 2026; Li et al., 2026a; Kim & Lee, 2026). Together, these approaches instantiate the trade-off through schedules, routing, or local gating. TIDE complements them with a shared two-scale design that coordinates the OPD–RL balance across training and update allocation across turns, using outcome-aware priority to connect the two levels.

5 Conclusion

We introduced TIDE, a two-scale approach for coordinating teacher supervision and reward optimization in multi-turn agents. Its global controller uses a smoothed discrepancy trend to transition from OPD to RL without prescribing a task-specific handoff step. Locally, relative action value and response-level discrepancy allocate teacher-guided and reward-driven updates across turns. Results on WebShop and ALFWorld, across the evaluated student scales, show improvements over fixed, scheduled, and component alternatives.

AI Assistance Disclosure

Large language model tools were used for language editing and organization of the manuscript, literature retrieval and citation discovery, conceptual discussion, and assistance with code-review and research-execution workflows. GPT-5.5 was additionally used only as the external turn-quality evaluator described in Appendix A.3; it did not participate in policy training, reward construction, checkpoint or hyperparameter selection, or test-time action selection. The authors reviewed all AI-assisted material and take responsibility for the final manuscript, experiments, and claims.

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*, volume 2024, pp. 21246–21263, 2024.

Xin Cheng, Shuo He, Lang Feng, HaiYang Xu, Ming Yan, Lei Feng, and Bo An. Beyond trajectory-level attribution: Graph-based credit assignment for agentic reinforcement learning. *arXiv preprint arXiv:2605.26684*, 2026.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.

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.

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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yibin Huang, Bin Xu, Hailong Cao, and Conghui Zhu. Bicaa: Bidirectional credit assignment for search-augmented agent. *arXiv preprint arXiv:2608.01321*, 2026.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms. *arXiv preprint arXiv:2410.01679*, 2024.

Jaehoon Kim and Dongha Lee. Opsd compresses what rlvr teaches: A post-rl compaction stage for reasoning models. *arXiv preprint arXiv:2605.06188*, 2026.

Qizhen Lan, Xi Xiao, Xiangchen Guan, Mengchen Fan, Moule Lin, Jung Im Choi, and Lijing Zhu. Trust is not enough: Influence calibration for on-policy self-distillation in agentic rl. *arXiv preprint arXiv:2608.14945*, 2026.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Boyan Li, Bingsen Chen, Chenghao Yang, Ping Nie, Chen Zhao, and Xi Ye. Sequential beats joint: On the interplay between on-policy distillation and rlvr. *arXiv preprint arXiv:2609.04108*, 2026a.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. Unifying group-relative and self-distillation policy optimization via sample routing. *arXiv preprint arXiv:2604.02288*, 2026b.

Wenze Lin, Jiale Zhao, Xitai Jiang, Songde Rao, Yining Li, Shenzhi Wang, Bingxiang He, and Gao Huang. On-policy distillation with verifiable reward. *arXiv preprint arXiv:2608.24696*, 2026.

Xinyu Liu, Kechen Jiao, Chunyang Xiao, Runsong Zhao, Junhao Ruan, Bei Li, Jiahao Liu, Qifan Wang, Xin Chen, Jingang Wang, et al. Teacher-guided policy optimization for on-policy reasoning distillation under large policy divergence. *arXiv preprint arXiv:2605.13230*, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, et al. Self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2605.15155*, 2026.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Changle Qu, Sunhao Dai, Hengyi Cai, Yuqi Zhou, Xinran Chen, Jun Xu, et al. Turnsight: Turn-level hindsight self-distillation for tool-integrated reasoning. *arXiv preprint arXiv:2608.04007*, 2026.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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.

Qitai Tan, Zefang Zong, Yang Li, and Peng Chen. Atod: Annealed turn-aware on-policy distillation for multi-turn autonomous agents. *arXiv preprint arXiv:2606.27814*, 2026.

Chen Wang, Zhaochun Li, Jionghao Bai, Yining Zhang, Hexuan Deng, Ge Lan, and Yue Wang. Distilled reinforcement learning for llm post-training. *arXiv preprint arXiv:2607.17247*, 2026a.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, et al. Skill-sd: Skill-conditioned self-distillation for multi-turn llm agents. *arXiv preprint arXiv:2604.10674*, 2026b.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, et al. Seed: Self-evolving on-policy distillation for agentic reinforcement learning. *arXiv preprint arXiv:2607.14777*, 2026.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

Shuo Yang, Jinyang Wu, Zhengxi Lu, Yuhao Shen, Fan Zhang, Lang Feng, Shuai Zhang, Haoran Luo, Zheng Lian, Zhengqi Wen, et al. Opid: On-policy skill distillation for agentic reinforcement learning. *arXiv preprint arXiv:2606.26790*, 2026b.

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.

Qinglin Ye, Zhiyuan Gu, Jingjie Xia, Yiheng Zhang, Kaiyan Zhao, Shunchao Zheng, Yuhang Mu, Wenchao Du, and Yiming Wang. Opdsearch+: On-policy distillation with rl refinement for search-augmented reasoning. *arXiv preprint arXiv:2608.24310*, 2026.

Yan Yu, Zhengxi Lu, Yizhou Liu, Yichen Pan, Aozhe Wang, Qipeng Chen, Hua Yang, Wenqi Zhang, Weiming Lu, Qianglong Chen, et al. Retireopd: Self-retiring on-policy distillation for agentic reinforcement learning. *arXiv preprint arXiv:2609.20784*, 2026.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

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*, 2026a.

Haoyu Zheng, Yun Zhu, Qing Wang, and Wenqiao Zhang. Trajectory-relative hindsight distillation for agentic reinforcement learning. *arXiv preprint arXiv:2608.07371*, 2026b.

A Supplementary Implementation Details

Section 3.1 gives the shared training configuration. Here we specify additional implementation details. We use one PPO epoch per update, a dual-clip bound of 3, and no reference-policy KL penalty.

Teacher construction and performance. For every benchmark, the teacher is a frozen Qwen2.5-7B-Instruct model trained with GRPO using the same task environment, prompts, action parser, reward function, rollout configuration, optimization hyperparameters, and training budget as the corresponding student run. The only difference is backbone scale: the teacher uses 7B parameters, whereas the student uses Qwen2.5-Instruct at 1.5B, 3B, or 7B parameters. The GRPO-trained teacher is frozen throughout TIDE training, used only to compute on-policy token log-probabilities, and never used at evaluation. We use the final checkpoint after the fixed training budget, without validation- or test-set-based checkpoint selection. Its task performance is reported by the 7B GRPO row in Table 2. For 1.5B and 3B students, the teacher is strictly larger in parameter count. For the 7B student, teacher and student have the same backbone scale; in this case, the teacher is a frozen task-trained GRPO policy rather than a larger-model teacher.

OPD signal implementation. For each student-sampled token, we subtract the behavior-policy log-probability from the frozen teacher's log-probability. Both quantities are fixed when forming the PPO advantage, so gradients flow only through the current-policy importance ratio. The resulting OPD term is a sampled-action distillation signal.

A.1 Baseline Descriptions

Vanilla. Zero-shot prompting with the base Qwen2.5-Instruct model. The model receives only the task description and environment observations, with no demonstrations or structured action guidance.

GRPO. Group Relative Policy Optimization (Shao et al., 2024) trains the student with outcome-based reward, normalizing advantages across the group of rollouts sampled for each prompt. No teacher supervision is used.

GiGPO. Group-in-Group Policy Optimization (Feng et al., 2026) supplements GRPO with a step-level relative advantage obtained by grouping matched interaction contexts within each prompt.

OPD. On-policy distillation (Agarwal et al., 2024) uses the frozen GRPO-trained teacher's log-probability on student-sampled tokens as an additional PPO advantage signal. It is the distillation-only baseline in Table 2 and uses the same frozen task-trained teacher and training configuration as ATOD and TIDE at each student scale.

OPSD. On-Policy Self-Distillation (Zhao et al., 2026) reinforces the student's verified on-policy solutions through self-distillation. It uses no frozen external teacher and serves as a self-improving distillation reference.

ATOD. Annealed Turn-Aware On-Policy Distillation (Tan et al., 2026) uses a prescribed linear OPD-to-RL annealing schedule. We reproduce ATOD under the same task environment, prompts, action parser, reward function, rollout configuration, optimization hyperparameters, training budget, decoding settings, and evaluation protocol as TIDE. Its values in Table 2 are mean $\pm$ sample standard deviation over three independently trained seeds.

SDAR. Self-Distilled Agentic Reinforcement Learning (Lu et al., 2026) combines distillation and RL through per-token gating within a shared student–teacher policy. We report its available 3B and 7B results as a method-specific hybrid reference.

A.2 Benchmark-specific Evaluation Protocol

Table 2 uses ALFWorld's official valid-seen split (140 tasks); its Avg. is the task-count-weighted success rate across the six task types. The 134-task valid-unseen split is reported only for controlled ablations and sensitivity analyses, since corresponding records are unavailable for some external baselines. Other evaluation settings are given in Section 3.1.

A.3 GPT-5.5 Process-quality Diagnostic

The GPT-5.5 scores in Table 1 are an external diagnostic only, following the use of LLM-based evaluators for fine-grained generation assessment (Liu et al., 2023): they are not used in training, reward construction, checkpoint selection, hyperparameter selection, or test-time action selection. We randomly sample 100 full trajectories from the WebShop test set and use the trajectory, rather than the turn, as the sampling unit. We score every executable turn in the sampled trajectories.

Blind judge input. For each turn, the judge receives the task description, at most two preceding observation–action pairs, the current observation, the admissible actions, and the parsed executed action. It does *not* receive the terminal reward, trajectory-success label, subsequent observations or actions, teacher/student log-probabilities, disagreement values, relative turn position, or the model's hidden reasoning. This restriction makes the diagnostic independent of the outcome quantities subsequently reported in Table 1.

Judge prompt and score. We query GPT-5.5 once per turn. The judge evaluates the executed action rather than proposing an alternative and assigns an integer score on a $[1, 3]$ scale. This coarse rubric is appropriate because the judge observes only the current decision context, rather than its future consequences. We do not use repeated judge calls or human annotations for this diagnostic, so it should be interpreted as supporting evidence for the motivation rather than a ground-truth process label. The following prompt is used verbatim, with bracketed fields instantiated for the evaluated turn:

User. Evaluate the *executed action* of an autonomous agent in a WebShop shopping task using only the task, recent interaction history, current observation, admissible actions, and executed action below. Do not infer future observations, terminal reward, purchase outcome, trajectory success, teacher preference, student probability, disagreement statistic, or hidden reasoning. Judge the action itself, not its action type.
Apply the following criteria. For a search action, assess whether its query targets the task's explicit constraints. For a product click, assess whether the visible title, price, or metadata makes the product a reasonable candidate. For an option, description, feature, or review click, assess whether it selects or verifies a task-relevant attribute. For back, previous, or next actions, assess whether navigation is justified by visible evidence or needed to continue a reasonable search. For `Buy Now`, assign a positive score only if the visible information verifies all explicit task constraints, including the selected variant when applicable.
Assign one integer score: 1 = clearly harmful, invalid, or inconsistent with the visible task constraints; 2 = reasonable but inconclusive information gathering, verification, or navigation; 3 = clearly advances the task based on visible evidence, or completes a purchase whose explicit constraints have been verified. When the visible context is insufficient to establish either an error or clear progress, assign 2. Return only valid JSON: {``score'': <1, 2, or 3>, ``reason'': ``<one concise sentence>"}.
Task: [task description]. Recent interaction history: [up to two preceding observation–action pairs]. Current observation: [current observation]. Admissible actions: [admissible actions]. Executed action: [parsed executed action].

A.4 Relative Action Value Details

We follow GiGPO's relative process-value comparison (Feng et al., 2026) with exact recent-history matching. The resulting relative action value Q_t is an outcome-conditioned credit heuristic over alternatives reached from matched recent states, related to rollout-based process-value estimation (Cui et al., 2025; Kazemnejad et al., 2024); it is not an independent causal attribution of a turn's contribution. Our configuration uses exact state-history matching with two preceding anchor states (history length 2), mean-centered group scores, length-weight exponent 1, and no trajectory-level base group.

Discounted return. Given a T-turn trajectory τ with task outcome $R(\tau)$ and a discount factor γ, the discounted return from turn t onward is

$$G_t = \gamma^{T-t} R(\tau). \tag{12}$$

We set $\gamma = 0.95$ in all experiments. The discounted return is not used directly as a local weight: before it contributes to Q_t, it is mean-centered within a matched-history comparison group. This removes the group-level return baseline, but it does not algebraically remove the effect of discounting when matched continuations have different remaining lengths. Thus, Q_t is not determined by raw turn position alone—it also depends on the alternatives sharing the same state history—but it is a time-discounted outcome-conditioned heuristic rather than position-free or causal turn attribution.

Exact recent observation-context grouping. Let s_t be the environment-provided anchor observation at turn t, before prompt formatting or the addition of interaction history. The policy input retains the two most recent observation–action pairs, whereas process-value grouping uses exact suffixes of recent anchor observations. For prompt group g and suffix length k, its comparison key is

$$K_{t,k} = \big(g, k, (s_{t-k+1}, \ldots, s_t)\big), \qquad k \in \{1, 2, 3\}. \tag{13}$$

Here $k = 3$ corresponds to the current anchor state plus the two preceding states (the configured history length is 2). We group turns with identical keys; textual states are matched by exact string equality, while arrays, lists, and dictionaries are recursively converted to hashable tuples. We do not use learned or semantic-similarity matching. Comparing downstream returns within a group yields a relative score from a matched recent context.

Per-group normalization. Within each group $K_{t,k}$ that contains at least two members, we compute a mean-centered process value:

$$p_{t,k} = G_t - \mu_{K_{t,k}}. \tag{14}$$

This is the configured `mean_norm` mode: we do not divide by the group standard deviation. $\mu_{K_{t,k}}$ is the mean discounted return in the group. Negative $p_{t,k}$ indicates a below-average outcome within

the matched context, whereas positive $p_{t,k}$ indicates a better-than-average outcome; a group with no return variation yields only zero scores.

Aggregation across history lengths. A turn may have nonzero scores from multiple suffix lengths. We aggregate them using a length-weighted combination:

$$Q_t = \sum_{k \in \mathcal{H}_t} \frac{(k+1)^\nu}{\sum_{q \in \mathcal{H}_t} (q+1)^\nu} p_{t,k}, \tag{15}$$

where $\mathcal{H}_t \subseteq \{1, 2, 3\}$ contains suffix lengths with a nonzero centered score for turn t, and ν controls the preference for longer suffixes. We set $\nu = 1$ and do not include a trajectory-level base group. A turn has no value-discriminative matched group when every exact-history group either contains only that turn or has no return variation; in this case, it is assigned raw $Q_t = 0$.

The trajectory-wise normalization and the local score z_t are defined in Equation 6. A turn with no value-discriminative matched group retains raw $Q_t = 0$ and is excluded from the value min–max range; its value factor is set to one as a disagreement-only OPD fallback. Thus, it receives disagreement-only OPD modulation but no value-driven RL advantage. The same fallback applies when the valid matched turns have no within-set value range. Disagreement is normalized over all turns; when its range is at most ϵ, its factor is one for every turn. All reported experiments use this fallback.

Local-ablation definitions. All local-allocation variants retain the global handoff, teacher, training budget, and combined OPD–RL objective. The full method uses relative action value as the RL advantage, normalized disagreement as the RL turn weight, and the product of normalized relative action value and disagreement as the OPD turn weight. The individual variants change these quantities as follows:

- *w/o Process Reward*: removes relative action value from both branches. RL uses the unmodulated trajectory-level GRPO advantage; normalized disagreement is used as the turn weight for both RL and OPD.
- *w/o Disagreement*: removes normalized disagreement from both branches. RL retains relative action value as its advantage but uses a unit turn weight, whereas OPD uses normalized relative action value as its turn weight.
- *w/o OPD Mod.*: retains the full RL branch but replaces the OPD turn weight with a unit weight.
- *w/o RL Mod.*: retains the full OPD branch and relative-action-value RL advantage but replaces the RL turn weight with a unit weight.
- *Additive Fusion*: retains the RL branch and replaces the OPD turn weight with the arithmetic mean of normalized relative action value and normalized disagreement.

Equivalently, with $q_t = \tilde{Q}_t$, $d_t = \tilde{D}_t$, $A^{\mathrm{RL}}_{t,j} = Q_t$, and $A^{\mathrm{OPD}}_{t,j} = \delta_{t,j}$ for the full method, the variants are

$$\begin{aligned}
\textit{w/o Process Reward}: &\quad A^{\mathrm{RL}}_{t,j} = A^{\mathrm{GRPO}}_{t,j}, \quad w^{\mathrm{RL}}_t = w^{\mathrm{OPD}}_t = d_t; \\
\textit{w/o Disagreement}: &\quad A^{\mathrm{RL}}_{t,j} = Q_t, \quad w^{\mathrm{RL}}_t = 1, \quad w^{\mathrm{OPD}}_t = q_t; \\
\textit{w/o OPD Mod.}: &\quad w^{\mathrm{OPD}}_t = 1; \qquad \textit{w/o RL Mod.}: \quad w^{\mathrm{RL}}_t = 1; \\
\textit{Additive Fusion}: &\quad w^{\mathrm{OPD}}_t = (q_t + d_t)/2, \quad w^{\mathrm{RL}}_t = d_t.
\end{aligned}$$

Algorithm 1 Trajectory-wise local-priority computation

Require: Turn values $Q_{1:T}$, valid-turn set $\mathcal{V}$, disagreements $D_{1:T}$, threshold ϵ

$\widetilde{Q}_t \leftarrow 1$ for $t \notin \mathcal{V}$
if $\mathcal{V} = \emptyset$ or $\max_{t\in\mathcal{V}} Q_t - \min_{t\in\mathcal{V}} Q_t \le \epsilon$ **then**
  $\widetilde{Q}_t \leftarrow 1$ for $t \in \mathcal{V}$
else
  $\widetilde{Q}_t \leftarrow (Q_t - \min_{u\in\mathcal{V}} Q_u)/(\max_{u\in\mathcal{V}} Q_u - \min_{u\in\mathcal{V}} Q_u + \epsilon)$ for $t \in \mathcal{V}$
end if
$\Delta_D \leftarrow \max_t D_t - \min_t D_t$
if $\Delta_D \le \epsilon$ **then**
  $\widetilde{D}_t \leftarrow 1$ for all t
else
  $\widetilde{D}_t \leftarrow (D_t - \min_u D_u)/(\Delta_D + \epsilon)$ for all t
end if
$z_t \leftarrow \widetilde{Q}_t \widetilde{D}_t$ for all t
return $z_{1:T}$

Normalization pseudocode.

Coverage of matched-history comparisons. We measure valid-group coverage as the fraction of turns that have at least one exact recent observation-context group with two or more members and nonzero return variance. This measures how often relative action value is available for local allocation. Figure 6 reports coverage for the 1.5B WebShop and ALFWorld TIDE training runs. Coverage is lower during early exploration, when repeated contexts are sparse, and then quickly becomes high and stable on both benchmarks. Thus, unmatched turns are concentrated early rather than dominating local modulation throughout training. Faint lines show per-update coverage, and bold curves show 10-update moving averages.

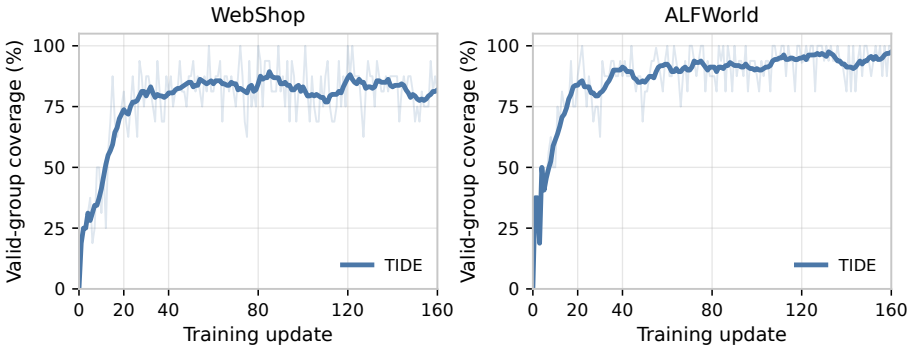


Figure 6: Valid matched-history coverage across 1.5B TIDE training on WebShop and ALFWorld.

B Supplementary Results and Analyses

B.1 Additional SearchQA Global-Handoff Evaluation

We additionally evaluate the global handoff on SearchQA (Dunn et al., 2017) with Qwen2.5-1.5B-Instruct students to assess whether the scheduling mechanism transfers to retrieval-augmented question answering (Lewis et al., 2020; Huang et al., 2026). This experiment does not apply local modulation. SearchQA contains NQ, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue, and Bamboogle. Agents retrieve three passages from Wikipedia-2018 with E5-base-v2 (Wang et al., 2022), retain four history turns, and take at most five interaction steps. We report exact-match accuracy on normalized final `<answer>` spans. Each update uses 128 questions with 8 rollouts per

question; PPO minibatches contain 512 trajectories. Results are mean ± sample standard deviation over three independently trained random seeds.

Table 4 compares GRPO, OPD, and TIDE. TIDE achieves the highest macro-average EM of 39.4%, exceeding OPD by 3.1 points and outperforming it on six of seven subsets. The result indicates that the global handoff remains beneficial in this setting; it does not test the local module.

Table 4: SearchQA exact-match accuracy with Qwen2.5-1.5B-Instruct students (%). Bold indicates the highest displayed student result.

Method	Type	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.
GRPO	RL	14.8 ±1.1	28.3 ±1.6	21.1 ±0.7	15.8 ±1.4	23.9 ±1.2	2.7 ±0.1	9.3 ±0.8	16.6 ±0.3
OPD	Distill	36.0 ±0.8	49.1 ±1.4	38.8 ±0.6	36.8 ±1.1	36.2 ±0.9	26.1 ±1.7	**31.2** ±1.0	36.3 ±0.5
TIDE	Hybrid	**40.8** ±1.2	**51.5** ±0.9	**44.3** ±0.5	**42.4** ±1.4	**38.3** ±1.0	**27.7** ±1.6	30.6 ±0.8	**39.4** ±0.3

B.2 Global Handoff Analysis

Motivation and setup. The global mechanism uses discrepancy as a practical event signal rather than following a prescribed training-time schedule. To characterize the realized schedule, we trace the handoff state r_k throughout training on WebShop, ALFWorld, and SearchQA.

Alternative schedule definitions. Success Handoff sets the handoff state r_k in Equation 5 to the observed batch success rate. Fixed 1:1 Mixture keeps unit OPD and RL coefficients throughout training. Linear and Cosine Handoff use prescribed handoff states to interpolate the OPD and RL coefficients. With normalized training progress $s_k = (k-1)/(K-1)$ over K updates, they set $r_k = s_k$ and $r_k = [1 - \cos(\pi s_k)]/2$, respectively.

Results. Figure 7 reports the realized task-level handoff trajectories. All begin with $r_k = 0$ and exhibit task-dependent transition windows and intermediate plateaus. The plot documents how the chosen event rule adapts its transition timing across tasks.

Parameter sensitivity. The plateau threshold ζ determines when a handoff step is triggered, and the step size η determines its magnitude. To test whether the handoff depends on a narrow parameter choice, we vary each independently across a $5\times$ range using Qwen2.5-1.5B-Instruct students on WebShop and ALFWorld. All other settings follow the main experimental configuration.

Table 5: Sensitivity of global handoff parameters (%).

(a) Plateau threshold ζ

ζ	WebShop		ALFWorld	
	SR ↑	Score ↑	Seen ↑	Unseen ↑
0.01	76.0 ±1.9	89.1 ±1.2	85.7 ±1.5	84.3 ±1.7
0.02	**77.2** ±1.8	**89.8** ±1.2	**86.0** ±0.8	**85.1** ±1.5
0.05	75.6 ±2.0	88.4 ±1.3	85.0 ±1.6	82.8 ±1.9

(b) Handoff step size η

η	WebShop		ALFWorld	
	SR ↑	Score ↑	Seen ↑	Unseen ↑
0.01	76.4 ±1.9	89.2 ±1.2	85.7 ±1.5	83.6 ±1.8
0.02	**77.2** ±1.8	**89.8** ±1.2	**86.0** ±0.8	**85.1** ±1.5
0.05	75.8 ±2.0	88.6 ±1.3	85.0 ±1.7	82.1 ±2.0

Results. The default setting, $\zeta = \eta = 0.02$, gives the strongest result in each reported column. Performance remains stable over the evaluated range: all variants remain within 1.6 points on WebShop SR and 3.0 points on ALFWorld unseen, indicating robustness to moderate changes in either parameter.

B.3 Additional 3B Ablations

Motivation and setup. We repeat the global-handoff and local-allocation ablations with Qwen2.5-3B-Instruct students to test whether the 1.5B findings persist at a larger student scale. Each variant uses the same teacher, rollout budget, training budget, and evaluation protocol as the 3B main result; global variants differ only in their handoff schedule, and local variants differ only in their turn-level allocation.

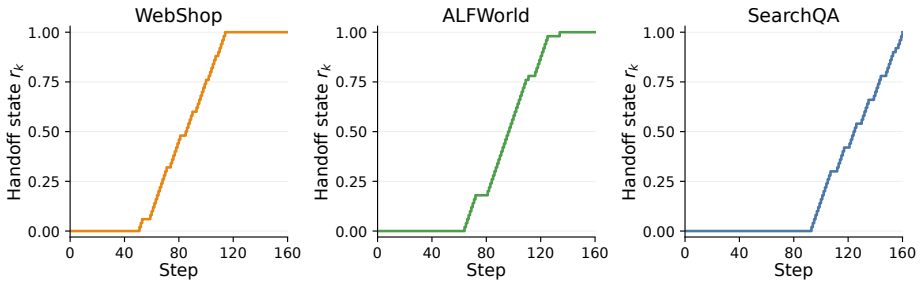


Figure 7: Global handoff states across WebShop, ALFWorld, and SearchQA.

Table 6: TIDE ablations with Qwen2.5-3B-Instruct students (%). Bold indicates the highest displayed value.

(a) Global handoff strategies

Method	WebShop		ALFWorld	
	SR ↑	Score ↑	Seen ↑	Unseen ↑
GRPO	63.3 ±2.0	79.8 ±1.1	74.0 ±0.8	60.4 ±2.2
Success Handoff	74.2 ±2.3	84.5 ±1.4	79.3 ±1.9	77.6 ±2.2
Fixed 1:1 Mixture	74.9 ±1.8	85.8 ±1.0	82.1 ±1.2	80.3 ±1.7
Linear Handoff	75.2 ±2.1	87.5 ±1.2	87.1 ±1.4	84.6 ±1.7
Cosine Handoff	76.0 ±1.6	88.4 ±0.9	86.4 ±1.4	85.1 ±1.5
TIDE	**79.0** ±1.5	**90.2** ±0.9	**89.3** ±0.7	**87.3** ±1.5

(b) Local signal allocation

Method	WebShop		ALFWorld	
	SR ↑	Score ↑	Seen ↑	Unseen ↑
w/o OPD Mod.	77.1 ±1.4	88.0 ±0.8	87.1 ±1.2	85.6 ±1.6
w/o RL Mod.	75.2 ±2.1	86.8 ±1.3	87.9 ±1.4	83.8 ±1.9
w/o Process Reward	74.8 ±2.3	86.2 ±1.5	83.6 ±1.9	83.1 ±2.2
w/o Disagreement	77.0 ±1.5	87.4 ±1.0	87.1 ±1.4	84.6 ±1.6
Additive Fusion	70.3 ±3.4	77.5 ±2.5	74.3 ±2.1	73.1 ±2.2
TIDE	**79.0** ±1.5	**90.2** ±0.9	**89.3** ±0.7	**87.3** ±1.5

Results. For global handoff, TIDE reaches 79.0% WebShop SR and 87.3% ALFWorld unseen success, exceeding the strongest preset schedule, Cosine Handoff, by 3.0 and 2.2 points, respectively. For local allocation, it again outperforms every variant. Removing process reward, disagreement, OPD modulation, or RL modulation lowers WebShop SR to 74.8%, 77.0%, 77.1%, and 75.2%, respectively; Additive Fusion is lowest at 70.3%. These results reproduce the two 1.5B findings at 3B: the discrepancy-triggered handoff improves over the evaluated schedules, and the joint local allocation is stronger than its component variants.