BRIDGE: BLACK-BOX ON-POLICY AGENT DISTILLATION BY BRIDGING EXPERT AND LEARNER TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Distilling capable large language model (LLM) agents into smaller models offers a practical path toward efficient deployment, yet existing approaches typically specialize in either imitation from expert trajectories or correction on student-generated trajectories, leaving the two complementary supervision regimes disconnected. More fundamentally, both paradigms often couple trajectory provenance with supervision direction, treating teacher-generated behavior as the learning target regardless of its realized quality. We introduce **BRIDGE**, a black-box agent distillation framework that bridges expert and learner trajectories within a unified preference-learning framework while decoupling trajectory source from preference direction. BRIDGE constructs comparisons from both expert- and learner-induced histories and determines their labels from realized trajectory quality, allowing either teacher- or student-generated behavior to provide the positive learning signal. It further explores informative recurrent states and terminates local comparisons upon environment-state reconvergence, yielding finer-grained credit assignment while limiting unnecessary policy divergence. BRIDGE requires only teacher-generated text and environment feedback, without access to teacher logits or internal representations. Across WEBSHOP, ALFWORLD, and SCIENCEWORLD, BRIDGE consistently outperforms strong imitation- and correction-based distillation baselines, and remains effective with DeepSeek-R1-Distill-Llama-8B students, demonstrating robust black-box distillation across heterogeneous model families. The anonymous code repository is available at `https://anonymous.4open.science/r/bridge123/`.

1 INTRODUCTION

Large language models (LLMs) have shown strong capabilities as autonomous agents, encompassing reasoning, tool use, and multi-turn interaction with external environments [23, 15]. However, frontier-scale agents remain expensive to deploy, motivating agent distillation: transferring their interactive capabilities to smaller and more efficient students [8, 12, 10]. Existing approaches largely follow a teacher-guided paradigm, in which a stronger teacher either provides expert trajectories for the student to *imitate*, or intervenes on student-generated trajectories to *correct* its behavior [8, 12, 11, 18, 9]. Both paradigms have substantially improved smaller models across a range of agent tasks.

Despite their success, imitation and correction provide complementary but imperfect supervision. As illustrated in Figure 1(a), imitation learns from coherent and successful expert trajectories, but exposes the student primarily to teacher-induced histories; once the student deviates at deployment, prediction errors may accumulate on unfamiliar states [14, 3]. Correction addresses this mismatch by supervising student-generated trajectories and therefore covers states the student is more likely to encounter. Figure 1(b) illustrates the reverse problem: the teacher must now provide guidance on potentially erroneous and unfamiliar student-induced histories, where its supervision may become less reliable [18, 9, 2]. More fundamentally, both paradigms often let model identity determine the training target: teacher-generated behavior is preferred simply because it comes from the teacher. As a result, the student may absorb imperfect teacher guidance. The same rule may also suppress superior student behavior as the student approaches or surpasses the teacher.

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.

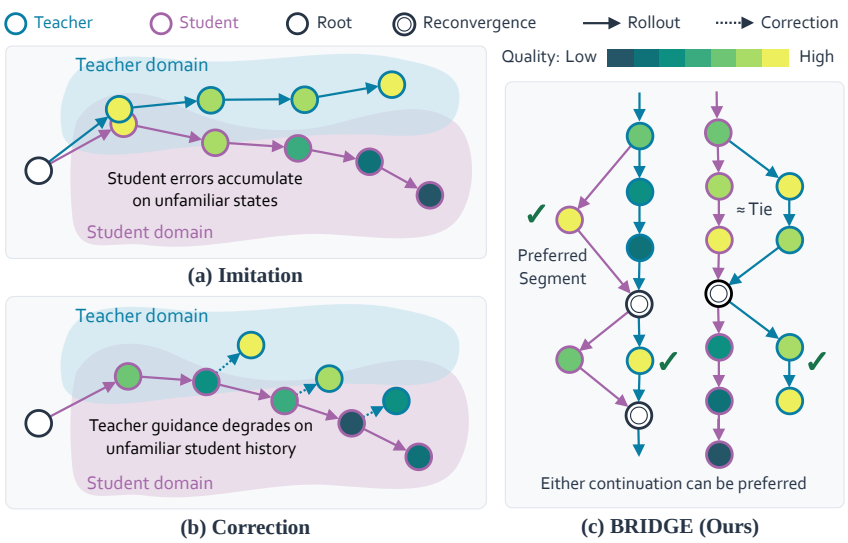


Figure 1: **Motivation and overview of BRIDGE. (a) Imitation** learns from expert trajectories but leaves the student vulnerable after deviating from familiar histories. **(b) Correction** covers student-induced histories but may require the teacher to provide guidance in unfamiliar and error-prone contexts. **(c) BRIDGE** compares teacher and student continuations from both history sources, determines preference by trajectory quality rather than model identity, and localizes credit at environment-state reconvergence. Either continuation can be preferred.

To address these limitations, we introduce **BRIDGE**, a preference-based agent distillation framework that bridges expert and learner trajectories to exploit their complementary supervision, as shown in Figure 1(c).[1] BRIDGE compares divergent segments of a teacher trajectory and a student trajectory by their trajectory quality. The better continuation becomes the preferred response regardless of which model generated it, and DPO [13] optimizes the student on the resulting preference pairs. This quality-based labeling avoids reinforcing teacher behavior by default and preserves better alternatives discovered by the student. To use both history sources, BRIDGE further branches from recurrent states on expert and learner trajectories: the student explores alternatives from expert histories, while the teacher provides alternatives from learner histories. These cross-policy continuations broaden supervision across both expert- and learner-induced histories, allowing the student to learn from reliable expert contexts while remaining exposed to states induced by its own policy. Finally, long-horizon comparisons can obscure which decisions caused the quality difference. BRIDGE therefore truncates a comparison when the two branches first reconverge to the same environment state and uses only the preceding divergent segments for preference learning. This design helps avoid attributing outcomes after reconvergence to earlier divergent behavior, providing more localized credit. Together, these mechanisms allow the student to benefit from teacher guidance without treating the teacher as an infallible target, while also learning from states induced by its own policy. Because BRIDGE relies only on trajectories and environment feedback, it requires neither teacher logits nor internal representations and naturally supports black-box teachers.

We evaluate BRIDGE on three interactive agent benchmarks: WEBSHOP [21], ALFWORLD [16], and SCIENCEWORLD [19]. Our main setting uses a GiGPO-trained [5] Qwen3-8B [20] agent as the teacher and Qwen3-4B as the student. BRIDGE achieves the highest task score on all three benchmarks, reaching 65.63, 95.62, and 65.46, respectively. These scores exceed the strongest competing Qwen3-4B baseline by 7.45, 5.84, and 3.18 points, and even surpass the Qwen3-8B teacher across all three environments. BRIDGE also produces substantially fewer responses with violations than SFT. To test heterogeneous black-box distillation, we further use DeepSeek-R1-Distill-Llama-8B [4] as the student. BRIDGE remains effective despite the change in model family and tokenizer, without modifying the teacher or accessing its token probabilities.

[1] We use *teacher*/*student* for model identities and *expert*/*learner* for trajectory roles; teacher-generated behavior is not necessarily preferred.

Our contributions are threefold. First, we identify a shared limitation of imitation- and correction-based agent distillation: both tend to conflate *where supervision is collected* with *which behavior should be preferred*. We separate these decisions by determining preference from trajectory quality rather than model identity, allowing either teacher- or student-generated behavior to provide the positive signal. Second, **BRIDGE** jointly exploits expert- and learner-induced histories through cross-trajectory continuation and environment-grounded preferences. It explores recurrent states and terminates comparisons at environment-state reconvergence, providing localized credit without predefined step-level preference labels. Third, BRIDGE improves task performance while maintaining low violation rates across three interactive environments, and it remains effective across student model families using only textual access to the teacher.

2 Related work

Agent distillation from expert trajectories. A straightforward approach to agent distillation is to transfer successful expert trajectories into smaller models through supervised fine-tuning. Agent Distillation trains compact students on complete reasoning and tool-use trajectories generated by stronger agents [8], while Structured Agent Distillation decomposes trajectories into reasoning and action spans to better preserve the structure of agent behavior [10]. Sub-goal Distillation instead augments expert trajectories with intermediate subgoals and fine-tunes separate planning and execution modules, demonstrating that structured supervision extracted from expert demonstrations can substantially improve smaller agents [7]. These approaches highlight the value of expert trajectories as dense and coherent sources of supervision, but largely use them as fixed targets for behavioral reproduction. BRIDGE retains their value as high-quality interaction histories while using them as starting points for alternative student continuations, thereby extending expert demonstrations into a broader space of behavioral comparisons.

On-policy distillation. On-policy distillation (OPD) reduces the mismatch between training and inference by collecting supervision on sequences generated by the student itself [11]. Its extension to interactive agents is particularly challenging because early decisions alter subsequent observations and progressively change the context in which teacher guidance is provided. TCOD introduces a temporal curriculum over interaction depth [18], Guided-OPD mixes teacher- and student-generated turns to constrain trajectory drift [9], and SAGE-OPD selectively applies teacher intervention according to its estimated utility [24]. FutureBridge-OPD further evaluates teacher guidance through its influence on subsequent student trajectories [2]. While conventional OPD typically relies on teacher token probabilities, Black-Box On-Policy Distillation shows that on-policy transfer can also be performed using only sampled teacher responses [22]. BRIDGE shares the motivation of learning from student-induced states but goes beyond learner-history-only correction: it jointly exploits expert- and learner-induced histories and, crucially, separates the history used to construct a comparison from the identity of its preferred continuation. This design naturally requires only generated teacher responses and therefore remains compatible with black-box teachers.

Preference learning for LLM agents. Preference learning replaces direct target imitation with relative comparisons between candidate behaviors [13], making it attractive for agents whose quality is often easier to assess at the trajectory level than at the token or action level. Agentic-DPO constructs state-conditioned preferences from expert trajectories by contrasting expert actions with student-generated alternatives [3]. Hierarchical Preference Learning (HPL) focuses on the granularity of agent supervision, combining trajectory-, action-group-, and step-level preferences to balance sparse long-horizon feedback against noisy local credit assignment [6]. These works highlight two central design questions for preference-based agent learning: how candidate behaviors should be compared, and at what temporal granularity preference credit should be assigned. BRIDGE addresses both through environment-grounded trajectory comparison: the preferred branch is determined by realized trajectory quality, while reconvergence between competing branches provides a natural boundary for localizing credit to the segment in which their behaviors actually differ.

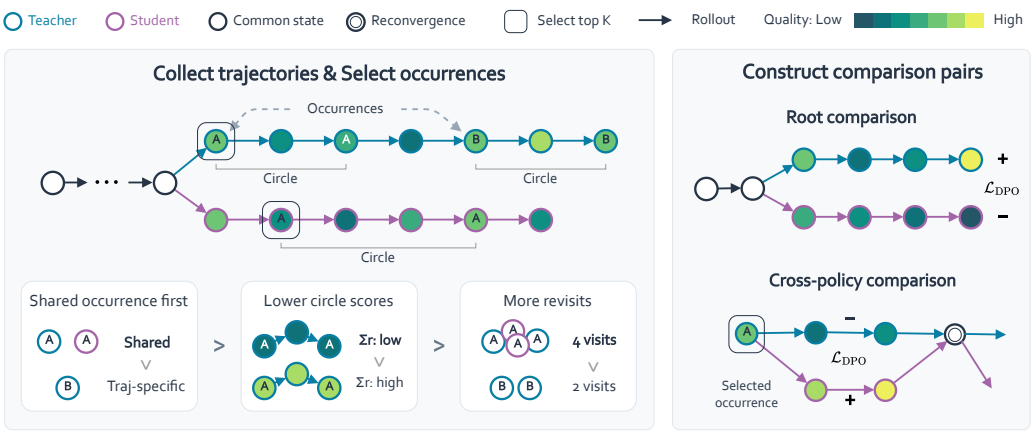


Figure 2: Overview of BRIDGE. Left: one teacher trajectory and one student trajectory are collected, and recurrent-state occurrences are selected from the two parent trajectories. Occurrences are prioritized by shared-state support, lower circle scores, and more repeated visits. Right: BRIDGE constructs two comparison types. The root comparison contrasts the teacher and student suffixes after their common-state prefix. Each selected occurrence also initiates a cross-policy comparison: the source continuation is retained, the other policy generates from the same history, and the two branches are compared until their first reconvergence. The resulting comparison pairs are ordered by quality and used for DPO.

3 METHOD

BRIDGE constructs preference supervision from teacher–student behavioral comparisons under both teacher- and student-induced interaction histories. Section 3.1 describes how BRIDGE builds the comparison set for each task, including an initial root comparison and additional cross-policy comparisons constructed along the parent trajectories. Section 3.2 converts these comparisons into quality-ordered preference pairs and trains the student with DPO. Finally, Section 3.3 describes the asynchronous rollout procedure used to collect task-level preference data efficiently while bounding policy staleness.

3.1 BRIDGING TEACHER AND STUDENT TRAJECTORIES

As illustrated in Figure 2, RIDGE constructs two types of teacher–student comparisons for each task. The *root comparison* directly contrasts the complete teacher and student trajectories after their common-state prefix. Moreover, *cross-policy comparisons* assess the two policies under additional interaction histories. The left panel of Figure 2 shows how the positions at which these comparisons begin are selected along the parent trajectories, while the right panel shows how the resulting root and cross-policy comparisons are constructed. For each selected position, the continuation on the source trajectory is retained while the other policy generates from the same history, until the two branches reconverge to the same environment state. The procedure can then continue along the source trajectory to construct subsequent comparisons. Together, the root and cross-policy comparisons form the comparison set used for preference learning.

To formalize this procedure, consider an interactive task x, where an agent acts according to its interaction history h and receives a step score r from the environment. The score reflects task feedback together with penalties for undesirable behaviors such as invalid, rejected, or repetitive interactions. The realized quality of a trajectory segment ξ is measured by $Q(\xi) = \sum_{t \in \xi} r_t$. For each task, the fixed teacher policy π_T provides one complete trajectory τ_T, and the trainable student policy π_θ one complete trajectory τ_S.

Initial root comparison. The complete trajectories τ_T and τ_S provide the root comparison. Their common-state prefix is absorbed into a shared history h_0, and the remaining suffixes form the teacher and student segments ξ_0^T and ξ_0^S. More generally, we denote the i-th comparison by $\langle \xi_i^T, \xi_i^S \rangle_{h_i}$,

where h_i denotes the common interaction history ending immediately before the two segments first differ. Both complete trajectories are then retained as parent trajectories for constructing additional cross-policy comparisons.

Recurrent-state occurrence selection. To construct cross-policy comparisons, BRIDGE must determine where along the parent trajectories the other policy should begin generating. It therefore searches both trajectories for environment states visited more than once, which we refer to as *recurrent states*. Such states arise naturally in modern long-horizon agents: a computer-use agent may return to the same application or browser state after inspecting another interface, while a software-engineering agent may return to the same repository and execution state after evaluating and reverting an alternative modification. Compared with states visited only once, recurrent states expose multiple observed continuations from the same environment configuration and therefore provide richer opportunities for behavioral exploration and comparison.

For each recurrent state, consecutive visits define candidate *circles*, whose cumulative step scores summarize the quality of the intervening behaviors. We call the earlier visit of each eligible circle an *occurrence*; each occurrence identifies a concrete position and interaction history from which a cross-policy comparison can begin. According to whether its underlying state is visited by both parent trajectories, an occurrence is either shared or trajectory-specific. Shared occurrences provide histories reached under both teacher- and student-induced trajectories, reducing the distribution mismatch associated with supervision concentrated on only one side. Trajectory-specific occurrences remain useful when the two policies visit substantially different parts of the environment, allowing BRIDGE to retain informative recurrent behavior even without cross-trajectory overlap.

BRIDGE ranks candidate occurrences lexicographically. It first prioritizes shared occurrences, then favors lower-scoring circles to focus comparisons on less successful intermediate behavior, and finally prefers states with more repeated visits across the two trajectories. The highest-ranked K occurrences are selected, or all eligible occurrences if fewer than K are available. These selected occurrences determine where the cross-policy comparisons described next are initiated.

Cross-policy continuation with reconvergence comparison. Having determined where cross-policy comparisons begin, BRIDGE next constructs them through cross-policy continuation and reconvergence, as illustrated in the right panel of Figure 2. Each selected occurrence specifies an interaction history and the continuation originally observed after it. BRIDGE performs *cross-policy continuation* by retaining this source continuation while asking the other policy to generate from exactly the same history. Thus, an occurrence on τ_T pairs the retained teacher behavior with a student-generated alternative, whereas an occurrence on τ_S pairs the retained student behavior with a teacher-generated alternative. In this way, BRIDGE constructs cross-policy comparisons under histories induced by either policy without assigning a preferred branch in advance.

After any common initial behavior has been absorbed into h_i, the first differing action marks the start of the comparison. Generation stops when the newly generated branch first reaches an environment state visited by the source continuation; we refer to this event as *reconvergence*. The matched state marks the end of the current comparison, and ξ_i^T and ξ_i^S denote the corresponding teacher and student segments between their first difference and reconvergence. BRIDGE then moves forward along the source continuation to the matched history and performs another cross-policy continuation. The procedure repeats until the source continuation is exhausted or the newly generated branch terminates, so a single selected occurrence can yield multiple cross-policy comparisons.

Let N_x denote the total number of root and cross-policy comparisons constructed for task x. Their union forms

$$\mathcal{C}(x) = \left\{ \left\langle \xi_i^T, \xi_i^S \right\rangle_{h_i} \right\}_{i=0}^{N_x - 1}. \tag{1}$$

At this stage, $\mathcal{C}(x)$ records only the provenance and realized behavior of the two branches. Section 3.2 determines their preference direction from trajectory quality.

3.2 QUALITY-BASED PREFERENCE LEARNING

BRIDGE assigns preference solely according to realized segment quality. For every non-tied comparison in $\mathcal{C}(x)$, the higher-scoring segment becomes the preferred response:

$$p_i = \begin{cases} \langle \xi_i^T, \xi_i^S \rangle_{h_i}, & Q(\xi_i^T) > Q(\xi_i^S), \\ \langle \xi_i^S, \xi_i^T \rangle_{h_i}, & Q(\xi_i^S) > Q(\xi_i^T). \end{cases} \tag{2}$$

Comparisons with equal scores are discarded. We denote the preferred and rejected segments in each retained pair by ξ_i^+ and ξ_i^-, respectively, giving the preference set

$$\mathcal{P}(x) = \left\{ \langle \xi_i^+, \xi_i^- \rangle_{h_i} \;\middle|\; i \in \{1, \ldots, N_x - 1\},\ Q(\xi_i^T) \neq Q(\xi_i^S) \right\}. \tag{3}$$

The mapping from $\langle \xi_i^T, \xi_i^S \rangle_{h_i}$ to $\langle \xi_i^+, \xi_i^- \rangle_{h_i}$ separates behavior provenance from preference direction: either teacher- or student-generated behavior can provide the positive training signal.

We optimize the resulting pairs with the standard DPO objective [13]. Let π_{ref} denote the reference policy and $\mathcal{P} = \bigcup_x \mathcal{P}(x)$. The objective is

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\langle \xi^+, \xi^- \rangle_h \sim \mathcal{P}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(\xi^+ \mid h)}{\pi_{\text{ref}}(\xi^+ \mid h)} - \beta \log \frac{\pi_\theta(\xi^- \mid h)}{\pi_{\text{ref}}(\xi^- \mid h)} \right) \right]. \tag{4}$$

Here, $\pi(\xi \mid h)$ denotes the likelihood assigned to the model-generated responses in the interaction segment, with environment observations encountered along the segment incorporated into the conditioning context. Both branches are scored only by the student and reference policies. BRIDGE therefore requires sampled teacher responses but no teacher logits or token probabilities, preserving the black-box setting regardless of whether the teacher-generated branch is preferred or rejected.

3.3 ASYNCHRONOUS ROLLOUT GENERATION

Each BRIDGE training iteration uses preference pairs from B tasks whose rollouts and local comparisons have completed and yielded non-empty preference sets. Since task durations vary with trajectory length and the number of local comparisons, BRIDGE collects rollouts asynchronously to reduce waiting for slow tasks, while retaining unfinished tasks and unconsumed results across iterations.

Asynchronous generation may leave completed tasks produced by slightly older student policies. DPO can optimize previously collected preference pairs and does not require data to be sampled from the current policy at every update [13, 17]; online preference optimization instead refreshes data from the evolving policy to improve distributional coverage [1, 17]. Let u denote the current student version and v_x the oldest version that generated a student response within task x. We define the policy lag of task x as $u - v_x$ and accept its preference pairs only if $u - v_x \leq L$. Tasks exceeding this limit are discarded, and collection continues in the normal sampling order without immediately retrying expired tasks. The learner proceeds once B eligible completed tasks are available, prioritizing those with older v_x. We use $L = 1$ in all experiments without tuning.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Environments and evaluation. We evaluate BRIDGE on three interactive agent benchmarks with distinct task and reward structures: WebShop [21], ScienceWorld [19], and ALFWorld [16]. WebShop requires agents to navigate an e-commerce website and select products that satisfy user-specified constraints, with reward provided only at the end of each episode. ScienceWorld evaluates multi-step scientific reasoning and experimentation in a text-based environment and provides intermediate step rewards that reflect task progress. ALFWorld evaluates embodied household planning and interaction, using a binary terminal reward for task success. To provide finer-grained comparison signals in the two sparse-reward environments, we additionally assign a penalty to violations, such as invalid actions, loop actions, or malformed outputs during pair construction. These auxiliary penalties are used only for comparison construction and do not affect the benchmark scores reported

at evaluation. We report Score and violation rate (VR): Score reflects task performance and is reported on a 0–100 scale, while VR denotes the percentage of agent responses exhibiting behavioral violations. Episodes are limited to 15 interaction steps in WebShop and 30 steps in ScienceWorld and ALFWorld. For each training run, we report the best evaluation checkpoint observed during training.

Models. Our primary setting distills an environment-specific Qwen3-8B teacher trained with GiGPO [5] into a Qwen3-4B student [20]. To examine whether BRIDGE remains effective when the teacher and student come from different model families, we additionally use DeepSeek-R1-Distill-Llama-8B [4] as the student. Unless otherwise specified, we construct a fixed pool of teacher trajectories for each environment and retain trajectories with task scores of at least 0.7 as the source expert data.

Baselines. In addition to reporting the raw student and teacher performance, we compare BRIDGE with representative approaches from the two major paradigms of teacher-guided agent learning. **SFT** represents imitation learning and fine-tunes the student on the same successful teacher trajectories available to BRIDGE. **OPD** [11] represents correction under learner-induced histories, using teacher token likelihoods to supervise student-generated behavior. We further include **TCOD** [18], which augments on-policy distillation with a temporal curriculum, using the configuration recommended in the original paper. Together, these baselines provide comparisons against both expert-trajectory imitation and on-policy correction.

Training details. BRIDGE is trained for 150 updates on all benchmarks, with $B = 32$ accepted tasks per update. We use AdamW with a learning rate of 1×10^{-6}. We set $K = 2$, selecting at most two recurrent-state occurrences for additional comparisons. Student rollouts are collected with a maximum policy lag of one update. Additional training, sampling, and implementation details are provided in Appendix A.

4.2 Main results

Table 1 reports the main results on WebShop, ALFWorld, and ScienceWorld. With Qwen3-4B as the student, BRIDGE achieves the highest task score on all three benchmarks, reaching 65.63, 95.62, and 65.46, respectively. Compared with SFT, the strongest baseline in task performance, BRIDGE improves the score by 7.45 points on WebShop, 5.84 on ALFWorld, and 3.18 on ScienceWorld. Both imitation- and correction-based baselines substantially improve over the vanilla student, confirming the value of teacher guidance, but their supervision is still tied to teacher identity: SFT directly reproduces teacher trajectories, while OPD and TCOD treat teacher predictions as the correction target on learner-induced histories. BRIDGE instead determines the preference direction from realized trajectory quality and learns from comparisons under both expert- and learner-induced histories. This allows it to retain useful teacher guidance without preferring teacher behavior by default, yielding consistent gains over SFT, OPD, and TCOD. Notably, BRIDGE also surpasses the Qwen3-8B teacher on all three benchmarks. Since SFT itself already exceeds the teacher on WebShop and ALFWorld, these results further indicate that teacher performance need not constitute an upper bound and motivate allowing better student-generated behavior to serve as the preferred signal.

BRIDGE also maintains low violation rates while improving task performance. With Qwen3-4B, it achieves the lowest VR on ALFWorld and ScienceWorld and the second-lowest on WebShop, where TCOD is slightly lower (0.59 vs. 0.79). In particular, the reduction on ScienceWorld is substantial, from 20.03 for SFT to 7.26 for BRIDGE. This suggests that the additional preference supervision does not simply optimize task return at the expense of interaction quality. It is also consistent with our use of reconvergence-based local comparisons, which restrict preference credit to the divergent behavior preceding state reconvergence rather than attributing subsequent shared outcomes to earlier decisions; we examine this effect further in the ablation study.

The results extend to heterogeneous black-box distillation with DeepSeek-R1-Distill-Llama-8B as the student. On WebShop, BRIDGE improves the task score from 59.33 with SFT to 60.10 despite changing both the model family and tokenizer, while achieving a VR of 2.30. On ALFWorld and ScienceWorld, BRIDGE similarly achieves the highest task scores of 97.78 and 65.50, with violation rates of 0.21 and 6.29, respectively. Unlike token-level distillation methods such as OPD and

Table 1: Main results on WebShop, ALFWorld, and ScienceWorld. Score denotes task performance and VR denotes violation rate; higher Score and lower VR are better. **Bold** and underlined entries indicate the best and second-best results, respectively.

Method	WebShop		ALFWorld		ScienceWorld	
	Score↑	VR (%) ↓	Score↑	VR (%) ↓	Score↑	VR (%) ↓
Qwen3-8B	58.15	1.89	85.76	4.30	63.27	24.28
Qwen3-4B						
Vanilla	47.38	5.50	47.44	17.67	42.41	27.52
SFT	58.18	2.52	89.78	3.38	62.28	20.03
OPD	56.97	2.44	87.59	4.82	53.87	35.56
TCOD	57.21	**0.59**	85.40	6.43	60.47	27.93
BRIDGE (Ours)	**65.63**	0.79	**95.62**	**0.66**	**65.46**	**7.26**
DeepSeek-R1-Distill-Llama-8B						
Vanilla	47.81	24.29	13.50	38.73	18.03	57.03
SFT	59.33	**0.47**	90.88	1.76	61.86	18.45
BRIDGE (Ours)	**60.10**	2.30	**97.78**	**0.21**	**65.50**	**6.29**

TCOD, BRIDGE requires only generated trajectories and environment feedback, making the same training mechanism directly applicable when teacher and student use different model architectures and tokenizations.

4.3 Analysis and Ablations

We analyze three aspects of BRIDGE: quality-based preference labeling, robustness to the quality of source teacher trajectories, and recurrent-state occurrence selection. Unless otherwise specified, all experiments in this section use Qwen3-4B as the student and follow the same teacher, training, and evaluation setup as the main experiments.

Table 2: Ablation on preference labeling and recurrent-state occurrence selection. Teacher$^+$ always treats the teacher-generated continuation as preferred, while K denotes the number of recurrent-state occurrences selected for additional comparisons.

Method	WebShop	ALFWorld	ScienceWorld
Teacher$^+$	58.53	87.23	61.52
$K = 0$	**65.92**	93.06	63.22
$K = 1$	65.66	94.89	64.99
$K = 2$	65.63	**95.62**	**65.46**

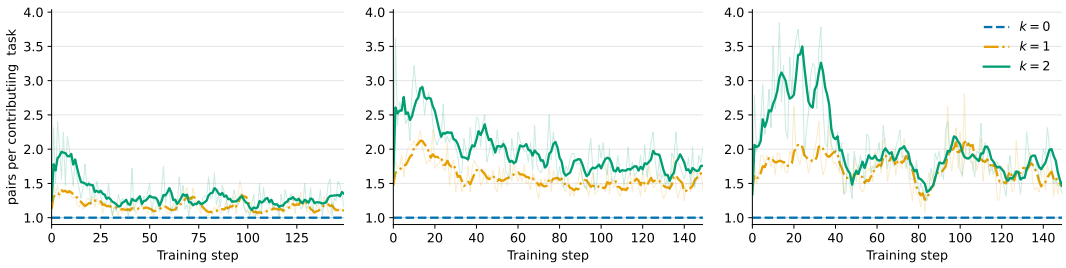
Quality-based preference labeling. We first examine whether preference direction should be determined by realized trajectory quality rather than generator identity. As shown in Table 2, forcing teacher-generated behavior to always serve as the preferred response (Teacher$^+$) substantially degrades performance, reducing the scores from 65.63 to 58.53 on WebShop, from 95.62 to 87.23 on ALFWorld, and from 65.46 to 61.52 on ScienceWorld.

This result shows that teacher-generated behavior is not necessarily the better learning target. As the student improves, better behavior can emerge from either side of a teacher–student comparison; labeling preferences by realized trajectory quality allows BRIDGE to retain such behavior instead of reinforcing the teacher by default.

Table 3: Sensitivity to the minimum score used to filter source teacher trajectories. Source traj. denotes the average score of the retained trajectories.

	WebShop			ALFWorld	
Min score	0	0.5	0.7	0	1
Source traj.	55.60	72.35	91.26	86.52	100.00
BRIDGE	65.39	65.58	65.63	**98.54**	95.62

Robustness to low-quality source trajectories. We further vary the minimum score used to retain source teacher trajectories. As shown in Table 3, BRIDGE is only weakly affected by large changes in the average quality of the source tra-



(a) WebShop (b) ALFWorld (c) ScienceWorld

Figure 3: Number of preference pairs per contributing task during training under different values of K. Increasing K enables BRIDGE to construct additional comparisons from selected recurrent-state occurrences, with the increase being more pronounced in ALFWorld and ScienceWorld.

jectory pool. On WebShop, lowering the threshold from 0.7 to 0 reduces the average source score from 91.26 to 55.60, yet BRIDGE changes only marginally, from 65.63 to 65.39. More strikingly, on ALFWorld, retaining lower-scoring teacher trajectories improves BRIDGE from 95.62 to 98.54 despite reducing the average source quality from 100.00 to 86.52. These results suggest that BRIDGE does not require the source pool to consist exclusively of high-quality demonstrations. Since preference direction is determined by realized trajectory quality rather than teacher identity, imperfect teacher behavior need not be reinforced and can still provide useful interaction histories for comparison. The improvement on ALFWorld further suggests that retaining a broader trajectory pool can sometimes be beneficial, potentially by increasing the diversity and coverage of available histories.

Effect of recurrent-state occurrence selection. We next vary the number of selected recurrent-state occurrences K. Setting $K = 0$ removes the additional cross-policy comparisons and retains only the root comparison. This variant already performs strongly, suggesting that quality-based preference learning alone accounts for a substantial portion of BRIDGE's improvement. Selecting recurrent-state occurrences provides further gains on the longer-horizon environments: on ALFWorld, performance increases from 93.06 at $K = 0$ to 94.89 and 95.62 at $K = 1$ and $K = 2$, respectively, while ScienceWorld improves from 63.22 to 64.99 and 65.46. Figure 3 helps explain this trend. Increasing K produces substantially more preference pairs per contributing task in ALFWorld and ScienceWorld, showing that recurrent states expose additional intermediate histories from which informative teacher–student comparisons can be constructed. In contrast, the three variants remain close on WebShop, where $K = 0$ slightly outperforms $K = 1$ and $K = 2$. This is consistent with its shorter interaction horizon: fewer useful recurrent-state occurrences are available, so increasing K yields only a small increase in pair density and correspondingly little additional performance gain. Overall, these results indicate that recurrent-state selection is most useful when longer trajectories provide sufficient opportunities for additional history-conditioned comparisons.

5 Conclusion

We introduced **BRIDGE**, a black-box agent distillation framework that bridges expert- and learner-induced interaction histories through quality-based preference learning. Rather than treating teacher-generated behavior as the default target, it separates trajectory provenance from preference direction and allows either teacher- or student-generated behavior to provide the preferred signal according to realized interaction quality. Cross-policy comparisons from recurrent states further extend supervision beyond complete trajectories, while environment-state reconvergence localizes preference credit to the behavioral segments in which the two policies differ. Across WebShop, ALFWorld, and ScienceWorld, the method consistently improves over imitation- and correction-based distillation baselines while requiring only teacher-generated text and environment feedback. Ablations show that quality-based preference labeling is central to these gains, with recurrent-state comparisons providing additional benefits in longer-horizon environments. Overall, the results suggest that effective agent distillation need not regard the teacher as a fixed behavioral target; expert and learner trajectories can instead serve as complementary sources of interaction histories and candidate behaviors whose learning value is determined by realized quality.

AI USE STATEMENT

We used generative AI tools to polish the language and improve the clarity of the manuscript, and to assist with code generation and experimental work. The authors manually reviewed the AI-assisted manuscript revisions, all AI-generated code, and all AI-assisted experiments and their results. The authors take full responsibility for the final manuscript, the correctness of the implementation and experimental results, and all claims and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This work studies agent distillation in the existing WEBSHOP, ALFWORLD, and SCIENCEWORLD benchmark environments. Our experiments do not involve human participants, the collection of new personal or sensitive data, or real-world transactions. We use publicly available data in accordance with their applicable licenses and usage agreements. Although the study is confined to benchmark environments, distillation may transfer biases or unsafe behaviors from the teacher to the student, and improved task performance does not by itself establish safety. Deployment beyond these benchmarks would require separate assessment of privacy, safety, and misuse risks, together with safeguards appropriate to the intended application.

REPRODUCIBILITY STATEMENT

To support reproducibility, we will provide all materials needed to reproduce our experiments at submission time through anonymized supplementary materials, including source code, training and evaluation configurations. The main text and appendix will detail the BRIDGE algorithm and preference construction procedure, model and environment versions, data splits and preprocessing, prompts, training hyperparameters, random seeds, computational resources, and evaluation protocols. The accompanying materials will include instructions for setting up the environments and running training and evaluation, as well as references to the publicly available models and benchmark data used in this work.

REFERENCES

[1] Daniele Calandriello, Zhaohan Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. Human alignment of large language models through online preference optimisation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5409–5435. PMLR, 2024.

[2] 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.

[3] Yixiong Chen and Alan Yuille. Agentic-dpo: From imitation to agentic policy optimization on expert trajectories. *arXiv preprint arXiv:2607.10601*, 2026.

[4] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[5] Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[6] Heyang Gao, Zexu Sun, Erxue Min, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Xu Chen. Solving the granularity mismatch: Hierarchical preference learning for long-horizon llm agents. In *International Conference on Learning Representations*, 2026.

[7] Maryam Hashemzadeh, Elias Stengel-Eskin, Sarath Chandar, and Marc-Alexandre Côté. Subgoal distillation: A method to improve small language agents. In *Proceedings of The 3rd Conference on Lifelong Learning Agents*, volume 274 of *Proceedings of Machine Learning Research*, pp. 1053–1075. PMLR, 2025.

[8] Minki Kang, Jongwon Jeong, Seanie Lee, Jaewoong Cho, and Sung Ju Hwang. Distilling llm agent into small models with retrieval and code tools. *Advances in Neural Information Processing Systems*, 38:106501–106538, 2026.

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

[10] Jun Liu, Zhenglun Kong, Peiyan Dong, Changdi Yang, Tianqi Li, Hao Tang, Geng Yuan, Wei Niu, Wenbin Zhang, Pu Zhao, Xue Lin, Dong Huang, and Yanzhi Wang. Structured agent distillation for large language model. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, 2026.

[11] 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.

[12] Yuanjie Lyu, Chengyu Wang, Jun Huang, and Tong Xu. From correction to mastery: Reinforced distillation of large language model agents. *arXiv preprint arXiv:2509.14257*, 2025.

[13] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[14] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

[15] Pascal J Sager, Benjamin Meyer, Peng Yan, Rebekka von Wartburg-Kottler, Layan Etaiwi, Aref Enayati, Gabriel Nobel, Ahmed Abdulkadir, Benjamin F Grewe, and Thilo Stadelmann. A comprehensive survey of agents for computer use: Foundations, challenges, and future directions. *Journal of Artificial Intelligence Research*, 85, 2026.

[16] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.

[17] Yuda Song, Gokul Swamy, Aarti Singh, J. Andrew Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. In *Advances in Neural Information Processing Systems*, volume 37, pp. 12243–12270, 2024.

[18] 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.

[19] 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. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.775.

[20] An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[21] 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, 2022.

[22] Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. Black-box on-policy distillation of large language models. *arXiv preprint arXiv:2511.10643*, 2025.

[23] Yingli Zhou, Wang Shu, Yaodong Su, Wenchuan Du, Yixiang Fang, and Xuemin Lin. A comprehensive survey on agent skills: Taxonomy, techniques, and applications. *arXiv preprint arXiv:2605.07358*, 2026.

[24] Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Bo Peng, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. Sage-opd: Selective agent-guided intervention for multi-turn on-policy distillation. *arXiv preprint arXiv:2606.19659*, 2026.

A Experimental details

A.1 Environments

ALFWorld. ALFWorld [16] is a text-based embodied environment derived from ALFRED, where agents complete household tasks through navigation and object manipulation. It contains six categories of tasks involving operations such as picking, placing, cleaning, heating, cooling, and examining household objects. We use both the seen and unseen evaluation splits for testing and set the maximum interaction horizon to 30 steps. The prompt template used for ALFWorld is shown below.

ALFWorld Task Prompt Template

```
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:

[Observation {i}]
{past_observation}
[Action {i}]
{history_action}
...

You are now at step {current_step}.
[Current Observation]
{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.
Once you've finished your reasoning, you should choose an admissible
action for current step and present it within <action> </action> tags.
```

WebShop. WebShop [21] is a simulated e-commerce environment in which an agent must search for and purchase a product that satisfies a natural-language user instruction. The agent interacts with the website through textual search and click actions, and task performance is measured by how well the selected product satisfies the requested attributes. We use the official test split for evaluation and set the maximum interaction horizon to 15 steps. The prompt template used for WebShop is shown below.

WebShop Task Prompt Template

```
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:

[Observation {i}]
{past_observation}
```

```
[Action {i}]
{history_action}
...

You are now at step {current_step}.
[Current Observation]
{current_observation}
Your admissible actions of the current situation are:
[
{admissible_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.
Once you've finished your reasoning, you should choose an admissible
action for current step and present it within <action> </action> tags.
```

ScienceWorld. ScienceWorld [19] is a text-based interactive environment for evaluating multi-step scientific reasoning and experimentation. Its tasks span a variety of elementary science concepts and require agents to interact with objects and the surrounding environment to accomplish a given scientific goal. Following our experimental setup, we enable the `teleportAction` and `openContainers` simplifications and set the maximum interaction horizon to 30 steps. The prompt template used for ScienceWorld is shown below.

ScienceWorld Task Prompt Template

```
You are a helpful assistant doing a scientific experiment in
ScienceWorld.
The environment has these rooms: kitchen, foundry, workshop, bathroom,
 outside, living room, bedroom, greenhouse, art studio, hallway.
You can teleport to any room in one step. All containers have already
been opened, so you can directly get items from containers.
Your 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:

[Observation {i}]
{past_observation}
[Action {i}]
{history_action}
...

You are now at step {current_step}.
[Current Observation]
{current_observation}
The available action templates are:
[
{admissible_actions}
].
Instantiate OBJ/LOC with concrete objects or rooms mentioned in the
observation or recent history.

Now it's your turn to take an action.
You should first reason step-by-step about the current situation.
Once you've finished your reasoning, you should choose an admissible
action for the current step and present it within <action> </action>
tags.
```

Table 4: Common training hyperparameters used across methods. Environment-specific values are reported in the order WebShop / ScienceWorld / ALFWorld.

Hyperparameter	Value
Student model	Qwen3-4B
Teacher model	Qwen3-8B (GiGPO)
Learning rate	1×10^{-6}
Maximum prompt length	8192
Maximum response length	1024
Thinking token budget	512
Maximum tokens per GPU	9216
Rollout temperature	1.0
Rollout top-p	1.0
Rollout top-k	disabled
Rollout score threshold	0.7
Maximum environment steps	15 / 30 / 30
Random seed	42

A.2 Baselines

We compare BRIDGE against three representative knowledge transfer paradigms for LLM agents:

Supervised fine-tuning (SFT). SFT serves as the standard imitation learning baseline. We fine-tune the student with the negative log-likelihood objective on high-quality trajectories collected from the teacher. Each trajectory is decomposed into interaction steps, where the prompt contains the interaction history up to the current state and the complete teacher response, including both reasoning and action, is used as the supervision target. Thus, SFT learns exclusively from teacher-induced interaction histories.

Vanilla on-policy distillation (OPD). We implement vanilla OPD [11] as the standard on-policy distillation baseline for multi-turn agents. The student interacts with the environment using its current policy and generates full-horizon trajectories. At each interaction step, the teacher evaluates the exact response tokens generated by the student under the same prompt, and the student is optimized using the resulting token-level teacher–student log-probability differences. Unlike BRIDGE, OPD therefore requires access to the teacher's token probabilities and applies teacher supervision exclusively on student-induced histories.

Temporal curriculum on-policy distillation (TCOD). We further compare against TCOD [18], a recent extension of OPD for multi-turn agent distillation. We adopt its Forward-to-Backward (F2B) variant, which initially restricts the student to a short rollout horizon from the beginning of each episode and progressively expands this horizon until the full interaction trajectory is covered. Within the exposed horizon, training follows the same token-level on-policy distillation procedure as OPD. Throughout the paper, TCOD refers to the F2B variant unless otherwise specified.

A.3 Training Hyperparameters

Unless otherwise specified, all methods share the training, model, and rollout configurations summarized in Table 4. For BRIDGE, OPD, and TCOD, we use one node with $8\times$ NVIDIA A800 GPUs for the teacher and two nodes with $8\times$ NVIDIA A800 GPUs each for the student. SFT is trained on two nodes with $8\times$ NVIDIA A800 GPUs each.

We use the same rollout generation configuration across all methods. Collected trajectories with an environment score below 0.7 are filtered out before being used for training.

BRIDGE. We train BRIDGE for 150 optimization steps in all three environments, with 32 accepted tasks per update. We use $K = 2$ recurrent-state occurrences and restrict the maximum policy

staleness to one update. Each collected batch is optimized for two epochs with a pair micro-batch size of 2 per GPU. For DPO optimization, we set $\beta = 1.0$.

SFT. We train SFT for 500 / 900 / 1600 optimization steps on WebShop / ScienceWorld / ALFWorld, respectively. Teacher rollouts passing the common score threshold are converted into step-level supervised examples. Each update uses a global batch of 64 teacher interaction turns and one optimization epoch. We use a cosine learning-rate schedule with a warmup ratio of 5%, and average the supervised loss over response tokens.

OPD. We train OPD for 150 / 250 / 250 optimization steps on WebShop / ScienceWorld / ALFWorld, respectively. Student trajectories are collected with 16 environment tasks per explorer batch. Each optimization update contains 64 student interaction turns and uses one training epoch. We allow a maximum policy staleness of two updates. Following our vanilla OPD implementation, the token-level teacher–student log-probability difference is clipped to $[-10, 10]$ before being used for optimization.

TCOD. TCOD uses the same optimization-step schedule, explorer batch size, training batch size, and maximum policy staleness as OPD. We use the Forward-to-Backward (F2B) curriculum, with curriculum intervals of 4 / 6 / 6 explorer batches for WebShop / ScienceWorld / ALFWorld, respectively. Unlike vanilla OPD, we do not clip the token-level teacher–student log-probability difference in TCOD.