# AUSO: Action-Level Unified Skill Optimization from Internalization to Utilization

Huizu Lin*
University of Science and Technology of China
Hefei, China
jordansancholhz@mail.ustc.edu.cn

Chengkai Huang*
University of New South Wales
Sydney, Australia
chengkay.huang@gmail.com

Tianqi Gao
Independent Researcher
China
tianqig358@gmail.com

Tao Huang†
University of Chinese Academy of Sciences
Beijing, China
thuang@iaii.ac.cn

Daijiao Liu
University of New South Wales
Sydney, Australia
daijiao.liu@student.unsw.edu.au

Tongxin Li
Xi'an Jiaotong-Liverpool University
Suzhou, China
tommy020929@gmail.com

Xiaoyan Sun
University of Science and Technology of China, Hefei Comprehensive National Science Center
Hefei, China
sunxiaoyan@ustc.edu.cn

Lina Yao
University of New South Wales
Sydney, Australia
lina.yao@unsw.edu.au



## Abstract

Skills play different roles as an agent's policy evolves: they should first provide learnable knowledge, then support capability formation, and finally be invoked only when they improve individual decisions. Existing methods rarely model this lifecycle. They either keep skills outside the model, fully internalize them, or select among internalization and utilization objectives through noisy task-level success rates. Such designs fragment training and assign uniform importance to actions within the same trajectory, even though skill guidance may help some decisions while distracting others. To solve these problems, we introduce **AUSO (Action-level Unified Skill Optimization)**, which unifies skill learning and skill use through a progressive, action-aware optimization process. At the beginning of training, AUSO jointly learns from teacher guidance and environmental outcomes, enabling the policy to acquire foundational skills without losing task-oriented feedback. It subsequently emphasizes outcome-based policy optimization to consolidate autonomous problem-solving ability. As the policy matures, AUSO evaluates each sampled action under both skill-conditioned and skill-free contexts. The resulting action-level information signal is coupled with the trajectory outcome advantage, allowing beneficial skill-sensitive actions to receive stronger updates and harmful ones to be suppressed. Therefore, skills gradually transition from an external source of supervision into decision knowledge whose utilization is adapted to its action-level benefit, while reinforcement learning remains the shared backbone across all stages. Experiments on ALFWorld, WebShop, and SearchQA show that AUSO consistently improves agent performance and out-of-distribution generalization over competitive baselines. Our code is available at https://github.com/JordanSancholhz/Action-Skill.

## CCS Concepts

• **Computing methodologies** → **Natural language processing**.

## Keywords

Agentic Reinforcement Learning, Agent Skills, Agentic Search, Large Language Model

## 1 Introduction

Large language models (LLMs) are increasingly evolving from static text generators into interactive agents that make sequential decisions in complex, long-horizon environments. Rather than producing a single response, these agents interleave reasoning, acting, and observation: they call external tools, retrieve information, navigate web interfaces, manipulate embodied environments, and revise plans based on feedback [13, 28, 46, 54]. Such interaction-centric settings are central to modern information access and decision-making tasks, including web shopping, embodied household execution, and search-based question answering [11, 24, 34, 44]. However, as task horizons grow, agents must repeatedly decide not only what action to take, but also which procedural knowledge, tool-use strategy, or environmental heuristic should guide that action.

Agent skills have emerged as a promising abstraction for this problem. They encode reusable procedural knowledge, task strategies, tool-use patterns, or environment-specific guidance, thereby providing agents with transferable priors for long-horizon decision-making [16, 37]. In prompt-based agents, skills are retrieved, composed, or injected as explicit guidance to support complex interactive task solving [41]. Recent post-training methods further optimize diverse skill-conditioned behaviors through fine-grained environment feedback, shifting skill usage from static prompts toward learnable components of agent policies [32, 55].

*Both authors contributed equally to this research.
†Corresponding author.

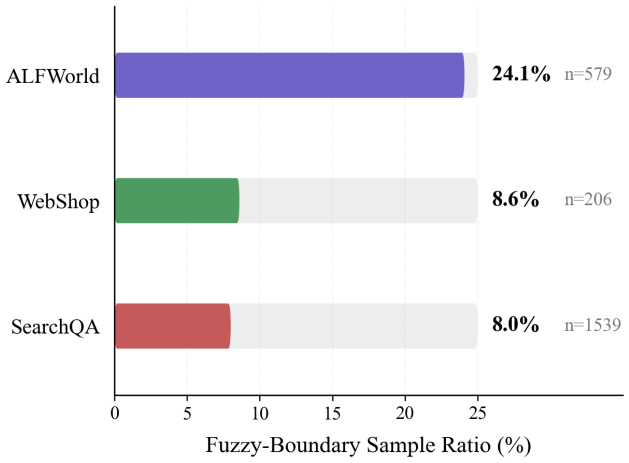


**Figure 1: Empirical evidence of ambiguous medium/easy skill-routing boundaries in Skill0.5. Samples across the success threshold can differ by as little as $1/8$, revealing the ambiguity of threshold-based routing.**

Existing skill-based agent learning methods can be broadly organized into three paradigms. The first paradigm fully externalizes skills: the agent keeps a persistent skill library outside the model and retrieves or updates skills during task execution, as in Voyager [37], SkillRL [41], and Skill1 [32]. This design preserves interpretability, modularity, and compositional reuse, but also introduces context overhead and retrieval noise. The second paradigm fully internalizes skills into the model parameters, aiming to eliminate inference-time skill prompts and enable autonomous execution, as in SKILL0 [25], LatentSkill [48], and related skill-to-parameter approaches [51]. This improves token efficiency and reduces dependence on external memory, but may lose the controllability and task-specific adaptability of explicit skills. The third paradigm jointly externalizes and internalizes skills. Skill0.5 [55] internalizes generalizable skills into the model while retaining explicit task-specific skill utilization, demonstrating that such a hybrid strategy can better balance cross-task transfer, inference efficiency, and task-level specialization across diverse environments.

Despite this progress, existing hybrid methods still leave a key optimization gap. Skill0.5 relies on task-level success rates to route samples into difficulty tiers and then applies different objectives to hard, medium, and easy tasks. Such routing makes the boundary between internalization and utilization sensitive to noisy aggregate outcomes: samples with very similar empirical success rates may receive different training treatment simply because they fall on opposite sides of a threshold. As shown in Figure 1, this issue is not rare in practice: across ALFWorld, WebShop, and SearchQA, a substantial number of samples lie around the medium/easy boundary, where their pass rates differ by at most one rollout unit but they are assigned to different optimization objectives. This observation suggests that the inherent boundaries introduced by trajectory-level routing motivate us to move beyond explicit skill selection and pursue a more unified learning paradigm, where skills are first internalized into the policy and then naturally utilized during subsequent interaction. More importantly, trajectory-level skill routing is broad and stable, but may be too coarse for pivotal states where skill guidance changes the quality of a specific decision. In contrast, action-level skill signals are more precise, but are harder to obtain because the environment usually provides only trajectory-level rewards. Such coarse outcome feedback cannot directly reveal which individual actions benefit from skill conditioning and which actions are distracted by it, echoing the credit-assignment challenge in reinforcement learning [29, 31]. As a result, existing methods struggle to optimize skill use at actual granularity of agent decision making both in skills internalization and utilization.

To address these limitations, we propose **AUSO (Action-level Unified Skill Optimization)**, a progressive reinforcement learning framework that unifies skill internalization and skill utilization through action-aware optimization. At its core, AUSO employs Jensen–Shannon divergence (JSD) as a unified information-gain signal to quantify how skill guidance changes the policy's action distribution, supporting both skill internalization and utilization at the action level. AUSO first allows the policy to learn from both teacher-provided skill guidance and environment outcomes, then consolidates autonomous problem-solving ability through outcome-based policy optimization, and finally evaluates each sampled action under both skill-conditioned and skill-free contexts. By coupling this action-level skill-sensitivity signal with trajectory-level advantage, AUSO strengthens updates for actions that genuinely benefit from skills while suppressing those where skill guidance is harmful. In this way, skills gradually transition from external supervision into decision knowledge utilized according to their action-level benefits, while reinforcement learning remains the shared backbone throughout the agent's policy evolution. Extensive experiments demonstrate that AUSO consistently outperforms strong baselines such as Skill0.5 across WebShop, ALFWorld, and Search-QA, showing robust improvements under both ID and OOD settings on WebShop and ALFWorld, as well as across single-hop and multi-hop tasks on Search-QA which shows effectiveness and generalization ability of AUSO across diverse agent tasks and skill settings. Our contributions are as follows:

- We propose AUSO, an Action-level Unified Skill Optimization, which unifies skill internalization and utilization in a progressive reinforcement learning process, enabling skills to evolve from external supervision into decision knowledge utilized according to their action-level benefits as the ability of policy model improves gradually.
- We introduce **action-level skill-sensitive optimization**, which unifies skill internalization and utilization at the action level under a shared JSD-based information-gain measure. It quantifies action-distribution discrepancies induced by skill guidance to facilitate the internalization of generalizable skills and subsequently guide adaptive skill utilization during interaction, overcoming the problem of coarse granularity of trajectory-level skill routing.
- We empirically validate our **AUSO** framework on ALFWorld, WebShop, and SearchQA, showing consistent improvements over competitive baselines in long-horizon agent performance and out-of-distribution generalization.

## 2 Related Work

### 2.1 LLM Agents

Large language models have increasingly been studied as interactive agents that solve long-horizon tasks through iterative reasoning, action execution, and feedback from external environments. Prompting methods such as chain-of-thought, ReAct, and Tree-of-Thoughts show that intermediate reasoning and deliberate search can substantially improve complex decision making [39, 45, 46]. Tool-augmented agents further extend LLMs beyond text-only generation by enabling API calls, search, calculation, and other external operations [20, 28]. These capabilities have motivated a broad set of agent benchmarks covering web navigation, online shopping, embodied household tasks, software engineering, and general agent evaluation [7, 10, 22, 34, 44, 54].

Another line of work improves LLM agents through memory [9, 47], reflection, experience reuse, or reinforcement learning from interaction. Reflexion uses verbal feedback from previous failures to guide future decisions without updating model parameters [33], while Voyager constructs reusable programs and curriculum-driven exploration in open-ended embodied environments [37]. Recent work on agentic reinforcement learning further emphasizes that post-training LLMs for interactive environments differs from standard static preference optimization [12, 14], because agents must collect trajectories, interact with environment states, and optimize behavior under delayed outcomes [7, 49]. However, most LLM-agent methods still treat external guidance, memory, and reusable knowledge as auxiliary modules rather than modeling how such knowledge should gradually transition into the policy itself. AUSO addresses this gap by optimizing skill learning and skill use in a unified reinforcement learning cycle process.

### 2.2 Agent Skills

Agent skills provide reusable procedural knowledge for solving complex tasks. In embodied and robotic settings, skills often correspond to executable primitives, affordance-aware subroutines, or programmatic task plans that ground high-level language instructions into feasible actions [2, 35]. In LLM agents, skills are commonly represented as natural-language strategies, environment-specific heuristics, tool-use patterns, or executable procedures stored in external libraries [37, 41]. This external skill paradigm improves interpretability and modular reuse, but it also introduces retrieval noise, context overhead, and difficulty in deciding which skills should be loaded for a particular decision [15, 17, 18].

Recent studies make this bottleneck explicit. Skill Retrieval Augmentation formulates large-scale skill retrieval as a distinct problem and shows that agents often struggle to determine when retrieved skills should actually be incorporated [36]. SkillComposer studies structured skill composition, where an agent must select not only which skills to use but also how many and in what order [53]. Skill-usage benchmarking further shows that skill benefits can degrade under realistic retrieval and refinement settings, especially when skills are noisy, mismatched, or not directly tailored to the task [23]. These works indicate that the external skills are useful but fragile: their effectiveness for dealing with complex tasks depends on accurate retrieval, composition, and selective use.

A complementary direction attempts to internalize skills into model parameters. SKILL0 trains agents to absorb skill knowledge and reduce dependence on explicit skill prompts at inference time [25], while LatentSkill and Skill-to-LoRA study parameter or latent-space forms of skill internalization [48, 51]. Other recent work studies continual experience internalization and test-time skill evolution, showing that the granularity and timing of experience injection are important for stable long-horizon agent learning [6, 26]. Between full externalization and full internalization, Skill0.5 jointly considers skill utilization and internalization through difficulty-aware routing [55]. AUSO follows this hybrid motivation, but differs by unifying skill internalization and utilization within a single action-level optimization framework, replacing trajectory-level routing with skill-sensitive learning that progressively internalizes generalizable skills and adapts skill utilization according to its benefit for each action.

## 3 Method

### 3.1 Problem Formulation

We consider a skill-augmented interactive agent that solves long-horizon tasks through multi-step interaction with an environment. Each task is formulated as an episodic decision process. At time step $t$, the agent observes an interaction history state $h_t = (o_1, a_1, \ldots, o_t)$ and generates an action $a_t \sim \pi_\theta(\cdot \mid h_t)$, where $\pi_\theta$ denotes the policy parameterized by an LLM. The environment then returns the next observation $o_{t+1}$, and the episode terminates after $T$ steps or when the task is completed. A trajectory is denoted as $\tau = (o_1, a_1, \ldots, o_T, a_T)$, and receives a trajectory-level outcome reward $R(\tau)$, such as task success or task score.

In skill-based agent learning, the policy may additionally condition on a skill $s$, which provides reusable procedural knowledge, task-solving guidance, or environment-specific strategies. We denote the skill-conditioned policy calculated as:

$$\pi_\theta^+(a_t \mid h_t, s), \tag{1}$$

and the policy without using a skill calculated as:

$$\pi_\theta^-(a_t \mid h_t). \tag{2}$$

Existing methods often decide whether a trajectory should be used for skill internalization or skill utilization according to empirical success rates estimated within a sliding window. However, such trajectory-level routing introduces fuzzy decision boundaries and assigns a uniform training role to all actions in the same trajectory, even though skill guidance may be beneficial for some actions but harmful for others. Our goal is therefore to train an agent whose skill learning follows a progressive lifecycle: the policy first internalizes useful skills, then improves through autonomous exploration, and ultimately learns to assess the benefit of skill guidance for each action rather than relying on coarse task-level routing. To realize this objective, we formulate skill-based agent training as a unified progressive policy optimization problem. Instead of selecting skill internalization or skill utilization through hard trajectory-level success-rate thresholds, AUSO organizes the learning process into a continuous policy evolution: the agent first internalizes useful skill knowledge, then improves through autonomous exploration, and finally learns to assess the benefit of skill guidance for each

action and adapt its utilization accordingly. To support this, AUSO further introduces action-level skill-aware weighting via JSD, so different actions receive different update strengths according to their outcome contribution and dependence on skill guidance.

## 3.2 ID and OOD settings

To evaluate our method under different generalization scenarios, we follow the ID/OOD setting of Skill0.5. Specifically, the task domain space is partitioned into ID domains $\mathcal{D}_{\text{id}}$ and OOD domains $\mathcal{D}_{\text{ood}}$, with their corresponding task sets denoted as $\mathcal{X}^{\text{id}}$ and $\mathcal{X}^{\text{ood}}$. Accordingly, we distinguish the skill space into general skills $\mathcal{S}_G$, which capture transferable knowledge shared across domains, and domain-specific skills $\mathcal{S}_S$. The latter are further partitioned into ID-specific skills $\mathcal{S}_S^{\text{id}}$ and OOD-specific skills $\mathcal{S}_S^{\text{ood}}$ according to their associated domains. During training, the agent only interacts with ID tasks $x \sim \mathcal{X}^{\text{id}}_{\text{train}}$, together with the accessible general skills $\mathcal{S}_G$ and ID-specific skills $\mathcal{S}_S^{\text{id}}$, while OOD tasks $\mathcal{X}^{\text{ood}}$ and their corresponding skills $\mathcal{S}_S^{\text{ood}}$ remain unseen. During evaluation, we evaluate the agent on both ID and previously unseen OOD tasks to assess its in-domain performance and cross-domain skill generalization. We adopt this setting for both ALFWorld and WebShop. Since Search-QA does not support the same ID/OOD partition, we follow the original setting of Skill0 for its evaluation.

## 3.3 Overview of Action-level Unified Skill Optimization

To train an agent whose skill learning follows a progressive lifecycle and enables action-level skill-aware optimization, we propose Action-level Unified Skill Optimization (AUSO) framework. Instead of treating skill internalization and skill utilization as two disjoint objectives selected by hard trajectory-level routing, AUSO keeps reinforcement learning via GRPO as the shared optimization backbone and changes how skill information is injected into the update. In the early stage, skills provide teacher-guided supervision weighted by action-level information gain [38], so that states where the student deviates more from skill-conditioned teacher behavior receive stronger corrective updates. In the later satge, as the policy becomes more capable through self-exploration, AUSO transitions from externally guided skill internalization to self-guided skill utilization, where the action-level influence of skill conditioning is measured via JSD and used to adaptively reweight policy gradients.

*3.3.1 Early-stage Teacher-guided Skill Internalization.* In the early stage, the main challenge is that outcome-only RL may provide no learning signal for difficult tasks. When all sampled trajectories fail, the group-normalized advantage becomes nearly zero according to the advantage of GRPO calculating algorithm:

$$A_{q,i}^{\text{GRPO}} = \frac{R_{q,i} - \bar{R}_q}{\sigma_q + \epsilon} \approx 0, \quad \text{if } R_{q,1} = \cdots = R_{q,G} = 0 \tag{3}$$

where $R_{q,i}$ means the $q^{th}$ trajectory and $i^{th}$ step in the trajectory. In this case, the policy cannot infer which actions should be corrected from the outcome reward alone. AUSO therefore introduces a teacher-guided action distribution supervision term, but only for such no-signal task groups.

For each time step t, the agent observes the history state:

$$h_t = (o_0, a_0, \ldots, o_t) \tag{4}$$

Given the current observation state $h_t$, we aim to internalize skill knowledge into the model rather than relying on explicit skill retrieval during inference. To this end, we construct a teacher–student paradigm from the same model without requiring additional environment rollouts. Specifically, the teacher is conditioned on the available general skills $\mathcal{S}_G$, while the student operates solely on the original observation without access to any skills. Accordingly, we compute two action distributions:

$$\pi_S(\cdot \mid h_t), \qquad \pi_T(\cdot \mid h_t, \mathcal{S}_G) \tag{5}$$

where $\pi_T$ represents the teacher distribution with $S_G$that provides guidance for transferring the generalizable knowledge encoded in $\mathcal{S}_G$, while $\pi_S$ denotes the skill-free student distribution. The teacher distribution is detached during optimization, such that the supervision exclusively updates the student model, progressively internalizing the general skills into its policy parameters.

We measure the discrepancy as the information gain between teacher and student by Jensen-Shannon divergence:

$$D_{t,k}^{T\to S} = \text{JSD}\left(\pi_T(\cdot \mid h_t, C_{\text{teacher}}) \,\|\, \pi_S(\cdot \mid h_t)\right) \tag{6}$$

computed over the top-$k$ teacher tokens. Since one environment action may contain multiple generated tokens, we aggregate token-level divergences within the action span $\mathcal{A}_t$:

$$D_t^{T\to S} = \frac{1}{|\mathcal{A}_t|} \sum_{k \in \mathcal{A}_t} D_{t,k}^{T\to S} \tag{7}$$

This prevents longer textual actions or reasoning tokens from receiving disproportionately large supervision weight.

To make teacher supervision action-aware while keeping the loss scale stable, AUSO first normalizes the teacher-student discrepancy among valid actions in the same task group:

$$z_t = \frac{D_t^{T\to S} - \mu_q}{\sqrt{\sigma_q^2 + \epsilon}} \tag{8}$$

where $D_t^{T\to S}$ denotes the action-level teacher-student divergence at action $t$, and $\mu_q$ and $\sigma_q^2$ are the mean and variance of such divergences over valid actions in task group $q$. We then transform the normalized discrepancy $z_t$ into a centered and bounded modulation:

$$\hat{u}_t = \frac{\tanh(z_t) - \mathbb{E}_{t\in q}[\tanh(z_t)]}{\max_{t\in q} \left|\tanh(z_t) - \mathbb{E}_{t\in q}[\tanh(z_t)]\right| + \epsilon} \tag{9}$$

Here, $\tanh(\cdot)$ bounds the modulation, the expectation term recenters the weights within the task group, and the denominator normalizes the maximum magnitude so that $\hat{u}_t \in [-1, 1]$.

Finally, the action-level teacher supervision weight is defined as

$$u_t = 1 + \beta_{\text{int}} \hat{u}_t, \qquad u_t \in [1 - \beta_{\text{int}}, 1 + \beta_{\text{int}}] \tag{10}$$

Following the stabilization principle of clipped policy optimization [31], we cap $\beta_{\text{int}}$ with an action-level internalization clip, so that teacher-guided JSD only mildly reweights actions instead of changing the overall optimization scale. This prevents high-discrepancy

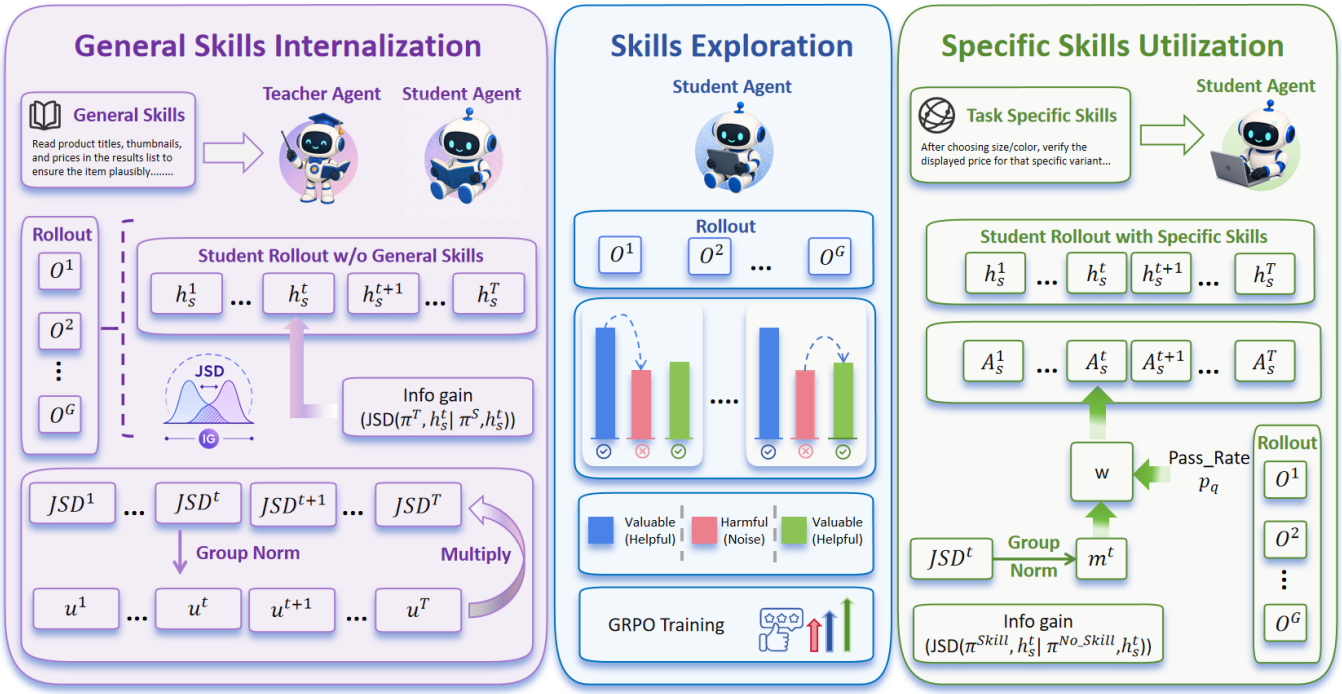


**Figure 2: Overview of Our AUSO (Action-Level Unified Skill Optimization from Internalization to Utilization): (1) General Skills Internalization: Internalize transferable general skills through action-level teacher–student guidance. (2) Skills Exploration: Explore and optimize diverse skills via GRPO. (3) Specific Skill Utilization: Evaluate the benefit of task-specific skills for each action to guide action-level skill utilization.**

actions from dominating the gradient while still assigning them stronger corrective supervision. The early-stage distillation loss is:

$$\mathcal{L}_{\text{JSD}} = \frac{1}{N} \sum_t u_t D_t^{T \to S} \tag{11}$$

We control the strength of teacher-guided internalization with a time-dependent coefficient:

$$\lambda_{\text{JSD}}(s) = \lambda_0 \alpha(s) \tag{12}$$

where $\lambda_0$ is the base JSD coefficient and $\alpha(s)$ determines how strongly teacher supervision contributes at training step $s$. Following the general principle of curriculum and scheduled supervision, auxiliary guidance should be introduced gradually and annealed as the learner becomes more capable [3, 4]. In AUSO, this is implemented with a short ramp-up period followed by a smooth decay: the teacher-guided signal is gradually activated during the first $r$ training steps, and then annealed as the policy shifts toward autonomous exploration. We therefore define $\alpha(s)$ as a ramp-up term multiplied by a smooth decay term:

$$\alpha(s) = \text{Ramp}(s) \cdot \left[1 - \text{SmoothStep}\left(\frac{s/T}{\rho_{\text{int}}}\right)\right] \tag{13}$$

where $s$ is the current training step, $T$ is the total number of training steps, and $\rho_{\text{int}}$ denotes the fraction of training allocated to the internalization phase. The ramp function is defined as:

$$\text{Ramp}(s) = \min\left(1, \frac{s}{r}\right) \tag{14}$$

where $r$ is the number of ramp-up steps. The smooth decay uses the standard SmoothStep function, a cubic Hermite interpolation commonly used for smooth transitions [19, 27]:

$$\text{SmoothStep}(x) = \tilde{x}^2(3 - 2\tilde{x}) \qquad \tilde{x} = \text{clip}(x, 0, 1) \tag{15}$$

With this schedule, teacher supervision is gradually activated at the beginning of training and then smoothly annealed, so that the policy first receives corrective skill guidance but is later encouraged to rely on environment-driven optimization. The early-stage objective is finally written as:

$$\mathcal{L}_{\text{early}} = \mathcal{L}_{\text{GRPO}} + \mathbb{I}[p_q = 0]\, \lambda_{\text{JSD}}(s) \mathcal{L}_{\text{JSD}} \tag{16}$$

where $p_q$ means the success rate of the student agent in task q.

*3.3.2 Outcome-driven Autonomous Exploration.* After early skill internalization, AUSO removes teacher-guided supervision and lets the agent improve through standard GRPO. In this stage, the policy samples multiple trajectories for each task and uses group-normalized outcome advantages to update itself:

$$\mathcal{L}_{\text{mid}} = \mathcal{L}_{\text{GRPO}} \tag{17}$$

This encourages the agent to consolidate autonomous problem-solving ability through environment-driven exploration.

*3.3.3 Later-stage Action-level Skill Utilization.* In the later stage, AUSO stops using the external teacher as a distillation target. The question is no longer how to imitate teacher behavior, but when

**Table 1: Performance comparison on WebShop and ALFWorld under in-distribution (ID) and out-of-distribution (OOD) settings.**

| Method | WebShop | | | | | | | | | ALFWorld | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ID | | | | | OOD | | | | ID | | | | OOD | | | |
| | App. | Elec. | Foot. | Other | Avg. | Access. | Beauty | Home | Avg. | Pick | Cool | Clean | Avg. | Look | Heat | Pick2 | Avg. |
| *Prompt-based Methods* | | | | | | | | | | | | | | | | | |
| Zero-shot | 4.4 | 4.6 | 3.4 | 1.0 | 3.5 | 2.5 | 3.7 | 5.5 | 3.9 | 28.6 | 12.0 | 18.5 | 20.7 | 38.5 | 12.5 | 12.5 | 18.9 |
| Few-shot | 14.2 | 15.1 | 13.5 | 24.0 | 16.5 | 18.8 | 29.6 | 5.5 | 16.9 | 62.9 | 44.0 | 63.0 | 57.5 | 46.2 | 31.2 | 8.3 | 24.5 |
| ReAct | 12.4 | 11.1 | 4.5 | 12.0 | 10.4 | 11.2 | 24.1 | 2.7 | 11.6 | 71.4 | 28.0 | 33.3 | 47.1 | 46.2 | 18.8 | 12.5 | 22.6 |
| Reflexion | 3.5 | 8.6 | 1.1 | 7.0 | 5.5 | 6.3 | 5.6 | 1.4 | 4.4 | 85.7 | 44.0 | 44.4 | 60.9 | 46.2 | 31.3 | 29.2 | 34.0 |
| Mem0 | 8.9 | 9.2 | 2.3 | 11.0 | 8.2 | 10.0 | 16.7 | 4.1 | 9.7 | 54.3 | 4.0 | 18.5 | 28.7 | 38.5 | 18.8 | 4.2 | 17.0 |
| ExpeL | 6.2 | 12.5 | 9.0 | 21.0 | 12.1 | 12.5 | 25.9 | 8.2 | 14.5 | 80.0 | 44.0 | 66.7 | 65.5 | 46.2 | 18.8 | 20.8 | 18.9 |
| MemP | 15.9 | 13.2 | 9.0 | 19.0 | 14.3 | 16.2 | 14.8 | 9.6 | 13.5 | 65.7 | 12.0 | 33.3 | 40.2 | 46.2 | 37.5 | 12.5 | 28.3 |
| SimpleMem | 11.5 | 13.8 | 6.7 | 13.0 | 11.7 | 10.0 | 20.4 | 5.5 | 11.1 | 71.4 | 16.0 | 44.4 | 47.1 | 53.8 | 18.8 | 20.8 | 28.3 |
| *RL-based Methods* | | | | | | | | | | | | | | | | | |
| RLOO | 34.9 | 23.9 | 41.5 | 32.1 | 31.1 | 31.4 | 46.5 | 22.5 | 32.9 | 91.4 | 80.0 | 81.5 | 85.1 | 61.5 | 56.3 | 20.8 | 41.5 |
| GRPO | 35.1 | 22.6 | 39.0 | 49.5 | 33.6 | 27.7 | 47.3 | 25.9 | 32.3 | 80.0 | 72.0 | 88.9 | 80.5 | 76.9 | 56.3 | 16.7 | 43.4 |
| *Memory-Augmented RL Methods* | | | | | | | | | | | | | | | | | |
| MemRL | 22.2 | 15.2 | 25.0 | 48.3 | 26.2 | 13.2 | 17.9 | 27.8 | 19.6 | 74.3 | 12.0 | 55.6 | 50.6 | 46.2 | 12.5 | 45.8 | 35.8 |
| EvolveR | 32.5 | 31.1 | 25.0 | 20.9 | 28.0 | 20.8 | 28.6 | 18.2 | 21.9 | 88.6 | 52.0 | 81.5 | 75.9 | 46.2 | 6.2 | 50.0 | 35.8 |
| Mem0+GRPO | 36.5 | 22.0 | 40.6 | 23.1 | 29.5 | 10.2 | 32.1 | 29.4 | 25.0 | 65.7 | 20.0 | 51.9 | 48.3 | 23.1 | 6.2 | 20.8 | 17.0 |
| SimpleMem+GRPO | 25.4 | 28.0 | 26.6 | 25.1 | 26.4 | 16.2 | 32.1 | 29.4 | 25.0 | 85.7 | 52.0 | 70.3 | 71.3 | 61.5 | 43.8 | 41.7 | 47.2 |
| *Skill-Augmented RL Methods* | | | | | | | | | | | | | | | | | |
| SkillRL | 36.0 | 34.2 | 41.4 | 49.3 | 38.1 | 36.3 | 48.5 | 27.6 | 36.7 | 91.4 | 84.0 | 96.3 | 90.8 | 69.2 | 75.0 | 12.5 | 45.3 |
| Skill0 | 39.2 | 33.0 | 38.1 | 37.9 | 35.2 | 42.1 | 38.6 | 26.5 | 35.4 | 94.3 | 76.0 | 81.5 | 85.1 | 46.2 | 50.0 | 29.2 | 39.6 |
| SLIM | 31.9 | 36.8 | 31.5 | 33.0 | 33.7 | 35.0 | 29.6 | 35.6 | 33.8 | 91.4 | 84.0 | 70.4 | 82.8 | 53.8 | 31.3 | 29.2 | 35.8 |
| Skill0.5 | 39.1 | 37.3 | 41.1 | 50.9 | 40.4 | 36.6 | 54.2 | 31.4 | 40.6 | 94.3 | 88.0 | 96.3 | 93.1 | 69.2 | 87.5 | 33.3 | 58.5 |
| **AUSO (Ours)** | 47.4 | 48.0 | 50.1 | 60.0 | 49.7 | 53.1 | 62.9 | 39.7 | 51.2 | 91.4 | 96.0 | 96.3 | 94.3 | 69.2 | 87.5 | 54.2 | 67.9 |

the student's own decision is meaningfully affected by skill information. For every state $H_t$ visited by the student, we compute two distributions from the same policy under two contexts:

$$\pi_\theta(\cdot \mid h_t, C = \text{Skill}), \qquad \pi_\theta(\cdot \mid h_t, C = \text{No-skill}) \tag{18}$$

Because both distributions are evaluated at the same visited state, this comparison does not require additional environment rollouts and avoids the mismatch that would arise from comparing two separately generated trajectories. Action-level skill information is:

$$I_t = \text{JSD}\left(\pi_\theta(\cdot \mid h_t, C = \text{Skill}) \,\|\, \pi_\theta(\cdot \mid h_t, C = \text{No-skill})\right) \tag{19}$$

Here, $I_t$ is not used as a distillation loss. It is an information measure: a larger value means that skill conditioning changes the student's next-action distribution more strongly at state $H_t$.

Since the raw scale of $I_t$ varies across tasks, AUSO standardizes it within each task group:

$$z_t^1 = \frac{I_t - \mu_q^t}{\sqrt{\sigma_{q,I}^2 + \epsilon}}, \qquad m_t = \tanh(\text{clip}(z_t^1, -3, 3)) \tag{20}$$

The normalized score $m_t \in [-1, 1]$ indicates whether an action is more or less skill-sensitive than the task-level average.

AUSO further modulates the action-level signal by a global uncertainty gate based on the normalized Bernoulli variance:

$$g(p_q) = K \cdot p_q(1 - p_q) \tag{21}$$

where K is a positive scaling constant. This form follows the common uncertainty principle that binary outcomes are most informative when their empirical probability is near 0.5 and least informative when the outcome is almost deterministic [30]. Thus, the gate is small when all rollouts fail or all rollouts succeed, because the group outcome provides little contrast for action-level credit assignment. We provide a detailed mathematical justification of this gate in Appendix A.1. It reaches its maximum when successful and failed rollouts coexist, where outcome differences are most informative for identifying which skill-sensitive actions matter. The final action weight is calculated as:

$$w_t = 1 + \beta(s) g(p_q) m_t \tag{22}$$

Same as the curriculum strategy in the early stage, we also introduce $\beta(s)$, which smoothly increases during training:

$$\beta(s) = \text{SmoothStep}\left(\frac{s - s_{\text{start}}}{T - s_{\text{start}}}\right) \tag{23}$$

where $s$ denotes the current training step, $s_{\text{start}}$ is the step at which utilization optimization stage is activated, and $T$ is the total number of training steps. The weighted advantage becomes:

$$A_{q,t}^{\text{AUSO}} = A_q^{\text{GRPO}} w_t. \tag{24}$$

This reweighting preserves the sign of the original RL advantage:

$$A_q^{\text{GRPO}} > 0 \Rightarrow A_{q,t}^{\text{AUSO}} > 0, \qquad A_q^{\text{GRPO}} < 0 \Rightarrow A_{q,t}^{\text{AUSO}} < 0 \tag{25}$$

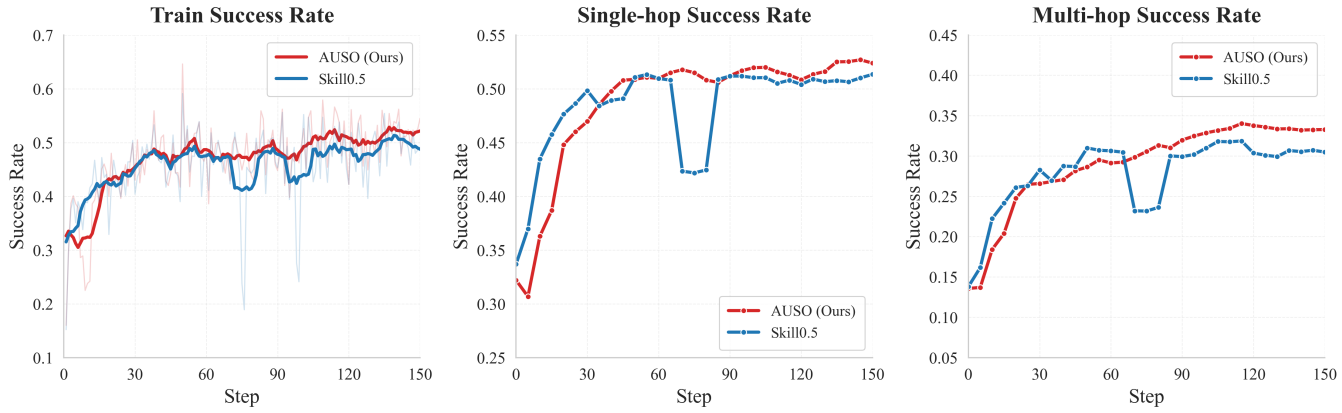


**Figure 3: Single-hop and multi-hop validation success rates and training success rates of our method and Skill0.5 with task-specific skills on Search QA benchmarks.**

Therefore, AUSO does not create artificial rewards or reverse the trajectory-level outcome signal. Instead, it redistributes update strength across actions according to the information gain: in successful trajectories, highly skill-sensitive actions receive stronger positive reinforcement; in failed trajectories, highly skill-sensitive actions receive stronger suppression; and low-information actions receive smaller updates. The later-stage objective is still standard on-policy RL, but with action-level skill-aware advantages:

$$\mathcal{L}_{\text{late}} = -\sum_{t} A^{\text{AUSO}}_{q,t} \log \pi_\theta(a_t \mid h_t, C = \text{Skill}) \tag{26}$$

*3.3.4 Unified Skill Lifecycle Objective.* The complete AUSO objective is formulated as:

$$\mathcal{L}_{\text{AUSO}}(\theta; s) = \mathcal{L}_{\text{GRPO}}\left(\theta; A^{\text{GRPO}}_{q,i}\left[1 + \beta(s)\, 4p_q(1-p_q)\, m_{q,i,t}\right]\right) + \lambda_0 \alpha(s) \sum_{q} \mathbb{I}[p_q = 0]\, \mathcal{L}^{(q)}_{\text{JSD}}. \tag{27}$$

Here, $\alpha(s)$ gradually removes teacher-guided supervision, while $\beta(s)$ progressively activates action-level skill utilization. AUSO retains outcome-driven GRPO as a persistent optimization backbone and progressively advances from external skill internalization to action-level skill utilization guided by skills benefit for each action.

# 4 Experiments

## 4.1 Experiment Settings

*4.1.1 Datasets.* We evaluate AUSO on three long-horizon agentic benchmarks: ALFWorld [34], WebShop [44], and SearchQA [11]. These benchmarks cover embodied decision-making, web-based interaction, and search-oriented question answering, allowing us to examine whether AUSO can improve both task ID performance and OOD (out-of-distribution) generalization across different forms of agent interacting with different environment.

**ALFWorld** is an embodied household task benchmark that aligns text-based environments with embodied manipulation scenarios. An agent must follow natural-language goals and complete multi-step household tasks through textual actions such as navigation, object interaction, and state manipulation. We use ALFWorld to evaluate whether skill guidance can improve long-horizon planning and action execution in embodied environments.

**Table 2: Performance comparison on Search-QA benchmark.**

| Method | Single-hop | | | Multi-hop | | | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | NQ | Triv | Pop | Hotp | 2Wk | MuS | Bam | |
| Zero-Shot | 10.4 | 32.4 | 22.3 | 15.8 | 15.4 | 7.2 | 19.2 | 17.5 |
| Few-Shot$^{\dagger}$ | 12.3 | 36.8 | 24.5 | 17.7 | 18.2 | 6.5 | 24.8 | 20.1 |
| Zero-Shot$^{\star}$ | 6.9 | 30.4 | 12.0 | 10.5 | 9.1 | 5.5 | 24.0 | 14.0 |
| Few-Shot$^{\star\dagger}$ | 10.5 | 31.9 | 18.7 | 14.2 | 14.4 | 6.9 | 24.8 | 17.3 |
| GRPO | 45.1 | 63.7 | 44.0 | 43.6 | 43.2 | 16.8 | 37.6 | 41.9 |
| OPSD | 8.8 | 8.6 | 17.5 | 2.5 | 4.2 | 1.2 | 6.2 | 4.5 |
| AgentOCR$^{*}$ | 43.1 | 61.0 | 45.4 | 40.8 | 38.3 | 15.7 | 36.8 | 40.1 |
| EvolveR | 43.5 | 63.4 | 45.9 | 38.2 | 42.0 | 15.6 | 54.4 | 43.1 |
| SkillRL | 45.9 | 63.3 | 45.9 | 43.2 | 40.3 | 20.2 | 73.8 | 47.1 |
| LatentSkill | 36.2 | 57.6 | 41.0 | 39.6 | 32.0 | 9.80 | 25.6 | 35.6 |
| Skill0 | 42.7 | 61.1 | 45.3 | 40.0 | 38.3 | 16.4 | 66.9 | 44.4 |
| Skill0.5 | 44.7 | 60.4 | 43.8 | 39.0 | 39.2 | 12.0 | 65.3 | 44.2 |
| Skill1 | 46.8 | 46.3 | 47.8 | 43.7 | 39.3 | 18.2 | 70.6 | 44.7 |
| **AUSO (Ours)** | 45.0 | 64.7 | 46.1 | 44.0 | 41.8 | 15.4 | 68.1 | 47.5 |

**WebShop** is a web-based shopping benchmark where an agent interacts with a simulated e-commerce website to find and purchase products satisfying user instructions. The agent must search, compare products, inspect attributes, and make purchase decisions through multi-step web actions. Following our out-of-distribution setting, we construct category-based ID and OOD splits according to Skill0.5: ID tasks are drawn from frequent categories such as apparel, electronics, footwear, and other products, while OOD validation tasks are drawn from held-out categories such as accessories, beauty and health, and home decor.

**SearchQA** is a search-based question answering benchmark where an agent must retrieve useful information and produce answers based on external evidence. Compared with embodied and web-shopping tasks, SearchQA emphasizes information seeking

**Table 3: Ablation studies of the average success rate on ALFWorld and WebShop to validate the effectiveness of all our core modules; the best performance of each metric is bolded, and the second-best result is underlined.**

| Variant | ALFWorld | | WebShop | |
|---|---|---|---|---|
| | ID Avg. | OOD Avg. | ID Avg. | OOD Avg. |
| w/o Int.-Action | 85.1 | <u>54.7</u> | 44.7 | <u>47.3</u> |
| w/o Ut.-Action | <u>89.7</u> | 49.1 | <u>47.5</u> | 45.9 |
| w/o Int.-Decrease | 83.9 | <u>54.7</u> | 43.9 | 44.4 |
| w/o Ut.-Increase | 86.2 | 39.6 | 47.3 | 43.5 |
| w/o Gate | 86.2 | 47.2 | 46.7 | 43.0 |
| **AUSO** | **94.3** | **67.9** | **49.7** | **51.2** |

**Table 4: Ablation Study for Period time for Internalize/Explore/Utilize in ALFWorld.**

| Variant | ALFWorld | |
|---|---|---|
| | ID Avg. | OOD Avg. |
| Ratio (3:5:2) | 85.1 | 62.3 |
| Ratio (3.3:3.3:3.3) | 87.4 | 41.5 |
| **AUSO (2:5:3)** | **94.3** | **67.9** |

and evidence aggregation, providing a complementary testbed for evaluating whether AUSO can improve skill use in knowledge-intensive decision processes.

*4.1.2 Baselines.* We compare AUSO with a diverse spectrum of agent learning methods. (1) **Prompt-based methods**: zero-shot and few-shot prompting, which directly condition the backbone model on task instructions without additional agent training [5]. (2) **Prompt-based agentic and memory-based methods**: ReAct [46], Reflexion [33], ExpeL [52], Mem0 [8], and SimpleMem [21], which improve multi-step decision making through in-context reasoning, verbal feedback, experiential memory, or external memory retrieval without updating the policy parameters. (3) **RL-based methods**: RLOO [1] and GRPO [31], which optimize LLM policies with outcome rewards and group-relative or leave-one-out advantage estimation. (4) **Memory-augmented RL methods**: MemRL [50] and Memory-R1 [42], which integrate reinforcement learning with persistent or learnable memory mechanisms. (5) **Skill-augmented RL methods**: SkillRL [41], Skill1 [32], SKILL0 [25], LatentSkill [48], and Skill0.5 [55], which represent recent attempts to train agents with external skills, internalized skills, or hybrid skill utilization and internalization.

*4.1.3 Implementation Details.* All experiments are conducted on a server equipped with **4 NVIDIA H200 GPUs**. Following Skill0.5 [55], we adopt **Qwen2.5-7B-Instruct** [43] as the backbone policy model for all experiments. For a fair comparison, all methods use the same backbone model, rollout configuration, maximum interaction length, and optimization budget within each benchmark. Specifically, for each task $q$, the policy performs $G = 8$ rollouts to construct a rollout group for group-based policy optimization. Across

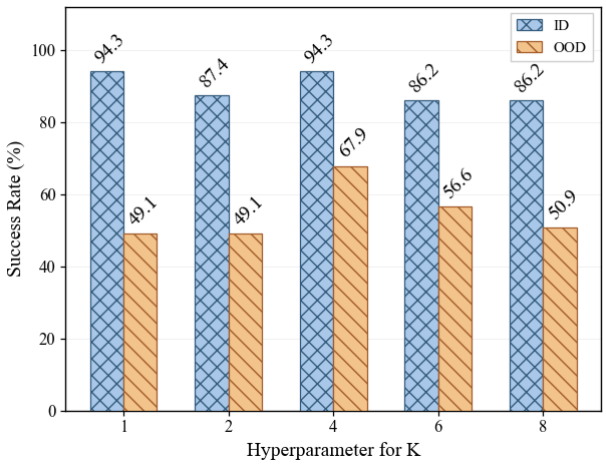


**Figure 4: The ID and OOD success rate for different hyperparameter $K$ in the ALFWorld benchmark.**

all benchmarks, including ALFWorld, WebShop, and SearchQA, we train each method for **150 optimization steps**. For ALFWorld and WebShop, we follow the category-based ID/OOD evaluation protocol and report both in-distribution (ID) validation performance and out-of-distribution (OOD) generalization performance. For SearchQA, which does not adopt an ID/OOD split, we report task success following Skill0 to evaluate. More training and evaluation details are provided in the supplemental materials.

## 4.2 Main Results

Tables 1 and 2 report the main results on WebShop, ALFWorld, and Search-QA. Overall, **AUSO consistently achieves the strongest performance across heterogeneous agent scenarios**. More importantly, its improvements are observed under both ID and OOD settings, demonstrating effective skill utilization and generalization beyond seen tasks and unseen tasks.

*WebShop.* On WebShop, AUSO achieves the best average performance under both ID and OOD settings, reaching **49.7** and **51.2**, respectively. Compared with Skill0.5, AUSO improves the two averages by **9.3** and **10.6** points. The gains are particularly pronounced on Electronics and Other under ID, as well as Access and Beauty under OOD. Notably, the larger improvement under OOD suggests that AUSO does not simply fit task-specific skill patterns. Instead, action-level optimization enables the agent to internalize useful skill knowledge and adapt its utilization according to the benefit of skill guidance for each action, resulting in stronger generalization to the unseen (OOD) tasks.

*ALFWorld.* On ALFWorld, AUSO achieves an ID average of **94.3**, outperforming Skill0.5 by **1.2** points and obtaining the best overall performance. More importantly, AUSO reaches an OOD average of **67.9**, substantially surpassing Skill0.5 (**58.5**) by **9.4** points. The improvement is especially clear on challenging unseen tasks such as *Pick2*, where the score increases from 33.3 to **54.2**. These results

demonstrate that action-level skill utilization optimization facilitates the acquisition of transferable behavioral knowledge rather than merely fitting seen task configurations.

*Search-QA.* The effectiveness of AUSO further extends to the knowledge-intensive reasoning tasks. AUSO achieves the highest overall average of **47.5**, outperforming SkillRL (**47.1**), Skill1 (**44.7**), and Skill0.5 (**44.2**). It achieves the best performance on TriviaQA (**64.7**) and HotpotQA (**44.0**), while remaining competitive across other single-hop and multi-hop tasks. These results demonstrate that AUSO generalizes beyond interactive environments and transfers action-level skill knowledge to diverse reasoning tasks.

*Training and Generalization Analysis.* Fig. 3 further compares the learning dynamics of AUSO and Skill0.5 on Search-QA. While the two methods exhibit comparable training success rates, AUSO achieves stronger validation performance as training progresses, particularly on multi-hop tasks where Skill0.5 shows noticeable fluctuations. This suggests that the advantage of AUSO does not simply come from better fitting training trajectories; instead, action-level optimization promotes more effective skill internalization and utilization, leading to improved generalization.

## 4.3 Ablation Study

We conduct comprehensive ablation studies on ALFWorld and WebShop to investigate the contribution of each component in AUSO. As shown in Table 3, removing any component consistently degrades the overall performance, demonstrating that different stages of action-level skill optimization are complementary. In particular, removing *Int.-Action* (Action-Level Internalization) leads to a substantial drop on ALFWorld, from **94.3/67.9** to 85.1/54.7 in ID/OOD settings, highlighting the importance of skill internalization. Similarly, removing *Ut.-Action* (Action-Level Utilization) or the adaptive increase/decrease strategies in utilization/internalization results in consistent degradation, indicating that dynamically controlling skill utilization is crucial for transferring learned skills to different interaction states. Removing the gating mechanism also causes clear performance drops on both benchmarks, further validating its role in coordinating action-level skill optimization.

We further investigate the period allocation among different optimization stages in Table 4. The adopted ratio of **2:5:3** achieves the best ID and OOD performance, suggesting that an appropriate balance between skill internalization, exploration, and utilization is important for effective policy optimization. Finally, Fig. 4 studies the influence of hyperparameter $K$. Performance peaks at $K = 4$, achieving **94.3** ID and **67.9** OOD success rates, while either smaller or larger values lead to degradation, especially under OOD evaluation. This indicates that a moderate $K$ provides an effective balance for skill optimization and generalization.

## 5 Conclusion

In this paper, we proposed AUSO, an information gain based action-level via JSD unified skill optimization framework that unifies skill internalization and skill utilization for LLM agents. AUSO progressively learns from external skill guidance, consolidates autonomous decision-making through reinforcement learning, and finally uses action-level skill-sensitive signals to emphasize decisions that benefit from skills while suppressing those where skill guidance is unhelpful. Experiments on ALFWorld, WebShop, and SearchQA demonstrate that AUSO consistently outperforms competitive baselines and achieves stronger out-of-distribution generalization. Further ablation studies verify the contribution of progressive scheduling, action-level optimization, and the uncertainty-based gate. These results show that modeling skills at the action level provides an effective way to transform external guidance into transferable policy knowledge for long-horizon llm agents.

## References


[1] Arash Ahmadian, Chris Cremer, Quentin Gallouedec, et al. 2024. Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. *arXiv preprint arXiv:2402.14740* (2024).

[2] Michael Ahn, Anthony Brohan, Noah Brown, et al. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. In *CoRL*.

[3] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In *NeurIPS*.

[4] Yoshua Bengio, Jerome Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. In *ICML*.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

[6] Jingwen Chen, Wenkai Yang, Shengda Fan, et al. 2026. Rethinking Continual Experience Internalization for Self-Evolving LLM Agents. *arXiv preprint arXiv:2606.04703* (2026).

[7] Wanyi Chen, Xiao Yang, Xu Yang, et al. 2026. Agent$^2$ RL-Bench: Can LLM Agents Engineer Agentic RL Post-Training? *arXiv preprint arXiv:2604.10547* (2026).

[8] Prateek Chhikara, Ziyang Xu, Charles Packer, et al. 2025. Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. *arXiv preprint arXiv:2504.19413* (2025).

[9] Hao Cong, Huizu Lin, Zihan Wang, Chengkai Huang, Quan Z. Sheng, and Lina Yao. 2026. Seeing and Reflecting: Multimodal Memory-Enhanced Agent Collaboration for Recommendation. *arXiv preprint arXiv:2607.07108* (2026).

[10] Xiang Deng, Yu Gu, Boyuan Zheng, et al. 2023. Mind2Web: Towards a Generalist Agent for the Web. *arXiv preprint arXiv:2306.06070* (2023).

[11] Matthew Dunn, Levent Sagun, Mike Higgins, et al. 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. *arXiv preprint arXiv:1704.05179* (2017).

[12] Tianqi Gao, Chengkai Huang, Zihan Wang, Cao Liu, Ke Zeng, and Lina Yao. 2026. Factorized Latent Reasoning for LLM-based Recommendation. *arXiv preprint arXiv:2604.26760* (2026).

[13] Chengkai Huang, Junda Wu, Yu Xia, Zixu Yu, Ruhan Wang, Tong Yu, Ruiyi Zhang, Ryan A Rossi, Branislav Kveton, Dongruo Zhou, et al. 2025. Towards agentic recommender systems in the era of multimodal large language models. *arXiv preprint arXiv:2503.16734* (2025).

[14] Chengkai Huang, Junda Wu, Zhouhang Xie, Yu Xia, Rui Wang, Tong Yu, Subrata Mitra, Julian McAuley, and Lina Yao. 2025. Pluralistic off-policy evaluation and alignment. *arXiv preprint arXiv:2509.19333* (2025).

[15] Chengkai Huang, Yu Xia, Rui Wang, Kaige Xie, Tong Yu, Julian McAuley, and Lina Yao. 2025. Embedding-informed adaptive retrieval-augmented generation of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*. 1403–1412.

[16] Hongtao Huang, Chengkai Huang, Junda Wu, Tong Yu, Julian McAuley, and Lina Yao. 2026. Listwise preference diffusion optimization for user behavior trajectories prediction. *Advances in Neural Information Processing Systems* 38 (2026), 159383–159408.

[17] Shuguang Jiao, Chengkai Huang, Shuhan Qi, Xuan Wang, Yifan Li, Quanchi Weng, Lingchuan Liu, Xunliang Cai, and Lina Yao. 2026. Doctor-RAG: A Failure-Aware Repair Framework for Agentic Retrieval-Augmented Generation. *arXiv preprint arXiv:2604.00865* (2026).

[18] Shuguang Jiao, Xinyu Xiao, Yunfan Wei, Shuhan Qi, Chengkai Huang, Quan Z Sheng, and Lina Yao. 2026. Prunerag: Confidence-guided query decomposition trees for efficient retrieval-augmented generation. In *Proceedings of the ACM Web Conference 2026*. 1923–1934.

[19] Khronos Group. 2026. SYCL Reference: smoothstep. https://github.khronos.org/SYCL_Reference/iface/common-functions.html. Accessed 2026-08-11.

[20] Minghao Li, Feifan Song, Bowen Yu, et al. 2023. API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs. *arXiv preprint arXiv:2304.08244* (2023).

[21] Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. 2026. SimpleMem: Efficient Lifelong Memory for LLM Agents. *arXiv preprint arXiv:2601.02553* (2026).

[22] Xiao Liu, Hao Yu, Hanchen Zhang, et al. 2023. AgentBench: Evaluating LLMs as Agents. *arXiv preprint arXiv:2308.03688* (2023).

[23] Yujian Liu, Jiabao Ji, An Li, et al. 2026. How Well Do Agentic Skills Work in the Wild: Benchmarking LLM Skill Usage in Realistic Settings. *arXiv preprint arXiv:2604.04323* (2026).

[24] Haowei Lou, Chengkai Huang, Hye-young Paik, Yongquan Hu, Aaron Quigley, Wen Hu, and Lina Yao. 2025. SpeechAgent: An end-to-end mobile infrastructure for speech impairment assistance. *arXiv preprint arXiv:2510.20113* (2025).

[25] Zhengxi Lu, Zhiyuan Yao, Jinyang Wu, et al. 2026. SKILL0: In-Context Agentic Reinforcement Learning for Skill Internalization. *arXiv preprint arXiv:2604.02268* (2026).

[26] Bo Mao, Jie Zhou, Yutao Yang, et al. 2026. Learning While Acting: A Skill-Enhanced Test-Time Co-Evolution Framework for Online Lifelong Learning Agents. *arXiv preprint arXiv:2606.04815* (2026).

[27] Ken Perlin. 1985. An Image Synthesizer. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*. 287–296.

[28] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, et al. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. In *NeurIPS*.

[29] John Schulman, Filip Wolski, Prafulla Dhariwal, et al. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[30] Burr Settles. 2009. *Active Learning Literature Survey.* Technical Report. University of Wisconsin–Madison.

[31] Zhihong Shao, Peiyi Wang, Qihao Zhu, et al. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300* (2024).

[32] Yaorui Shi, Yuxin Chen, Zhengxi Lu, et al. 2026. Skill1: Unified Evolution of Skill-Augmented Agents via Reinforcement Learning. *arXiv preprint arXiv:2605.06130* (2026).

[33] Noah Shinn, Federico Cassano, Edward Berman, et al. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *NeurIPS*.

[34] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, et al. 2021. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *ICLR*.

[35] Ishika Singh, Valts Blukis, Arsalan Mousavian, et al. 2023. ProgPrompt: Program Generation for Situated Robot Task Planning Using Large Language Models. *Autonomous Robots* (2023).

[36] Weihang Su, Jianming Long, Qingyao Ai, et al. 2026. Skill Retrieval Augmentation for Agentic AI. *arXiv preprint arXiv:2604.24594* (2026).

[37] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, et al. 2023. Voyager: An Open-Ended Embodied Agent with Large Language Models. *arXiv preprint arXiv:2305.16291* (2023).

[38] Xihang Wang, Zihan Wang, Chengkai Huang, Quan Z Sheng, and Lina Yao. 2026. MEG-RAG: Quantifying Multi-modal Evidence Grounding for Evidence Selection in RAG. *arXiv preprint arXiv:2604.24564* (2026).

[39] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

[40] B. L. Welch. 1947. The Generalization of "Student's" Problem when Several Different Population Variances are Involved. *Biometrika* 34, 1–2 (1947), 28–35.

[41] Peng Xia, Jianwen Chen, Hanyang Wang, et al. 2026. SkillRL: Evolving Agents via Recursive Skill-Augmented Reinforcement Learning. *arXiv preprint arXiv:2602.08234* (2026).

[42] Sikuan Yan, Xiufeng Yang, Zuchao Huang, et al. 2026. Memory-R1: Enhancing Large Language Model Agents to Manage and Utilize Memories via Reinforcement Learning. In *ACL*.

[43] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).

[44] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. *arXiv preprint arXiv:2207.01206* (2022).

[45] Shunyu Yao, Dian Yu, Jeffrey Zhao, et al. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *arXiv preprint arXiv:2305.10601* (2023).

[46] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *ICLR*.

[47] Juexiang Ye, Xue Li, Yang Xinyu, Chengkai Huang, Lanshun Nie, Lina Yao, and Dechen Zhan. 2026. MemWeaver: Weaving Hybrid Memories for Traceable Long-Horizon Agentic Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*. 12928–12956.

[48] Aofan Yu, Chenyu Zhou, Tianyi Xu, et al. 2026. LatentSkill: From In-Context Textual Skills to In-Weight Latent Skills for LLM Agents. *arXiv preprint arXiv:2606.06087* (2026).

[49] Guibin Zhang, Hejia Geng, Xiaohang Yu, et al. 2025. The Landscape of Agentic Reinforcement Learning for LLMs: A Survey. *arXiv preprint arXiv:2509.02547* (2025).

[50] Shengtao Zhang, Jiaqian Wang, Ruiwen Zhou, Junwei Liao, Yuchen Feng, Zhuo Li, Yujie Zheng, Weinan Zhang, Ying Wen, Zhiyu Li, Feiyu Xiong, Yutao Qi, Bo Tang, and Muning Wen. 2026. MemRL: Self-Evolving Agents via Runtime Reinforcement Learning on Episodic Memory. *arXiv preprint arXiv:2601.03192* (2026).

[51] Tianyi Zhang and Zhonghao Qi. 2026. Skill-to-LoRA: From Using Skills to Learning Behaviors for Token-Efficient LLM Agents. *arXiv preprint arXiv:2606.16769* (2026).

[52] Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. ExpeL: LLM Agents Are Experiential Learners. In *AAAI*.

[53] Xinyu Zhao, Zhen Tan, Vaishnav Tadiparthi, et al. 2026. Generative Skill Composition for LLM Agents. *arXiv preprint arXiv:2606.32025* (2026).

[54] Shuyan Zhou, Frank F. Xu, Hao Zhu, et al. 2023. WebArena: A Realistic Web Environment for Building Autonomous Agents. *arXiv preprint arXiv:2307.13854* (2023).

[55] Jiapeng Zhu, Jianxiang Yu, Yibo Zhao, et al. 2026. Skill0.5: Joint Skill Internalization and Utilization for Out-of-Distribution Generalization in Agentic Reinforcement Learning. *arXiv preprint arXiv:2605.28424* (2026).


# A Training Algorithm

## A.1 Analysis of the Global Uncertainty Gate

*Statistical Justification of the Uncertainty Gate.* We provide a statistical interpretation of the global uncertainty gate $g(p_q) = K \cdot p_q(1-p_q)$ used in the skill utilization stage of AUSO. The key intuition is that action-level credit assignment is more reliable when both successful and failed trajectories are sufficiently represented within the rollout group. For a task $q$, suppose the rollout group contains $G$ trajectories, with empirical success rate $p_q$. The expected numbers of successful trajectories $G_+$ and failed trajectories $G_-$ are therefore be represented as:

$$G_+ = Gp_q, \qquad G_- = G(1-p_q) \tag{28}$$

Let $I$ denote the action-level skill information gain score. To determine whether skill-sensitive actions are associated with successful outcomes, consider the difference between the expected information scores of successful and failed trajectories:

$$\Delta_I = \mathbb{E}[I \mid R=1] - \mathbb{E}[I \mid R=0]. \tag{29}$$

Its empirical estimator is calculated as:

$$\widehat{\Delta}_I = \bar{I}_+ - \bar{I}_-, \tag{30}$$

where $\bar{I}_+$ and $\bar{I}_-$ denote the average action-level information scores computed from successful and failed trajectories, respectively.

Motivated by the standard variance analysis for the difference between two sample means [40], we analyze the statistical reliability of this success–failure information contrast. To this end, we consider the conditional variances of the action-level information score under successful and failed outcomes:

$$\sigma^2_{I,+} = \mathrm{Var}(I \mid R=1), \qquad \sigma^2_{I,-} = \mathrm{Var}(I \mid R=0) \tag{31}$$

The variance of the estimated information contrast can then be characterized as:

$$\begin{aligned} \mathrm{Var}(\widehat{\Delta}_I) &= \frac{\sigma^2_{I,+}}{n_+} + \frac{\sigma^2_{I,-}}{n_-} \\ &= \frac{\sigma^2_{I,+}}{Gp_q} + \frac{\sigma^2_{I,-}}{G(1-p_q)} \end{aligned} \tag{32}$$

To isolate the effect of the rollout outcome composition from task-dependent variations in the information-score variance, we consider the homoscedastic approximation:

$$\sigma^2_{I,+} \approx \sigma^2_{I,-} \approx \sigma^2_I \tag{33}$$

Under this approximation, Eq. (32) becomes:

$$\mathrm{Var}(\widehat{\Delta}_I) \approx \frac{\sigma_I^2}{G p_q(1-p_q)} \tag{34}$$

For a fixed rollout budget $G$ and information-score variance $\sigma_I^2$, the precision of the estimated success–failure contrast, defined as the inverse of its variance, therefore satisfies:

$$\mathrm{Precision}(\widehat{\Delta}_I) = \frac{1}{\mathrm{Var}(\widehat{\Delta}_I)} \propto p_q(1-p_q) \tag{35}$$

This result shows that the reliability of the success–failure contrast depends directly on the composition of rollout outcomes. When $p_q$ approaches either 0 or 1, one outcome group becomes absent or severely underrepresented, making the contrast increasingly unreliable.

*Boundedness and Direction Preservation.* Recall that the action-level modulation weight is:

$$w_t = 1 + \beta(s) g(p_q) m_t \tag{36}$$

where $|m_t| \le 1$ and $0 \le \beta(s) < 1$. Since $0 \le g(p_q) \le 1$, we have:

$$1 - \beta(s) \le w_t \le 1 + \beta(s) \tag{37}$$

Therefore, $w_t > 0$, and the resulting action-level advantage

$$A_{q,t}^{\mathrm{AUSO}} = w_t A_q^{\mathrm{GRPO}} \tag{38}$$

preserves the sign of the original GRPO advantage:

$$\mathrm{sign}\left(A_{q,t}^{\mathrm{AUSO}}\right) = \mathrm{sign}\left(A_q^{\mathrm{GRPO}}\right) \tag{39}$$

Moreover,

$$A_q^{\mathrm{GRPO}} = 0 \quad \Longrightarrow \quad A_{q,t}^{\mathrm{AUSO}} = 0 \tag{40}$$

so the action-level information signal cannot introduce an artificial optimization direction unsupported by the original outcome reward.

Finally, because the action-level modulation is centered within each task, $\mathbb{E}_{t\in q}[m_t] = 0$, we obtain:

$$\mathbb{E}_{t\in q}[w_t] = 1 + \beta(s) g(p_q) \mathbb{E}_{t\in q}[m_t] = 1 \tag{41}$$

Therefore, the global uncertainty gate does not alter the average optimization scale within a task. Instead, it controls the reliability of action-level modulation according to the available success–failure contrast, while $m_t$ redistributes the resulting credit across the individual actions of agent in the same trajectory.

## A.2 A Unified View of Internalization and Utilization

*Analysis of the Unified Procedure.* Although skill internalization and skill utilization play different roles during training, AUSO implements both through the same action-level counterfactual information operator. The distinction between the two stages lies not in how action-level information is measured, but in how the resulting information signal is used for optimization. For an action $a_{q,t}$ taken at interaction history $h_{q,t}$, let $C_{q,t}^+$ and $C_{q,t}^-$ denote two counterfactual context conditions. We define the unified action-level information operator for the task $q$ as:

$$\mathcal{I}_{q,t}\left(C_{q,t}^+, C_{q,t}^-\right) = \mathrm{JSD}\left(\pi_\theta\left(\cdot \mid h_{q,t}, C_{q,t}^+\right) \middle\| \pi_\theta\left(\cdot \mid h_{q,t}, C_{q,t}^-\right)\right) \tag{42}$$

Since one environment action may contain multiple generated tokens, the implemented action-level score averages token-level divergences over the action span:

$$\mathcal{I}_{q,t} = \frac{1}{|a_{q,t}|} \sum_{j=1}^{|a_{q,t}|} \mathrm{JSD}\left(\pi_\theta\left(\cdot \mid h_{q,t}, a_{q,t,<j}, C_{q,t}^+\right) \middle\| \pi_\theta\left(\cdot \mid h_{q,t}, a_{q,t,<j}, C_{q,t}^-\right)\right) \tag{43}$$

The two learning stages instantiate this operator with different counterfactual contexts:

$$\left(C_{q,t}^+, C_{q,t}^-\right) = \begin{cases} \left(C_q^{\mathrm{general}} \cup C_q^{\mathrm{specific}}, C_q^{\mathrm{specific}}\right), & x \in \mathcal{H}, \\ \left(C_q^{\mathrm{specific}}, \varnothing\right), & q \in \mathcal{E}. \end{cases} \tag{44}$$

Here, $\mathcal{H}$ and $\mathcal{E}$ denote the task regimes corresponding to skill internalization and skill utilization, respectively.

For tasks in $\mathcal{H}$, the information score measures the discrepancy between a privileged skill-conditioned teacher and the student. It is therefore used as a direct supervision signal:

$$\mathcal{L}_{\mathrm{int}}(q) = \frac{1}{T_q} \sum_{t=1}^{T_q} w_{q,t}^{\mathrm{int}} \mathcal{I}_{q,t}\left(C_q^{\mathrm{general}} \cup C_q^{\mathrm{specific}}, C_q^{\mathrm{specific}}\right) \tag{45}$$

For tasks in $\mathcal{E}$, the same operator instead measures how strongly the current policy depends on task-specific skill conditioning:

$$\mathcal{I}_{q,t}^{\mathrm{util}} = \mathcal{I}_{q,t}\left(C_q^{\mathrm{specific}}, \varnothing\right) \tag{46}$$

This information is not optimized directly. Instead, it is converted into the action-level modulation:

$$w_{q,t}^{\mathrm{util}} = 1 + \beta(s) g(p_q) m_{x,t}, \tag{47}$$

which reweights the original GRPO advantage:

$$\mathcal{L}_{\mathrm{util}}(q) = \mathcal{L}_{\mathrm{GRPO}}\left(q; w_{q,t}^{\mathrm{util}} A_q\right) \tag{48}$$

Therefore, the unified procedure can be summarized as:

$$\mathcal{I}_{q,t} \longrightarrow \begin{cases} \text{direct skill supervision}, & q \in \mathcal{H}, \\ \text{GRPO credit modulation}, & q \in \mathcal{E}. \end{cases} \tag{49}$$

while tasks in the intermediate regime are optimized by standard GRPO without additional skill-based correction.

*Unified Routed Objective.* Let $\rho_{\mathrm{int}}$, $\rho_{\mathrm{exp}}$, and $\rho_{\mathrm{util}}$ denote the curriculum proportions allocated to skill internalization, autonomous exploration, and skill utilization, respectively:

$$\rho_{\mathrm{int}} + \rho_{\mathrm{exp}} + \rho_{\mathrm{util}} = 1. \tag{50}$$

Rather than routing individual tasks to different objectives, AUSO progressively changes the contribution of skill information along this training curriculum. The complete objective is written as

$$\mathcal{L}_{\mathrm{AUSO}}(\theta; s) = \mathcal{L}_{\mathrm{GRPO}}\left(\theta; A_{q,t}^{\mathrm{GRPO}}\left[1 + \beta(s) g(p_q) m_{q,t}\right]\right) + \lambda_0 \alpha(s) \sum_q \mathbb{I}[p_q = 0] \mathcal{L}_{\mathrm{JSD}}^{(q)}, \tag{51}$$

where $\alpha(s)$ progressively decreases the contribution of teacher-guided skill internalization, while $\beta(s)$ progressively activates action-level skill utilization. The curriculum proportions determine when

these transitions occur, rather than routing individual tasks into discrete optimization regimes.

Thus, GRPO remains the persistent optimization backbone throughout training, while the role of skill information evolves progressively from external skill internalization, through autonomous exploration, to action-level information-gain-guided skill utilization.

*Smooth reduction to the common backbone.* The utilization branch continuously reduces to standard GRPO when the action-level evidence is uninformative. Specifically,

$$w_{q,t}^{\text{util}} = 1 \tag{52}$$

whenever any of the following conditions holds:

$$p_q \in \{0, 1\}, \qquad m_{q,t} = 0, \qquad \beta(s) = 0. \tag{53}$$

Therefore, we can get the results under each of these conditions:

$$\mathcal{L}_{\text{util}}(q) = \mathcal{L}_{\text{GRPO}}(q) \tag{54}$$

Furthermore, because:

$$\lim_{\beta_{(s)} \to 0} w_{q,t}^{\text{util}} = 1, \tag{55}$$

then we can get the following results:

$$\lim_{\beta_{(s)} \to 0} \mathcal{L}_{\text{util}}(q) = \mathcal{L}_{\text{GRPO}}(q). \tag{56}$$

Thus, utilization is not an independent or competing optimization objective; it is a continuous, bounded refinement of the same GRPO objective. Similarly, when the internalization coefficient vanishes:

$$\lambda_{\text{int}}(s) = 0 \tag{57}$$

the hard auxiliary correction disappears. When both auxiliary mechanisms are inactive, we can get:

$$\mathcal{L}_{\text{unified}} \longrightarrow \mathcal{L}_{\text{GRPO}} \tag{58}$$

*Unification result.* Above establishes three forms of unification:

(1) **Signal unification:** internalization and utilization are both derived from same counterfactual action-level JSD operator;
(2) **Optimization unification:** all routing branches update a single policy under one mutually exclusive routed objective;
(3) **Backbone unification:** the proposed objective continuously reduces to standard GRPO when the auxiliary information is absent or unreliable.

Hence, the framework does not combine unrelated auxiliary losses heuristically. Instead, it instantiates a single counterfactual-information principle in two complementary directions: reducing dependence on general skills through internalization and preserving sensitivity to task-specific skills through utilization.

## A.3 Why Jensen–Shannon Divergence for Action-Level Information Gain?

*Action-level information gain based on JSD.* At each environment state $h_t$, AUSO evaluates two counterfactual executions of the same policy: one conditioned on the retrieved skill and one without skill conditioning. Let $C \in \{0, 1\}$ denote the skill context, where $C = 1$ represents skill-conditioned execution and $C = 0$ represents skill-free execution. Because the language-model action is generated autoregressively, we define the action distribution at the $\ell$-th response-token position as:

$$P_{t,\ell}(a) = \pi_\theta(a \mid h_{t,\ell}, C = 1), \qquad Q_{t,\ell}(a) = \pi_\theta(a \mid h_{t,\ell}, C = 0). \tag{59}$$

Here, $h_{t,\ell}$ denotes the token-level history within environment action $u_t$. In this paper, we assign a uniform prior over the two contexts:

$$p(C = 0) = p(C = 1) = \frac{1}{2} \tag{60}$$

The corresponding marginal token distribution is

$$M_{t,\ell}(a) = p(a \mid h_{t,\ell}) = \frac{1}{2} P_{t,\ell}(a) + \frac{1}{2} Q_{t,\ell}(a). \tag{61}$$

The conditional mutual information between the skill context and the token-level action is:

$$\begin{aligned} I(C; A_{t,\ell} \mid h_{t,\ell}) &= \sum_{c \in \{0,1\}} p(c) D_{\text{KL}} \left( p(A_{t,\ell} \mid h_{t,\ell}, C = c) \,\middle\|\, p(A_{t,\ell} \mid h_{t,\ell}) \right) \\ &= \frac{1}{2} D_{\text{KL}} \left( P_{t,\ell} \,\middle\|\, M_{t,\ell} \right) + \frac{1}{2} D_{\text{KL}} \left( Q_{t,\ell} \,\middle\|\, M_{t,\ell} \right) \\ &= D_{\text{JS}} \left( P_{t,\ell} \,\middle\|\, Q_{t,\ell} \right) \end{aligned} \tag{62}$$

Therefore, we can get:

$$\boxed{D_{\text{JS}} \left( P_{t,\ell} \,\middle\|\, Q_{t,\ell} \right) = I(C; A_{t,\ell} \mid h_{t,\ell})} \tag{63}$$

This identity shows that token-level JSD measures how much information about the skill context is expressed in the local action distribution. In particular, when:

$$D_{\text{JS}} \left( P_{t,\ell} \,\middle\|\, Q_{t,\ell} \right) = 0 \iff P_{t,\ell} = Q_{t,\ell} \tag{64}$$

which means that skill conditioning does not change the token-level decision at that position.

*Symmetry.* The two distributions $P_{t,\ell}$ and $Q_{t,\ell}$ correspond to two counterfactual decisions of the same policy. They do not form a teacher–student pair with a predefined reference direction. A one-sided KL divergence is direction-dependent:

$$D_{\text{KL}}(P_{t,\ell} \| Q_{t,\ell}) \neq D_{\text{KL}}(Q_{t,\ell} \| P_{t,\ell}) \tag{65}$$

in general. Choosing one of these two directions would therefore introduce an arbitrary asymmetry. In contrast, JSD satisfies

$$D_{\text{JS}}(P_{t,\ell} \| Q_{t,\ell}) = D_{\text{JS}}(Q_{t,\ell} \| P_{t,\ell}), \tag{66}$$

more suitable for comparing two counterfactual policy decisions.

*Boundedness.* Because JSD is equal to the mutual information between a binary context variable and the token action, it is bounded by the entropy of the context:

$$0 \leq D_{\text{JS}}(P_{t,\ell} \| Q_{t,\ell}) = I(C; A_{t,\ell} \mid h_{t,\ell}) \leq H(C) = \log 2. \tag{67}$$

The action-level average is consequently also bounded:

$$0 \leq \text{IG}(u_t) \leq \log 2. \tag{68}$$

This bounded scale is useful for stable action-level modulation. By contrast, the KL divergence satisfies

$$D_{\text{KL}}(P_{t,\ell} \| Q_{t,\ell}) \in [0, \infty] \tag{69}$$

and may become arbitrarily large when $Q_{t,\ell}$ assigns very small probability to an action favored by $P_{t,\ell}$.

*Top-k approximation used in implementation.* Computing the exact JSD over the entire vocabulary is expensive. Therefore, for each token position, AUSO constructs a coarse-grained support

$$\mathcal{S}_{t,\ell} = \mathrm{TopK}(P_{t,\ell}) \cup \mathrm{TopK}(Q_{t,\ell}) \cup \{\mathrm{tail}\}. \tag{70}$$

The first two terms contain the union of the top-$k$ tokens selected by the skill-conditioned and skill-free policies. All remaining vocabulary probability mass is grouped into one aggregate tail category. Let $\widetilde{P}_{t,\ell}$ and $\widetilde{Q}_{t,\ell}$ denote the resulting normalized distributions on $\mathcal{S}_{t,\ell}$. The implemented token-level score is then

$$\widehat{\mathrm{IG}}_{t,\ell} = D_{\mathrm{JS}}\left(\widetilde{P}_{t,\ell} \,\middle\|\, \widetilde{Q}_{t,\ell}\right), \tag{71}$$

and the implemented action-level score is

$$\widehat{\mathrm{IG}}(u_t) = \frac{1}{L_t} \sum_{\ell} m_{t,\ell} \widehat{\mathrm{IG}}_{t,\ell}. \tag{72}$$

The top-$k$ union and the aggregate tail preserve the normalization of both distributions while avoiding the construction of two full vocabulary-sized probability tensors.

*Role in AUSO.* JSD is used as the unified action-level information signal throughout both internalization and utilization, providing a consistent measure of how skill knowledge influences individual decisions. During internalization, it quantifies the discrepancy between skill-conditioned teacher behavior and the skill-free counterfactual behavior, thereby identifying actions for which external skill guidance provides stronger corrective information. During utilization, the same measure characterizes how strongly the current policy decision depends on the retrieved skill, allowing the policy to selectively emphasize skill-sensitive actions. Therefore, the same action-level signal naturally connects the transition from external skill supervision to autonomous skill utilization, while providing symmetry, boundedness, support robustness, and a direct information-theoretic interpretation.

Importantly, $\mathrm{IG}(u_t)$ measures the influence of skill conditioning on the action distribution rather than directly representing a task-success probability. A large information gain indicates that introducing the skill substantially changes the policy's decision, but does not necessarily imply that this change is beneficial. The actual usefulness of such skill-induced changes is ultimately determined by the reinforcement-learning objective and the resulting task reward. In this way, JSD identifies *where* skills exert meaningful influence, while environment feedback determines *whether* that influence contributes to successful behavior.

## B Experimental Setup

*Environments and ID/OOD protocol.* We evaluate our method on ALFWorld, WebShop, and SearchQA. Across all benchmarks, policy optimization uses only in-distribution (ID) training tasks, while out-of-distribution (OOD) tasks are reserved for validation and final evaluation. ID and OOD performance is reported separately.

*ALFWorld.* Following Skill0.5, we treat {Pick & Place, Clean & Place, Cool & Place} as ID task types and {Examine in Light, Heat & Place, Pick Two & Place} as OOD task types. Training samples are drawn only from the three ID types in the standard training split. For evaluation, the standard `eval_in_distribution` games are partitioned by task type into disjoint ID and OOD subsets, and

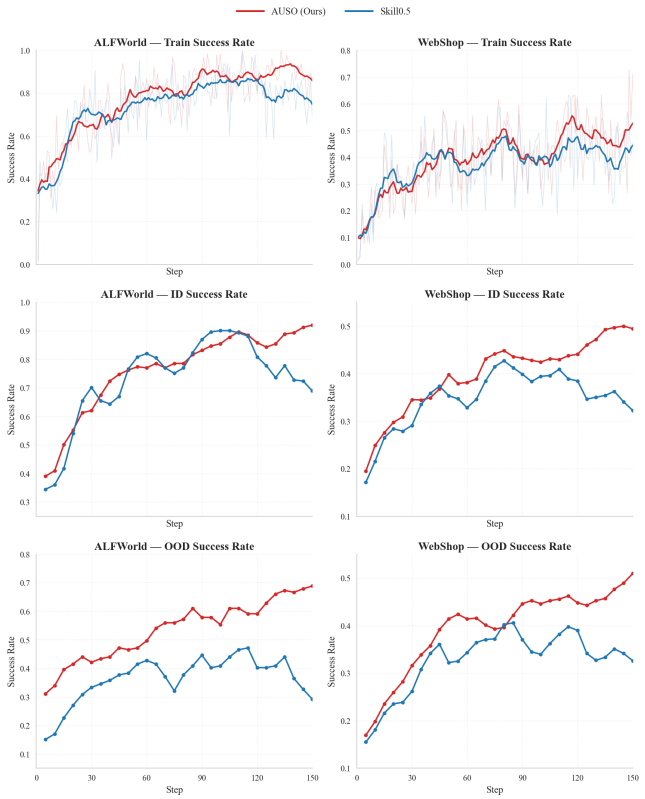


**Figure 5: ID and OOD validation success rates and training success rates of our method and Skill0.5 with task-specific skills on Alfworld and Webshop benchmarks.**

all available games are evaluated. ID tasks use the ID skill bank, whereas OOD tasks use the corresponding OOD-specific skill bank. We report per-task success rates and the macro-average within each domain. The interaction horizon is 30 actions.

*WebShop.* We partition the 12,087 human-annotated goals into seven product domains. The ID domains are {Apparel, Electronics, Footwear, Other}, and the OOD domains are {Accessories, Beauty & Health, Home Decor}. Following Skill0.5, we downsample the dominant Other category using farthest-point sampling. This produces 3,320 ID training goals, 454 ID evaluation goals, and 207 OOD evaluation goals. Training uses only ID goals from the original training range (indices 1,500 and above). The original test and development ranges (indices 0–1,499) are merged and then separated into ID and OOD evaluation subsets. OOD goals never contribute to policy updates. We report exact task success, per-category success rates, and ID/OOD averages. The interaction horizon is 15 actions.

*SearchQA..* The policy is trained only on the training splits of NQ and HotpotQA, which form the ID group. Evaluation covers seven datasets: NQ and HotpotQA are ID, while TriviaQA, PopQA, 2WikiMultiHopQA, MuSiQue, and Bamboogle are OOD. The five OOD datasets are not used for policy optimization. We construct a deterministic and domain-balanced validation set with random seed 42 and select the checkpoint at the last step. The selected


**Figure 6: Training curves of AUSO vs. GRPO on ALFWorld: AUSO achieves higher success rate and reward with lower policy entropy and shorter episode length, indicating more confident and sample-efficient learning.**

checkpoint is then evaluated on the full test collection. All datasets query the same Wikipedia corpus and complete SearchQA skill library; ID/OOD labels are used only to identify training exposure and aggregate metrics. We report per-dataset accuracy and ID/OOD averages. Each episode may issue at most four search actions, with the top three passages returned per query.

*Implementation details.* All experiments initialize from Qwen2.5-7B-Instruct and use GRPO with learning rate $1\times10^{-6}$, group size $G = 8$, and one policy epoch per update. We train each benchmark for 150 optimization steps and validate every five steps. The task batch sizes are 16, 16, and 128 for ALFWorld, WebShop, and SearchQA, respectively. Skill banks remain fixed during training, allowing us to evaluate policy learning and OOD skill utilization without online changes to the external skill corpus.

# C Detail Experimental Results

## C.1 Training Dynamics of AUSO on ALFWorld

As shown in Fig. 5, we first compare the training dynamics of AUSO with Skill0.5 on ALFWorld and WebShop. To verify that the superior performance of AUSO on ALFWorld is not simply attributable to a longer training process, we extend Skill0.5 to the same training budget of 150 steps for a controlled comparison. Under this identical optimization budget, AUSO still achieves stronger performance, particularly on the ID and OOD evaluations, demonstrating that simply increasing the number of training steps is insufficient to match the gains brought by AUSO. As training proceeds, AUSO maintains or further improves its performance, whereas Skill0.5 tends to plateau or even degrade in the later stages despite continued optimization. This trend is especially evident under the more challenging OOD settings of both ALFWorld and WebShop, where AUSO exhibits a more stable improvement and the performance gap gradually becomes more pronounced. he consistent advantage on ID tasks indicates that the improvement in generalization does not come at the expense of skill-covered scenarios. These observations rule out additional training steps as the primary source of improvement and instead demonstrate that AUSO learns more effective and robust skill-conditioned behaviors under same training budget.

We further investigate the optimization behavior of AUSO against standard GRPO in Fig. 6. AUSO achieves a consistently higher training success rate and mean reward throughout most of the optimization process, indicating that successful task behaviors are acquired earlier and reinforced more effectively. Meanwhile, AUSO exhibits lower policy entropy and progressively shorter episode lengths than GRPO, particularly during the middle and later stages of training. The reduced policy entropy suggests that the policy gradually forms more confident action distributions instead of remaining dependent on uncertain exploration, while the shorter trajectories indicate that AUSO learns to complete tasks with fewer redundant or ineffective interactions. Importantly, these improvements occur consistently across multiple optimization metrics rather than only in the final success rate. Taken together, the two analyses provide complementary evidence that the gains of AUSO originate from its more effective action-level skill optimization rather than a larger training budget, enabling the agent to acquire successful behaviors more efficiently, utilize skill information more effectively, and generalize more robustly to unseen task scenarios.

## C.2 Case Study: AUSO is More Sample-Efficient than Skill0.5

**AUSO — Same Query [SUCCESS, 2 steps]**

**Query:** *Frank Lamson-Scribner was adopted by a family near which town in Kennebec County?*

**Step 0:** `search: Frank Lamson-Scribner adopted family Kennebec County town`

**Success:** Answer-bearing passage retrieved

**Observation:** "was adopted by the Virgil Scribner family near Manchester, Maine"

**Step 1:** `answer: Manchester`

**Success:** Minimal answer span; terminate

**Skill0.5 — Same Query [FAIL, 4 steps]**

**Step 0:** Retrieve the same answer-bearing passage.

**Success:** Sufficient evidence already available

**Step 1:** `search: manchester maine adopted by virgil scribner family`

**Failure:** Redundant search

**Step 2:** `search: manchester maine adopted by virgil scribner family near`

**Failure:** Irrelevant retrieval

**Step 3:** `answer: Manchester, Maine`

**Failure:** Over-answering under exact match

*SearchQA (Multi-hop).* For the same multi-hop SearchQA question, both AUSO and Skill0.5 successfully retrieve the answer-bearing passage in the first search step, indicating that their performance difference does not originate from initial retrieval quality. Given the sufficient evidence that Frank Lamson-Scribner "was adopted by the Virgil Scribner family near Manchester, Maine,"

**Table 5: Representative cases comparing AUSO and Skill0.5.**

| Domain | Method | Query / Task | #Steps | Correct | Final output / action |
|---|---|---|---|---|---|
| SearchQA (multi-hop) | AUSO (Ours) | *Frank Lamson-Scribner was adopted by a family near which town in Kennebec County?* | 2 | ✓ | `Manchester` |
| | Skill0.5 | | 4 | × | `Manchester, Maine` |
| ALFWorld (OOD) | AUSO (Ours) | *put a spraybottle in garbagecan* | 4 | ✓ | `put spraybottle 2 in/on garbagecan 1` |
| | Skill0.5 | *put a cool lettuce in countertop* | 25 | × | `go to diningtable 1` |

AUSO immediately identifies the relevant information, extracts the minimal answer span `Manchester`, and terminates the interaction after only two steps. In contrast, Skill0.5 fails to recognize that the retrieved evidence is already sufficient and continues issuing redundant follow-up queries about the adoption location. These unnecessary searches not only increase the interaction cost but also introduce additional and less relevant information into the reasoning context, eventually leading Skill0.5 to return the over-specified answer `Manchester, Maine`, which fails under exact-match evaluation. This comparison highlights that successful SearchQA reasoning depends not only on retrieving relevant evidence, but also on determining when the available evidence is sufficient and how it should be converted into a precise final answer. By making more effective action-level decisions, AUSO avoids unnecessary retrieval, terminates at an appropriate point, and produces a concise answer consistent with the required answer granularity. This case therefore illustrates how AUSO can translate retrieved evidence into more efficient and precise search-and-answer trajectories than Skill0.5.

**AUSO — "Put a spraybottle in garbagecan" [SUCCESS, 4 steps]**

**Step 0:** `go to countertop 1`
**Success:** Locate target object
**Step 1:** `take spraybottle 2`
**Success:** Acquire object
**Step 2:** `go to garbagecan 1`
**Success:** Navigate to target receptacle
**Step 3:** `put spraybottle 2 in/on garbagecan 1`
**Success:** Goal completed

**Skill0.5 — "Put a cool lettuce in countertop" [FAIL, 25 steps]**

**Trajectory summary:**
`open fridge` → `inspect countertop` → `inspect cabinet` → `revisit locations` → `go to diningtable 1`
**Final observation:** `Nothing happens`
**Failure:** Target lettuce still not found
**Failure:** Repeated exploration of already-checked locations.
**Failure:** Step budget exhausted.

*ALFWorld (OOD).* A similar advantage is observed in the OOD ALFWorld case reported in Table 5. AUSO successfully completes the task *"put a spraybottle in garbagecan"* within only four interaction steps. Specifically, it first navigates to the countertop to locate the target object, directly acquires the spraybottle, moves to the target garbagecan, and executes the required placement action. The entire trajectory remains compact and goal-directed, with each action making explicit progress toward task completion. In contrast, Skill0.5 fails to complete its OOD task even after exhausting the 25-step interaction budget. Its trajectory repeatedly explores previously inspected locations, including the fridge, countertop, and cabinet, without successfully locating the target lettuce or establishing a reliable direction for subsequent actions. Eventually, the agent moves to an unrelated location and terminates without satisfying the task objective. This substantial difference in trajectory efficiency suggests that AUSO is better able to translate its learned skills into effective action-level decisions in unseen scenarios, reducing redundant exploration and maintaining consistent progress toward the goal. Together with the SearchQA case, these examples demonstrate that AUSO improves not only final task success but also the efficiency and precision of intermediate decision making across different interactive environments.

Overall, these cases illustrate that AUSO not only improves task success but also produces more economical and goal-directed interaction trajectories. On SearchQA, it avoids unnecessary retrieval once sufficient evidence has been obtained and terminates with a concise answer, whereas redundant follow-up searches can introduce irrelevant information. On ALFWorld, AUSO follows a compact locate–acquire–navigate–place sequence instead of repeatedly revisiting ineffective states. These observations suggest that AUSO improves not only final task performance but also the efficiency and precision of intermediate decision making.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

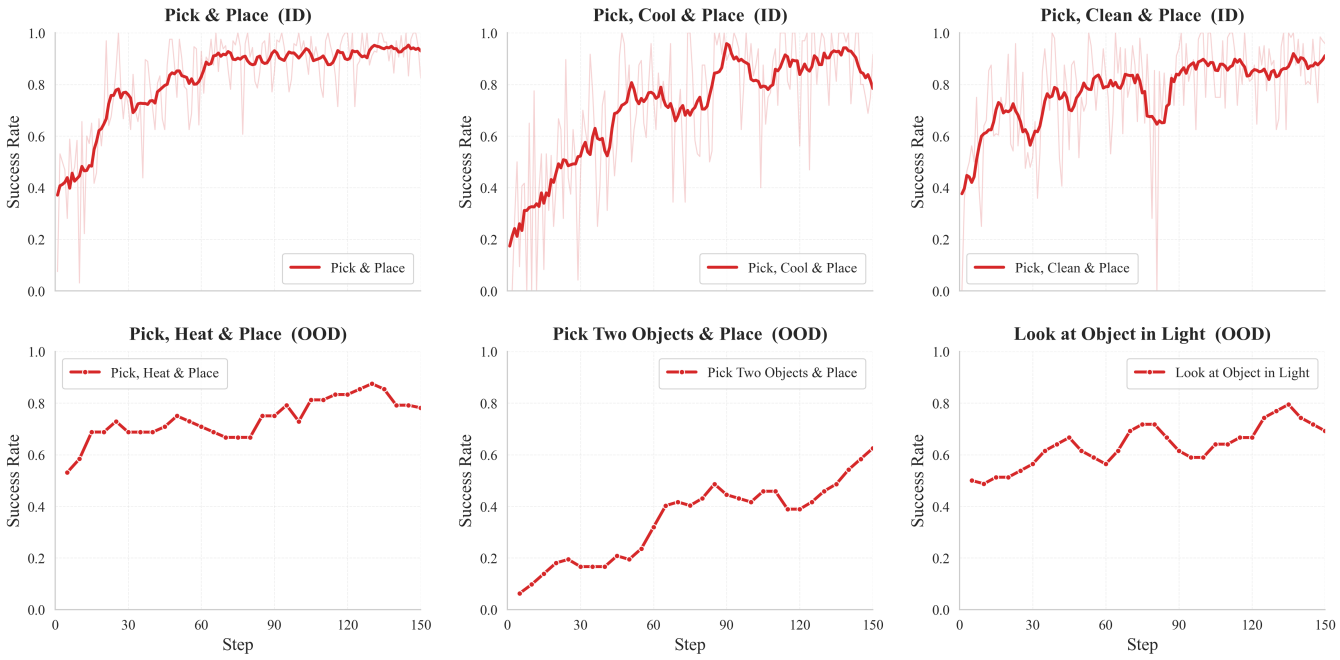


**Figure 7: Success rate of each action on the ALFWorld benchmark for AUSO (ID for training part and OOD for validation)**

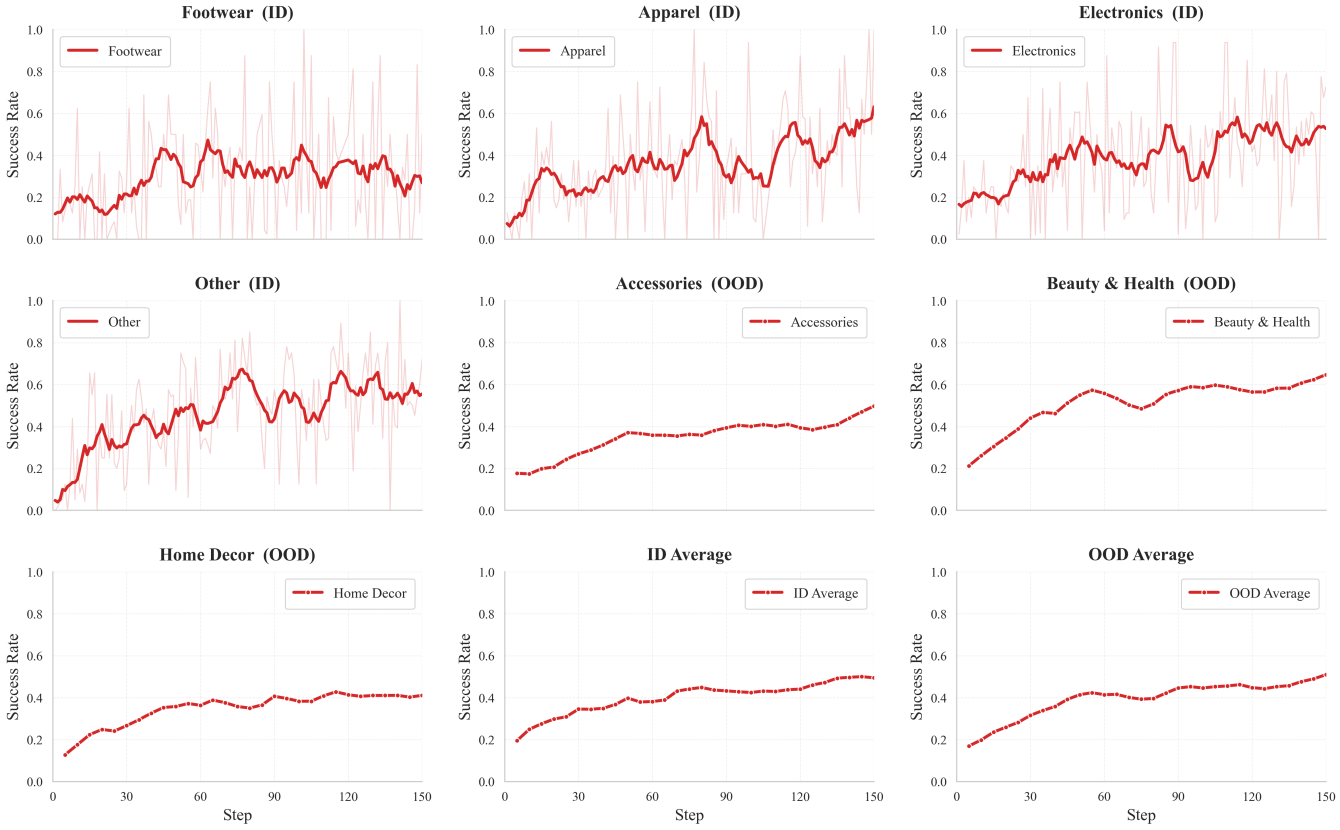


**Figure 8: Success rate of each action on the Webshop benchmark for AUSO (ID for training part and OOD for validation)**

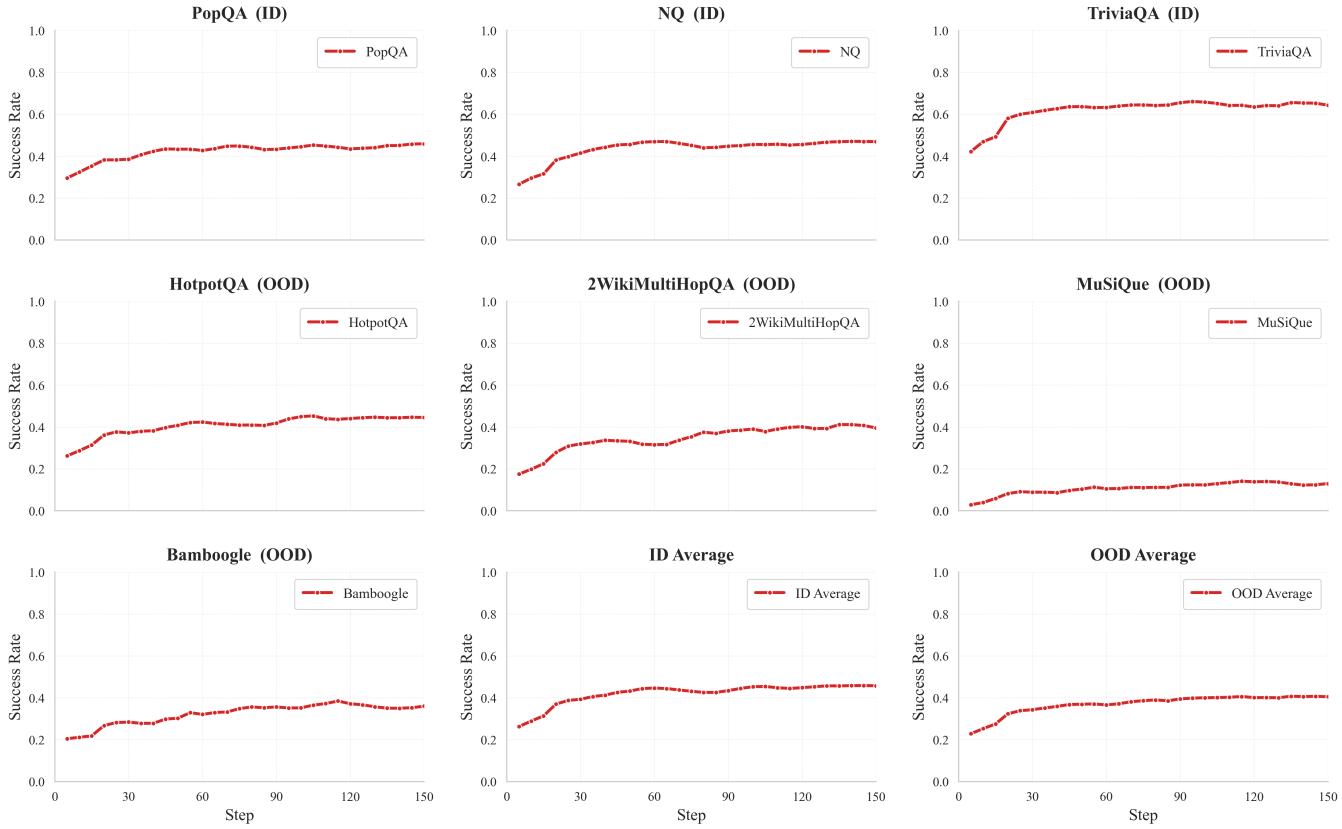


**Figure 9: Success Rate of each Action in validation set in SearchQA benchmark for our AUSO method**