Hierarchical Self-Distillation for Agentic Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

On-policy self-distillation (OPSD) has emerged as a promising post-training approach that complements sparse outcome rewards in agentic reinforcement learning with dense guidance from a privileged self-teacher. Nevertheless, we reveal that the utility of teacher supervision in multi-turn OPSD is hierarchically non-uniform: step-level supervision utility is influenced by both accumulated state drift and future teacher–student divergence, while at the token level, under a limited support budget, a fixed-size support cannot maintain consistent fidelity to full-distribution supervision as teacher–student discrepancy varies. To address this, we propose **HiSD**, a hierarchical self-distillation framework that combines progress-aware step supervision with support-adaptive token supervision, allocating step-level distillation according to relative divergence progress and adapting the distributional scope of token-level supervision based on teacher–student discrepancy. Across ALFWorld, Search-QA, and WebShop with Qwen2.5-Instruct models, HiSD consistently outperforms GRPO and recent OPSD-based baselines. Notably, on tasks with strong dependencies across interaction steps, HiSD outperforms SDAR by an average of 16.0 points on Pick2 and 2.4 points on multi-hop QA, while achieving 159% higher teacher budget utilization on WebShop. Code available: https://anonymous.4open.science/r/HiSD.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has become a prevailing paradigm for post-training language models and agents, but its policy updates still rely primarily on sparse outcome rewards from the environment (Guo et al., 2025; Yu et al., 2026a; Jin et al., 2025; Feng et al., 2026). On-policy distillation (OPD) complements these rewards with dense token-level guidance from a stronger teacher on trajectories sampled by the student, with both policies evaluated under the same student-induced prefixes (Agarwal et al., 2024; Jin et al., 2026). On-policy self-distillation (OPSD) further constructs the teacher from the same policy by conditioning it on privileged information unavailable to the student, replacing the parametric advantage of a stronger external teacher with an information advantage that allows the student to learn from privileged guidance grounded in its own on-policy experience (Zhao et al., 2026; He et al., 2026).

Despite these advantages, extending OPSD from single-turn reasoning tasks such as mathematical reasoning and code generation to multi-turn agentic settings introduces a distinct challenge (Zhong et al., 2026; Wu et al., 2026; Lu et al., 2026). In reasoning tasks, even as the student gradually deviates from the teacher distribution, subsequent predictions unfold within a continuously extending autoregressive context, allowing teacher supervision to correct the deviation within only a few subsequent tokens (Jiang & Ferraro, 2026; Zhang et al., 2026). In multi-turn agent interactions, by contrast, an erroneous action changes the environment state and continues to influence later decisions, causing teacher–student divergence to accumulate across interaction steps, as shown in Figure 1(a). Beyond this cumulative state drift, Figure 1(b) further shows that teacher intervention in higher-divergence windows yields better subsequent student alignment, with the advantage becoming more pronounced as the takeover window expands. This indicates that step-level supervision utility is influenced not only by accumulated historical drift but also by future discrepancy.

Expanding each agent step along its token-generation dimension, several recent methods instantiate token-level distillation using only the sampled token, providing a computationally efficient approx-

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.

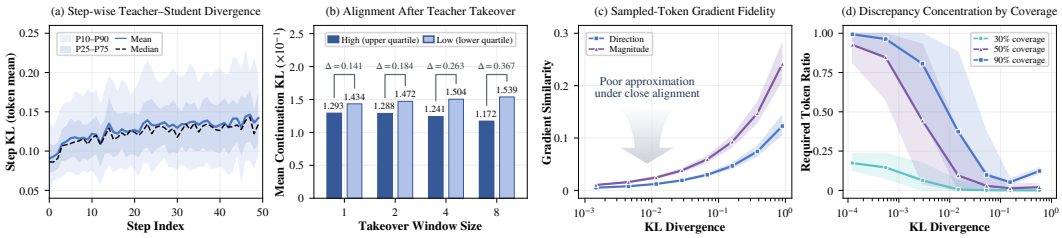


Figure 1: The motivation of HiSD. (a) Teacher–student divergence grows across interaction steps as state drift accumulates. (b) Lower continuation KL indicates greater potential utility of teacher intervention in future interaction windows, with High and Low referring to the corresponding takeover-window KL. (c) Under close alignment, sampled-token gradients poorly approximate the direction and magnitude of the full-distribution KL gradient. (d) The required ratio of teacher-ranked candidates at fixed full-distribution gradient coverage levels (30%, 50%, and 90%) generally decreases as KL divergence grows. Detailed experimental settings are provided in Section B.4.

imation to full-distribution supervision (Lu et al., 2026; Wang et al., 2026a). However, Figure 1(c) shows that when the policies of teacher and student are closely aligned, the sampled-token gradient exhibits weaker directional agreement and larger magnitude distortion relative to the full-distribution gradient, which is theoretically established in Appendix A.1. Given the limited fidelity of sampled-token supervision, Figure 1(d) further examines how many teacher-ranked candidates are needed to cover a fixed proportion of the full-distribution gradient signal. Specifically, the gradient signal is more diffuse under low divergence, requiring more candidates for the same coverage, but becomes increasingly concentrated in the teacher's high-probability region as divergence grows. These observations indicate that the required support size is closely tied to teacher–student divergence, shrinking as the discrepancy becomes more concentrated.

The above analyses reveal that effective multi-turn on-policy distillation requires jointly calibrating the strength of teacher supervision across interaction steps and its distributional scope within token prediction. Based on these insights, we propose **HiSD**, a hierarchical self-distillation framework that couples **progress-aware step supervision** with **support-adaptive token supervision**, as shown in Figure 2. At the step level, HiSD determines the distillation weight for each decision by comparing the divergence accumulated over preceding interactions with that remaining from current decision onward. When remaining divergence dominates, HiSD assigns a large distillation weight to preserve the teacher's guidance role, whereas it progressively reduces this weight as accumulated divergence becomes dominant and the student-induced history increasingly constrains the reliability of subsequent guidance, shifting the balance from distillation toward outcome-driven GRPO. To approximate full-distribution supervision under a limited support budget, HiSD adjusts the token-level support size according to instantaneous KL divergence. Large divergence yields a compact support focused on teacher's high-probability candidates, whereas closer alignment expands support to preserve fine-grained probability relations among more candidates. By matching support scope to discrepancy concentration, HiSD efficiently approximates full-distribution supervision for each token.

We conduct extensive experiments on ALFWorld (Shridhar et al., 2020), Search-QA (Jin et al., 2025), and WebShop (Yao et al., 2022a) using Qwen2.5-Instruct models, with HiSD achieving average gains over GRPO of 10.2, 6.5, and 8.5 points, respectively, and continuing to improve in both success rate and interaction efficiency even after surpassing SDAR, a strong OPSD-based baseline. In-depth studies further validate the two key designs: step-level adaptation is particularly effective on tasks with pronounced dependencies across interaction steps, whereas token-level adaptation captures systematic variation in distributional scope across tokens with different semantic roles (e.g., reasoning-pattern and action/entity-related tokens), with their combination also stabilizing the balance between RL and OPSD throughout training.

2 Related Work

2.1 LLM-based Multi-turn Agents.

Large language models have demonstrated strong capabilities as multi-turn agents across diverse interaction settings, including API and tool calling (Schick et al., 2023; Qin et al., 2024), web

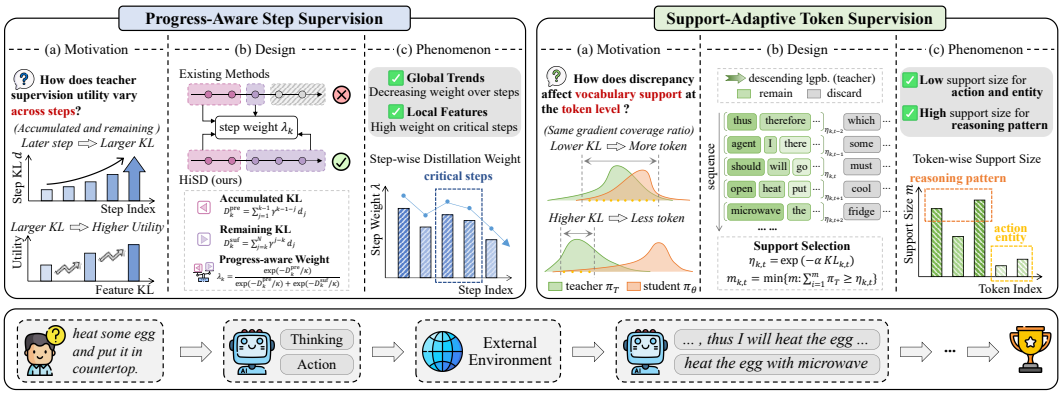


Figure 2: Overview of HiSD. Progress-aware Step Supervision (left) leverages accumulated and remaining teacher–student divergence to reduce the distillation strength of steps with substantial accumulated drift while emphasizing locally critical steps. Support-adaptive Token Supervision (right) adjusts the teacher support size according to token-level divergence, preserving fine-grained distributions under close alignment and focusing on high-probability tokens otherwise.

navigation and information searching (Nakano et al., 2021; Yao et al., 2022a), and embodied or open-ended environments (Ahn et al., 2022; Wang et al., 2023). Early work typically relied on carefully designed prompting frameworks to elicit interleaved reasoning and action generation from frozen pretrained models (Yao et al., 2022b; Shinn et al., 2023). Recent work has extended RLVR-based policy optimization from single-turn reasoning to long-horizon agent training, enabling agents to learn directly through interactions with the environment, although such training remains sample-inefficient in sparse-reward settings (Wang et al., 2025; Feng et al., 2026).

2.2 Reliable On-Policy (Self-)Distillation.

Although on-policy (self-)distillation provides dense teacher supervision on student-generated sequences while avoiding the distribution mismatch induced by offline distillation (Gu et al., 2024), the resulting supervision is not uniformly reliable (Fu et al., 2026; Li et al., 2026b). Some methods derive distillation signals from the sampled token alone (Zhao et al., 2026; Lu et al., 2026), others from distributions over the full vocabulary or a fixed support (Agarwal et al., 2024; Fu et al., 2026), while TIDE (Yu et al., 2026b) switches between these supervision forms. However, these methods are designed for single-turn reasoning and overlook divergence accumulation induced by environment state transitions. Existing agentic OP(S)D methods regulate multi-turn supervision through global training schedules (Wang et al., 2026b; Li et al., 2026a; Tan et al., 2026) or reweight steps using local reliability (Zhong et al., 2026; Zhou et al., 2026). In contrast, our method jointly adapts step-level weights to relative divergence progress and token-level supports to teacher–student discrepancy.

3 Method

3.1 Problem Setup: Multi-Turn Agent with On-Policy Self-Distillation

Given a task $x \sim \mathcal{D}$, an agent interacts with the environment $\mathcal{E}$ for N steps, yielding a trajectory $\tau = (x, \{(a_k, o_k)\}_{k=1}^{N})$. Before the k-th decision, the interaction history is

$$\mathcal{H}_k = (x, a_1, o_1, \ldots, a_{k-1}, o_{k-1}),$$

where a_j and o_j denote the agent action and the corresponding environment observation at step j, respectively. Given $\mathcal{H}_k$, the policy π_θ generates the k-th action as

$$a_k = (y_{k,1}, \ldots, y_{k,T_k}) \sim \pi_\theta(\cdot \mid \mathcal{H}_k),$$

where T_k denotes the number of tokens in a_k. Specifically, the t-th token $y_{k,t}$ is autoregressively generated conditioned on $c_{k,t} = (\mathcal{H}_k, y_{k,<t})$:

$$y_{k,t} \sim \pi_\theta(\cdot \mid c_{k,t}).$$

Group Relative Policy Optimization For each task x, GRPO samples a group of G trajectories $\{\tau_i\}_{i=1}^{G}$ from the policy $\pi_{\theta_{\text{old}}}$ and computes the trajectory-level advantage $\hat{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G})}$ based on the environment reward R. With $\mathcal{T}_i$ denoting the set of all agent-generated token positions in trajectory τ_i, GRPO loss is defined as

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathcal{T}_i|}\sum_{(k,t)\in\mathcal{T}_i}\left(\min\left(\rho_{i,k,t}(\theta)\hat{A}_i, \text{clip}\left(\rho_{i,k,t}(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i\right)\right)\right],$$

where $\rho_{i,k,t}(\theta) = \frac{\pi_\theta(y_{i,k,t}|c_{i,k,t})}{\pi_{\theta_{\text{old}}}(y_{i,k,t}|c_{i,k,t})}$ denotes token-level importance ratio and ϵ is the clipping range.

On-Policy Self-Distillation In OPSD, the teacher augments each context $c_{k,t}$ generated by the student policy π_θ with privileged information r, instantiated as task-relevant skills in our method:

$$\pi_T\left(\cdot \mid c_{k,t}, r\right) \triangleq \text{sg}\left[\pi_\theta\left(\cdot \mid c_{k,t}, r\right)\right],$$

where $\text{sg}[\cdot]$ denotes stop-gradient.

To provide dense supervision at tractable cost, we construct a candidate support $\mathcal{C}_{k,t}$ of fixed size K, rather than the full vocabulary, by retaining the highest-probability tokens from the teacher, whose privileged distribution serves as the guidance for the student. The teacher and student distributions are then renormalized over Top-K support as follows:

$$\bar{\pi}_T\left(v \mid c_{k,t}, r\right) = \frac{\pi_T(v \mid c_{k,t}, r)}{\sum_{u\in\mathcal{C}_{k,t}}\pi_T(u \mid c_{k,t}, r)},\ \bar{\pi}_\theta\left(v \mid c_{k,t}\right) = \frac{\pi_\theta(v \mid c_{k,t})}{\sum_{u\in\mathcal{C}_{k,t}}\pi_\theta(u \mid c_{k,t})},\ v \in \mathcal{C}_{k,t}. \quad (1)$$

3.2 Progress-Aware Step Supervision

In multi-turn interaction, teacher–student divergence evolves along the trajectory through successive decisions and environment feedback, resulting in different degrees of accumulated drift across steps.

For the k-th interaction step, we measure the instantaneous teacher–student divergence at token (k,t) using reverse KL over the fixed candidate support $\mathcal{C}_{k,t}$:

$$\delta_{k,t}^{\leftarrow} = \text{sg}\left[D_{\text{KL}}\left(\bar{\pi}_\theta(\cdot \mid c_{k,t}) \,\|\, \bar{\pi}_T(\cdot \mid c_{k,t}, r)\right)\right].$$

We first define the step-level divergence d_k as the average token-level divergence within step k. Based on d_k, we separately aggregate the divergence accumulated before step k, denoted by D_k^{pre}, and the divergence remaining from the current step onward, denoted by D_k^{suf}:

$$d_k = \frac{1}{T_k}\sum_{t=1}^{T_k}\delta_{k,t}^{\leftarrow}, \qquad D_k^{\text{pre}} = \sum_{j=1}^{k-1}\gamma^{k-1-j}d_j, \qquad D_k^{\text{suf}} = \sum_{j=k}^{N}\gamma^{j-k}d_j,$$

where $\gamma \in (0,1]$ is a temporal decay factor that emphasizes nearby steps.

We define the distillation weight of step k by normalizing its prefix divergence against the decayed divergence accumulated on both sides of the current decision:

$$\lambda_k = \frac{\exp(-D_k^{\text{pre}}/\kappa)}{\exp(-D_k^{\text{pre}}/\kappa) + \exp\left(-D_k^{\text{suf}}/\kappa\right)} \quad (2)$$

where $\kappa > 0$ controls the sensitivity of the weighting to divergence progression. HiSD thereby allocates teacher supervision according to each decision's position within the trajectory-level divergence progression, rather than relying solely on local divergence or absolute step index.

3.3 Support-Adaptive Token Supervision

Within each interaction step, applying a candidate support $\mathcal{C}_{k,t}$ of fixed size K for each token may not provide an appropriate granularity for token-level distillation. We therefore formulate adaptive support selection within $\mathcal{C}_{k,t}$ as a **divergence-constrained compression** of the candidate-normalized

teacher distribution. Specifically, we seek to minimize the supervision support while constraining the distortion induced by truncation within a budget determined by current teacher–student divergence:

$$\min_{\mathcal{S}\subseteq\mathcal{C}_{k,t}} |\mathcal{S}| \qquad \text{s.t.} \qquad D_{\mathrm{KL}}\left(\widetilde{\pi}_T \,\|\, \bar{\pi}_T\right) \le \alpha\, \delta^{\leftarrow}_{k,t}. \tag{3}$$

Here, for any $\mathcal{S} \subseteq \mathcal{C}_{k,t}$, $\widetilde{\pi}_T$ and $\widetilde{\pi}_\theta$ denote the candidate-normalized teacher and student distributions further truncated and renormalized over $\mathcal{S}$:

$$\widetilde{\pi}_T\left(v \mid c_{k,t}, r\right) = \frac{\bar{\pi}_T(v \mid c_{k,t}, r)}{\sum_{u\in\mathcal{S}} \bar{\pi}_T(u \mid c_{k,t}, r)}, \qquad \widetilde{\pi}_\theta\left(v \mid c_{k,t}\right) = \frac{\bar{\pi}_\theta(v \mid c_{k,t})}{\sum_{u\in\mathcal{S}} \bar{\pi}_\theta(u \mid c_{k,t})}, \quad v \in \mathcal{S}. \tag{4}$$

The coefficient $\alpha > 0$ controls the allowable distortion: a larger α relaxes the distortion budget and permits stronger support compression, whereas a smaller α imposes a tighter constraint.

The left-hand side of Eq. (3) measures the distortion induced when the teacher distribution over $\mathcal{C}_{k,t}$ is approximated by its renormalized restriction to $\mathcal{S}$. Substituting Eq. (4) into this term yields

$$D_{\mathrm{KL}}\left(\widetilde{\pi}_T \,\|\, \bar{\pi}_T\right) = -\log \sum_{v\in\mathcal{S}} \bar{\pi}_T(v \mid c_{k,t}, r).$$

Therefore, by eliminating $\widetilde{\pi}_T$, the constraint in Eq. (3) can be equivalently rewritten as an inequality on the probability mass assigned to $\mathcal{S}$ under $\bar{\pi}_T$:

$$\sum_{v\in\mathcal{S}} \bar{\pi}_T(v \mid c_{k,t}, r) \ge \exp(-\alpha\, \delta^{\leftarrow}_{k,t}).$$

We further define $\eta_{k,t}$ as the adaptive threshold induced by the distortion constraint, which specifies the minimum teacher probability mass that must be retained for distillation:

$$\eta_{k,t} = \exp(-\alpha\, \delta^{\leftarrow}_{k,t}).$$

To satisfy this threshold with minimum support, we sort the tokens in $\mathcal{C}_{k,t}$ in descending order of their candidate-normalized teacher probabilities $\bar{\pi}_T$:

$$\bar{\pi}_T\left(v^{(1)}_{k,t} \mid c_{k,t}, r\right) \ge \bar{\pi}_T\left(v^{(2)}_{k,t} \mid c_{k,t}, r\right) \ge \cdots \ge \bar{\pi}_T\left(v^{(K)}_{k,t} \mid c_{k,t}, r\right),$$

where $v^{(i)}_{k,t}$ denotes the token with the i-th largest teacher probability within $\mathcal{C}_{k,t}$. The support size is then determined by the number of ranked tokens required for their cumulative teacher probability to reach $\eta_{k,t}$:

$$m_{k,t} = \min\left\{ m : \sum_{i=1}^{m} \bar{\pi}_T\left(v^{(i)}_{k,t} \mid c_{k,t}, r\right) \ge \eta_{k,t} \right\}. \tag{5}$$

The resulting adaptive support is $\mathcal{S}_{k,t} = \{v^{(1)}_{k,t}, \ldots, v^{(m_{k,t})}_{k,t}\}$. We constrain $|\mathcal{S}_{k,t}| \ge 2$ to retain nontrivial relative-probability supervision. Instantiating Eq. 4 with $\mathcal{S} = \mathcal{S}_{k,t}$ yields the truncated and renormalized teacher and student distributions $\widetilde{\pi}_T$ and $\widetilde{\pi}_\theta$ for distillation. Thus, the optimization problem in Eq. (3) is equivalently reformulated as selecting candidates from $\mathcal{C}_{k,t}$ in descending order of teacher probability until their cumulative probability reaches $\eta_{k,t}$. As teacher–student divergence grows, $\eta_{k,t}$ decreases, permitting stronger compression within $\mathcal{C}_{k,t}$, whereas closer alignment raises the threshold and expands supervision toward full candidate support.

3.4 Unified Hierarchical Distillation Objective

To facilitate learning from the guidance conveyed by the teacher-selected candidates, we perform token-level self-distillation utilizing forward KL rather than reverse KL between the renormalized teacher and student distributions over the adaptive support $\mathcal{S}_{k,t}$:

$$\widetilde{\delta}^{\rightarrow}_{k,t} = D_{\mathrm{KL}}\left(\widetilde{\pi}_T(\cdot \mid c_{k,t}, r) \,\|\, \widetilde{\pi}_\theta(\cdot \mid c_{k,t})\right).$$

The hierarchical self-distillation objective integrates the support-adaptive token-level supervision with the progress-aware step weighting:

$$\mathcal{L}_{\mathrm{HiSD}} = \mathbb{E}\left[\frac{1}{N} \sum_{k=1}^{N} \lambda_k \left(\frac{1}{T_k} \sum_{t=1}^{T_k} \widetilde{\delta}^{\rightarrow}_{k,t} \right) \right].$$

The overall objective combines GRPO with hierarchical self-distillation, with λ_{base} controlling their global trade-off:

$$\mathcal{L} = \mathcal{L}_{\mathrm{GRPO}} + \lambda_{\mathrm{base}} \mathcal{L}_{\mathrm{HiSD}}.$$

Table 1: Main results on ALFWorld, Search-QA, and WebShop. The best and second-best results within each backbone are highlighted in **bold** and underlined, respectively. Results marked with † are reproduced by us; the others are reported in GiGPO (Feng et al., 2026) or SDAR (Lu et al., 2026).

Method	ALFWorld							Search-QA								WebShop	
	Pick	Look	Clean	Heat	Cool	Pick2	Avg	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam	Avg	Score	Acc
Closed-Source Model																	
GPT-4o	75.3	60.8	31.2	56.7	21.6	49.8	48.0	–	–	–	–	–	–	–	–	31.8	23.7
Gemini-2.5-Pro	92.8	63.3	62.1	69.0	26.6	58.7	60.3	–	–	–	–	–	–	–	–	42.5	35.9
Qwen2.5-1.5B-Instruct																	
Vanilla†	17.2	0.0	12.9	5.3	0.0	5.0	8.6	15.8	31.9	15.3	15.1	22.4	2.5	4.8	15.4	7.9	1.5
ReAct†	17.4	20.5	15.7	6.2	7.7	2.0	12.8	17.1	34.7	16.5	16.4	24.4	2.6	4.4	16.6	20.6	3.9
Reflexion†	35.3	22.2	21.7	13.6	19.4	3.7	21.8	19.9	35.2	20.5	18.3	25.6	3.2	4.0	18.1	26.4	6.3
GRPO†	80.0	72.2	76.0	72.7	54.5	72.7	71.9	23.4	41.3	22.3	21.9	34.1	3.2	**11.3**	22.5	74.4	61.7
Skill-GRPO†	59.7	50.0	51.5	64.3	35.9	57.8	53.1	21.4	43.0	20.5	19.7	31.6	3.5	5.2	20.7	69.5	57.0
SSPO†	83.3	41.7	85.7	89.5	**79.2**	50.0	74.2	23.3	**45.5**	21.4	21.3	**36.8**	3.5	4.4	22.3	78.8	64.8
RLSD†	76.5	50.0	**92.3**	61.9	75.0	**73.7**	75.0	23.2	42.7	21.3	20.9	31.6	3.7	8.5	21.7	80.4	62.9
Skill-SD†	88.2	50.0	**92.3**	76.2	70.0	52.6	76.6	25.0	42.2	23.2	24.1	30.3	3.3	7.3	22.2	79.6	63.7
SDAR†	92.0	69.0	56.3	50.0	33.8	50.0	63.3	23.8	44.9	23.0	20.0	33.0	2.6	3.2	21.5	77.8	62.9
HiSD (Ours)	**92.9**	**88.9**	78.5	**91.7**	72.7	66.8	**80.5**	**25.4**	42.2	**23.6**	**26.2**	34.8	**4.3**	5.2	**23.1**	**86.4**	**70.7**
Qwen2.5-3B-Instruct																	
Vanilla	44.4	11.1	6.2	15.4	28.6	12.5	21.9	24.6	48.1	31.0	26.3	25.3	7.2	59.7	31.7	6.7	0.8
ReAct†	16.1	11.8	17.9	13.3	15.0	23.5	16.4	24.1	47.5	30.4	25.3	24.9	6.8	60.1	31.3	5.9	0.0
Reflexion†	30.3	27.8	30.8	28.6	26.3	22.2	28.1	24.1	43.3	34.1	27.6	25.5	10.0	64.9	32.8	13.2	1.9
GRPO	91.2	62.5	96.2	61.9	65.0	47.4	75.0	39.3	60.6	41.1	37.4	34.6	15.4	26.4	36.4	79.8	63.3
Skill-GRPO	88.9	71.4	58.8	70.6	40.7	29.2	60.2	43.5	58.8	43.0	36.8	32.2	11.7	12.5	34.1	77.3	60.9
SSPO†	90.9	68.8	72.7	58.3	**76.9**	47.4	72.7	37.6	56.6	40.1	37.6	34.8	13.8	**67.3**	41.1	78.9	62.5
RLSD	87.9	**75.0**	90.9	75.0	73.1	68.4	79.7	41.5	58.6	42.3	40.4	40.2	16.8	66.9	43.8	84.4	66.4
Skill-SD	88.2	50.0	96.2	52.4	65.0	57.9	73.4	44.4	60.4	44.0	39.5	40.4	15.4	64.9	44.1	75.9	64.0
SDAR	97.1	62.5	**100**	61.9	75.0	84.2	84.4	44.8	58.1	44.3	38.6	36.2	15.7	66.1	43.4	85.0	68.0
HiSD (Ours)	**97.6**	66.7	91.9	**79.2**	68.9	**100**	**88.3**	**47.0**	**63.7**	**47.9**	**44.7**	**43.7**	**17.2**	66.1	**47.2**	**85.7**	**71.1**
Qwen2.5-7B-Instruct																	
Vanilla	36.1	22.2	3.1	0.0	0.0	0.0	12.5	25.2	50.8	29.5	29.0	29.0	10.4	63.7	33.9	5.9	1.6
ReAct	48.5	35.4	34.3	13.2	18.2	17.6	31.2	19.1†	37.5†	10.3†	18.2†	26.7†	7.4†	16.8†	19.4†	46.2	19.5
Reflexion	62.0	41.6	44.9	30.9	36.3	23.8	42.7	25.6†	52.2†	29.6†	27.6†	24.7†	8.8†	12.1†	25.8†	58.1	28.8
GRPO	91.2	87.5	96.2	81.0	65.0	57.9	81.2	45.1	63.7	44.0	43.6	43.2	16.8	37.6	42.0	80.9	72.6
Skill-GRPO	88.5	66.7	65.2	61.1	57.7	73.1	69.5	45.2	63.7	45.7	43.1	43.3	19.6	21.4	40.3	80.4	71.9
SSPO†	85.7	54.5	88.9	83.3	69.2	60.0	75.8	48.0	63.8	47.4	44.0	41.8	14.4	63.3	46.1	84.1	74.6
RLSD	**100**	87.5	92.3	58.8	**80.0**	65.2	82.0	46.8	63.0	44.4	45.5	**48.9**	**21.5**	**73.0**	49.0	87.4	77.3
Skill-SD	93.9	**93.8**	90.9	**100**	69.2	68.4	85.1	47.1	64.5	47.8	44.2	42.1	20.2	69.0	47.8	86.1	76.5
SDAR	94.7	75.0	**100**	86.7	68.2	78.9	85.9	46.3	63.5	48.2	43.8	48.4	19.6	**73.0**	49.0	89.4	**82.8**
HiSD (Ours)	94.4	88.9	93.3	85.7	76.2	**94.4**	**89.8**	**49.3**	**65.6**	**49.6**	**47.7**	47.8	21.3	69.4	**50.1**	**90.5**	81.2

4 Experiments

4.1 Experimental Setup

Benchmarks and Implementation. We evaluate HiSD on three representative multi-turn agent settings covering text-based embodied tasks (**ALFWorld**) (Shridhar et al., 2020), simulated online shopping (**WebShop**) (Yao et al., 2022a), and search-augmented question answering. For ALFWorld and WebShop, we set the maximum environment steps per episode to 50 and 15, with a maximum prompt length of 2048 and 4096, respectively. For **Search-QA**, we follow the Search-R1 (Jin et al., 2025) setup, including single-hop datasets, NQ, TriviaQA, and PopQA, and multi-hop datasets, HotpotQA, 2Wiki, MuSiQue, and Bamboogle, with NQ and HotpotQA in-domain and the rest held out. The agent retrieves evidence using E5 for at most 4 steps, with a maximum prompt length of 4096. We adopt the SDAR (Lu et al., 2026) training framework with privileged information from SkillRL (Xia et al., 2026), training Qwen2.5-Instruct models for 150 steps on 8 GPUs. For HiSD, we set $\lambda_{\text{base}} = 0.01$, $\kappa = 1.0$, and $\gamma = 0.9$, and perform adaptive support selection over the teacher's Top-256 candidates with $\alpha = 3.0$.

Baselines. We compare HiSD against four categories of baselines: (1) **Closed-source LLMs**: GPT-4o (Hurst et al., 2024) and Gemini-2.5-Pro (Comanici et al., 2025), serving as strong proprietary

Table 2: Component and mechanism ablations of HiSD using Qwen2.5-1.5B-Instruct. Δ denotes the decrease from HiSD (Full); darker cells indicate larger drops within each metric.

Configuration	ALFWorld		Search-QA		WebShop			
	Succ.	Δ	Avg.	Δ	Scor.	Δ	Succ.	Δ
HiSD (Full)	**80.5**	–	**23.1**	–	**86.4**	–	**70.7**	–
Component Ablations								
GRPO+OPSD	64.1	↓ 16.4	21.6	↓ 1.5	80.4	↓ 6.0	64.5	↓ 6.2
+ w/ Step-wise Adaptation	69.5	↓ 11.0	21.8	↓ 1.3	81.3	↓ 5.1	63.7	↓ 7.0
+ w/ Token-wise Adaptation	68.0	↓ 12.5	22.3	↓ 0.8	81.9	↓ 4.5	65.2	↓ 5.5
Step-level Coefficient Strategies								
Prefix Only	77.3	↓ 3.2	22.7	↓ 0.4	81.8	↓ 4.6	67.6	↓ 3.1
Suffix Only	74.2	↓ 6.3	22.2	↓ 0.9	83.6	↓ 2.8	63.7	↓ 7.0
Relative Position	72.7	↓ 7.8	22.0	↓ 1.1	80.8	↓ 5.6	61.3	↓ 9.4
Token-level Support Strategies								
Random Selection	63.3	↓ 17.2	19.6	↓ 3.5	85.9	↓ 0.5	67.2	↓ 3.5
Top Selection	68.8	↓ 11.7	21.6	↓ 1.5	84.4	↓ 2.0	68.8	↓ 1.9
Student Candidate Support	76.6	↓ 3.9	21.3	↓ 1.8	79.4	↓ 7.0	64.8	↓ 5.9

model references. (2) **Prompt-based Agentic Methods**: ReAct (Yao et al., 2022b) interleaves reasoning and acting, while Reflexion (Shinn et al., 2023) further introduces self-reflection. (3) **RL-based Methods**: GRPO (Guo et al., 2025) and its skill-augmented variant, Skill-GRPO. (4) **OPSD-based Methods**: SSPO (Wu et al., 2026) and RLSD (Yang et al., 2026) instead use teacher–student discrepancy to reweight RL advantages at the step and token levels, respectively. Skill-SD (Wang et al., 2026a) and SDAR (Lu et al., 2026) incorporate self-distillation as an auxiliary token-level loss, with Skill-SD employing importance-weighted distillation and SDAR adaptively gating teacher signals based on the discrepancy between student and teacher.

4.2 Main Results

Overall Performance Across Model Scales. HiSD achieves strong performance even with limited model capacity, with its 1.5B variant substantially outperforming the closed-source models on both ALFWorld and WebShop. The consistent underperformance of Skill-GRPO relative to GRPO across model scales indicates that sparse reward is insufficient for effective skill learning, with direct skill augmentation even impairing policy optimization. In contrast, OPSD-based methods improve skill-augmented training by providing dense teacher supervision. However, these methods exhibit unstable performance across model scales, with token-level SDAR underperforming on small models and step-level SSPO on large models. By comparison, HiSD adaptively balances OPSD with RL and improves fine-grained distribution alignment, achieving the strong performance on three benchmarks.

Effectiveness on Long-Horizon Tasks. Beyond its strong overall performance, HiSD achieves more substantial improvements on subtasks requiring longer decision horizons. Specifically, on ALFWorld, HiSD surpasses SDAR by an average of 16.0 points on Pick2, which involves manipulating multiple objects, compared with only 0.4 points on the simpler Pick task. A similar phenomenon is observed on Search-QA: across model scales, the performance margin of HiSD over SDAR averages 1.9 points on single-hop QA but widens to 2.4 points on multi-hop QA, which requires iterative query formulation, retrieval, and evidence aggregation. These results demonstrate the advantage of hierarchical adaptation across steps and tokens over token-level adaptation alone, particularly for tasks requiring long decision horizons.

4.3 Ablation Study

Complementarity of Hierarchical Adaptation. Compared with GRPO+OPSD, Step-wise Adaptation provides significant gain on ALFWorld (+5.4), which features long-horizon interaction with continually evolving environment states, whereas Token-wise Adaptation attains larger gains on Search-QA (+0.7) and WebShop (+1.5 in score), which require query formulation and fine-grained product and option selection, respectively. Notably, Step-wise Adaptation improves Webshop Score by 0.9 but reduces success rate by 0.8, indicating that step-level supervision reallocation improves partial matching but remains insufficient for task success with a fixed (K=256) token support. By

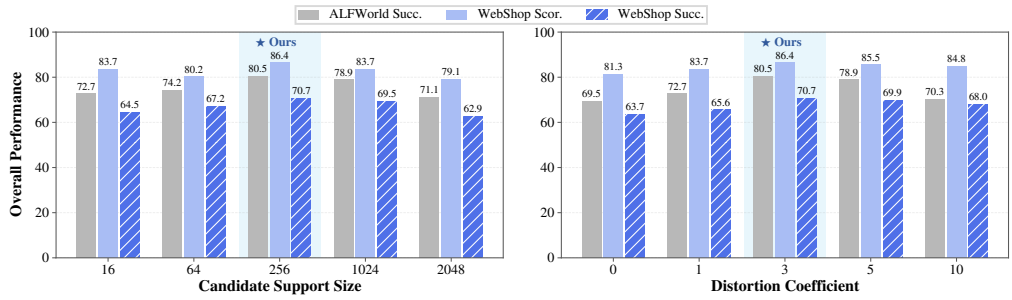


Figure 3: Ablation studies on the candidate support size K and distortion coefficient α.

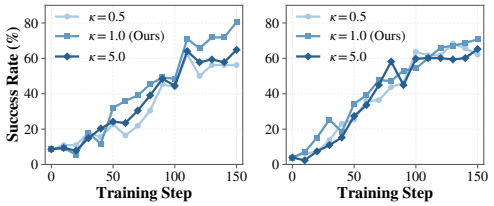


Figure 4: Effect of the temperature κ on ALFWorld (left) and WebShop (right).

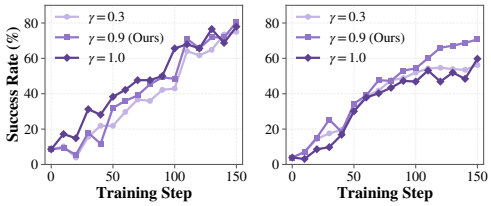


Figure 5: Effect of the decay factor γ on ALFWorld (left) and WebShop (right).

combining these complementary components, the hierarchical adaptation mechanism delivers substantially larger gains than either adaptation alone across all benchmarks.

Ablation of Step-level Strategies. Compared with Relative Position, which employs a simple linear decay to approximate the relative value of distillation at each interaction step within a trajectory, Prefix Only incorporates trajectory-specific accumulated divergence, improving ALFWorld and WebShop success by 4.6 and 6.3 points, respectively. Suffix Only nevertheless outperforms Prefix Only on WebShop Score (+1.8), reflecting the complementary value of remaining discrepancy for matching product constraints. HiSD instead combines prefix and suffix discrepancies to measure relative divergence progress within the trajectory, outperforming all variants by jointly accounting for accumulated drift and remaining discrepancy.

Ablation of Token-level Strategies. Top Selection outperforms Random Selection on ALFWorld (+5.5), Search-QA (+2.0), and WebShop success (+1.6). However, restricting random sampling to the teacher's top-256 candidates rather than the full vocabulary retains valuable distillation targets, allowing Random Selection to achieve a higher WebShop Score (+1.5). HiSD improves over Top Selection by adapting support size to token-level discrepancy, matching supervision granularity to each decision. Student Candidate Support retains the support adaptation mechanism but underperforms HiSD, as its candidate set omit tokens assigned low probability by the student despite receiving high probability under the teacher distribution, thereby weakening the teacher's guidance.

4.4 Hyperparameter Sensitivity Analysis

Temperature κ and Decay Factor γ. Figures 4 and 5 present ablations of the step-level hyperparameters κ and γ, respectively, on ALFWorld and WebShop. Across both tasks, $\kappa = 1.0$ sustains improvement throughout late-stage training by balancing sensitivity to prefix and suffix divergence. For the temporal decay factor, $\gamma = 0.9$ avoids both the aggressive discounting of distant divergence when $\gamma = 0.3$ and the absence of decay when $\gamma = 1.0$.

Candidate Support Size K and Distortion Coefficient α. Sensitivity to the token-level hyperparameters is shown in Figure 3. The left panel shows that performance does not continue to improve with K, instead declining after peaking at $K = 256$ despite the several-fold increase in computational overhead. The right panel further shows that with an overly small distortion coefficient α, support size varies little with discrepancy and degenerates at $\alpha = 0$ into a fixed support containing all 256 candidates, whereas the support becomes excessively concentrated on high-probability tokens when α is too large, discarding informative secondary candidates and fine-grained probability relations.

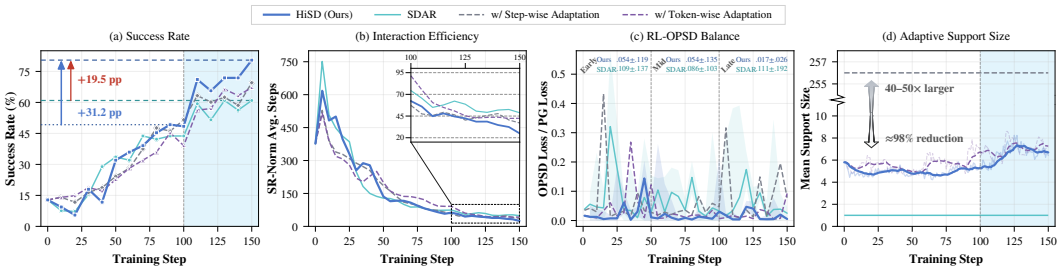


Figure 6: Training dynamics of HiSD, SDAR, and the step-wise adaptation and token-wise adaptation variants on ALFWorld using Qwen2.5-1.5B-Instruct. (a) Arrows indicate HiSD's gains in success rate over SDAR (+19.5 pp) and during late-stage training (+31.2 pp). (b) Success-rate-normalized average interaction steps, where lower values indicate that task success is achieved with fewer interactions. (c) Ratio of OPSD loss to policy-gradient (PG) loss, indicating the relative scale of distillation supervision during policy optimization. (d) Mean support size.

4.5 Training Dynamics

Overall. As shown in Figures 6(a) and (b), respectively, HiSD achieves the highest success rate without increasing the interaction length, while both its success rate and interaction efficiency show further improvement trends even after achieving the best results. Figure 6(c) further shows that, unlike SDAR and the single-level variants, HiSD exhibits substantially slight fluctuations in the ratio of OPSD loss to policy-gradient loss, maintaining a stable balance between reinforcement learning and distillation, which explains why it continues to benefit from extended training.

Adaptation Analysis. In Figures 6(a) and (b), during early-to-middle training, the Step-wise Adaptation achieves higher success rates and greater interaction efficiency than the Token-wise Adaptation by concentrating supervision on reliable prefixes whose interaction states are less affected by accumulated drift. Figure 6(d) shows that, as teacher–student discrepancy decreases during training, the adaptive support size gradually expands to preserve increasingly fine-grained distribution, enabling both the Token-wise Adaptation and HiSD to continue improving throughout late training. These complementary roles allow HiSD to combine efficient early learning from reliable trajectory segments with sustained late-stage refinement through fine-grained token-level supervision.

4.6 Computational Efficiency.

We present an empirical comparison of computational efficiency on WebShop. Figure 7(a) illustrates that HiSD's adaptive support selection lowers the overhead of distributional supervision, reducing total training-step time by 11.4% relative to GRPO+OPSD with fixed support size 256. Although HiSD incurs 27.8% more training time than SDAR with sampled-token supervision, its hierarchical adaptation enables more efficient distillation, yielding 234% and 159% higher teacher budget utilization than GRPO+OPSD and SDAR respectively, as shown in Figure 7(b).

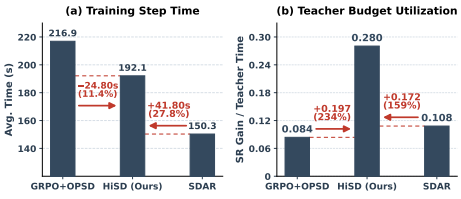


Figure 7: (a) Average wall-clock time per training step. (b) Teacher budget utilization, measured as the success-rate gain over GRPO per unit of teacher forward time.

5 Conclusion

In this work, we study the hierarchical non-uniformity of privileged teacher supervision in multi-turn OPSD, where supervision utility and the appropriate distributional scope vary across interaction steps and token predictions, respectively. HiSD addresses this through progress-aware step supervision and support-adaptive token supervision. Extensive experiments on diverse multi-turn agent benchmarks demonstrate that HiSD improves both task success and interaction efficiency, with continued gains during late-stage training even after surpassing the strongest baseline results. Our findings highlight the importance of complementary adaptation across interaction steps and within token predictions, suggesting that teacher guidance should adapt to the hierarchical structure of agent decision making.

AI USE STATEMENT

We used generative AI tools only to refine the language and expression of the manuscript, primarily to improve clarity, conciseness, and readability. All AI-assisted text was carefully reviewed and revised by the authors. We take full responsibility for the final content of this work.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Yuxuan Jiang and Francis Ferraro. Bridging reasoning trajectories in on-policy distillation via near-future guidance. *arXiv preprint arXiv:2606.00305*, 2026.

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

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

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

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

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*, 2026a.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b.

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

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

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

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

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, volume 2024, pp. 9695–9717, 2024.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

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

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

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

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

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

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*, 2026b.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025.

Haoze Wu, Chuqiao Kuang, Tianyi Zhuang, and Xiaoguang Li. Beyond outcome rewards: Step-level self-distilled policy optimization for deep search agents. *arXiv preprint arXiv:2608.12764*, 2026.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, et al. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026.

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

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

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

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026a.

Zichao Yu, Chengzhi Yu, Shengze Xu, Yujin Han, Bingqing Jiang, Xu Wang, and Difan Zou. Mismatch matters: On-policy distillation beyond token agreement. *arXiv preprint arXiv:2608.09836*, 2026b.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D Lyng, Sanjit Singh Batra, and Robert E Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25553–25569, 2026.

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

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

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 Theoretical Analysis

We adopt the notation in Section 3 and fix a token position (k, t). Let $\mathcal{V}$ denote the vocabulary and $\mathcal{C} = \mathcal{C}_{k,t} \subseteq \mathcal{V}$ the candidate set. The student and privileged teacher distributions over $\mathcal{V}$ are

$$p_{\mathcal{V}}(\cdot) = \pi_\theta(\cdot \mid c_{k,t}), \qquad q_{\mathcal{V}}(\cdot) = \mathrm{sg}[\pi_T(\cdot \mid c_{k,t}, r)]\,.$$

For any nonempty set $\mathcal{A} \subseteq \mathcal{V}$, define

$$p_{\mathcal{A}}(v) = \frac{p_{\mathcal{V}}(v)}{\sum_{w \in \mathcal{A}} p_{\mathcal{V}}(w)}, \qquad q_{\mathcal{A}}(v) = \frac{q_{\mathcal{V}}(v)}{\sum_{w \in \mathcal{A}} q_{\mathcal{V}}(w)}, \qquad v \in \mathcal{A}.$$

In particular, $p_{\mathcal{C}}$ and $q_{\mathcal{C}}$ correspond to the candidate-normalized distributions $\bar{\pi}_\theta$ and $\bar{\pi}_T$, respectively. Let z_v denote the student logit of token v and write $z_{\mathcal{A}} = (z_v)_{v \in \mathcal{A}}$, so that $p_{\mathcal{A}} = \mathrm{softmax}(z_{\mathcal{A}})$. Teacher distributions and token sets are held fixed during differentiation.

A.1 Relative Error of Sampled-Token Gradients

The forward-KL objective over $\mathcal{C}$ and its gradient with respect to $z_{\mathcal{C}}$ are

$$L_{\mathcal{C}} = D_{\mathrm{KL}}(q_{\mathcal{C}} \| p_{\mathcal{C}}), \qquad g_{\mathcal{C}} = \nabla_{z_{\mathcal{C}}} L_{\mathcal{C}} = p_{\mathcal{C}} - q_{\mathcal{C}}. \tag{A.1}$$

For each $v \in \mathcal{C}$, let $e_v \in \mathbb{R}^{|\mathcal{C}|}$ denote the corresponding one-hot vector. For an on-policy token $Y \sim p_{\mathcal{C}}$, consider the importance-corrected single-token loss

$$L_Y = -\,\mathrm{sg}\left[\frac{q_{\mathcal{C}}(Y)}{p_{\mathcal{C}}(Y)}\right] \log p_{\mathcal{C}}(Y)$$

and its gradient

$$\widehat{g}_Y = \nabla_{z_{\mathcal{C}}} L_Y = \mathrm{sg}\left[\frac{q_{\mathcal{C}}(Y)}{p_{\mathcal{C}}(Y)}\right] (p_{\mathcal{C}} - e_Y). \tag{A.2}$$

Proposition 1 (Unbiasedness of the sampled gradient)**.** *The estimator $\widehat{g}_Y$ in Eq. equation A.2 is an unbiased estimator of $g_{\mathcal{C}}$:*

$$\mathbb{E}_{Y \sim p_{\mathcal{C}}}[\widehat{g}_Y] = g_{\mathcal{C}}.$$

Proof. Taking the expectation over $Y \sim p_{\mathcal{C}}$ gives

$$\begin{aligned}
\mathbb{E}_{Y \sim p_{\mathcal{C}}}[\widehat{g}_Y] &= \sum_{v \in \mathcal{C}} p_{\mathcal{C}}(v) \frac{q_{\mathcal{C}}(v)}{p_{\mathcal{C}}(v)} (p_{\mathcal{C}} - e_v) \\
&= \sum_{v \in \mathcal{C}} q_{\mathcal{C}}(v)(p_{\mathcal{C}} - e_v) \\
&= p_{\mathcal{C}} \sum_{v \in \mathcal{C}} q_{\mathcal{C}}(v) - \sum_{v \in \mathcal{C}} q_{\mathcal{C}}(v) e_v \\
&= p_{\mathcal{C}} - q_{\mathcal{C}} = g_{\mathcal{C}}.
\end{aligned}$$

□

Proposition 2 (Relative error of the sampled gradient)**.** *The mean-squared error of the estimator in Eq. equation A.2 is*

$$\mathbb{E}_{Y \sim p_{\mathcal{C}}}\left[\|\widehat{g}_Y - g_{\mathcal{C}}\|_2^2\right] = \sum_{v \in \mathcal{C}} \frac{q_{\mathcal{C}}(v)^2}{p_{\mathcal{C}}(v)} \|p_{\mathcal{C}} - e_v\|_2^2 - \|p_{\mathcal{C}} - q_{\mathcal{C}}\|_2^2. \tag{A.3}$$

Consequently, for $p_{\mathcal{C}} \neq q_{\mathcal{C}}$, its relative mean-squared error is

$$\begin{aligned}
\mathcal{R}(p_{\mathcal{C}}, q_{\mathcal{C}}) &:= \frac{\mathbb{E}_{Y \sim p_{\mathcal{C}}}\left[\|\widehat{g}_Y - g_{\mathcal{C}}\|_2^2\right]}{\|g_{\mathcal{C}}\|_2^2} \\
&= \frac{\displaystyle\sum_{v \in \mathcal{C}} \frac{q_{\mathcal{C}}(v)^2}{p_{\mathcal{C}}(v)} \|p_{\mathcal{C}} - e_v\|_2^2 - \|p_{\mathcal{C}} - q_{\mathcal{C}}\|_2^2}{\|p_{\mathcal{C}} - q_{\mathcal{C}}\|_2^2}.
\end{aligned} \tag{A.4}$$

Proof. By Proposition 1,

$$\begin{aligned}
&\mathbb{E}_{Y\sim p_{\mathcal{C}}}\left[\|\widehat{g}_Y - g_{\mathcal{C}}\|_2^2\right] \\
&\quad= \mathbb{E}_{Y\sim p_{\mathcal{C}}}\left[\|\widehat{g}_Y\|_2^2 - 2\widehat{g}_Y^\top g_{\mathcal{C}} + \|g_{\mathcal{C}}\|_2^2\right] \\
&\quad= \mathbb{E}_{Y\sim p_{\mathcal{C}}}\|\widehat{g}_Y\|_2^2 - 2\left(\mathbb{E}_{Y\sim p_{\mathcal{C}}}[\widehat{g}_Y]\right)^\top g_{\mathcal{C}} + \|g_{\mathcal{C}}\|_2^2 \\
&\quad= \mathbb{E}_{Y\sim p_{\mathcal{C}}}\|\widehat{g}_Y\|_2^2 - 2\|g_{\mathcal{C}}\|_2^2 + \|g_{\mathcal{C}}\|_2^2 \\
&\quad= \mathbb{E}_{Y\sim p_{\mathcal{C}}}\|\widehat{g}_Y\|_2^2 - \|g_{\mathcal{C}}\|_2^2 \\
&\quad= \sum_{v\in\mathcal{C}} p_{\mathcal{C}}(v)\frac{q_{\mathcal{C}}(v)^2}{p_{\mathcal{C}}(v)^2}\|p_{\mathcal{C}} - e_v\|_2^2 - \|p_{\mathcal{C}} - q_{\mathcal{C}}\|_2^2 \\
&\quad= \sum_{v\in\mathcal{C}} \frac{q_{\mathcal{C}}(v)^2}{p_{\mathcal{C}}(v)}\|p_{\mathcal{C}} - e_v\|_2^2 - \|p_{\mathcal{C}} - q_{\mathcal{C}}\|_2^2.
\end{aligned}$$

This proves Eq. equation A.3. Dividing by $\|g_{\mathcal{C}}\|_2^2 = \|p_{\mathcal{C}} - q_{\mathcal{C}}\|_2^2 > 0$ gives Eq. equation A.4. □

Corollary 1 (Inverse-KL scaling near alignment)**.** *For fixed $p_{\mathcal{C}}$ with $|\mathcal{C}| \geq 2$, as $q_{\mathcal{C}} \to p_{\mathcal{C}}$ with $q_{\mathcal{C}} \neq p_{\mathcal{C}}$,*

$$\mathcal{R}(p_{\mathcal{C}}, q_{\mathcal{C}}) = \Theta\left(\frac{1}{D_{\mathrm{KL}}(q_{\mathcal{C}}\|p_{\mathcal{C}})}\right).$$

Here, Θ means that $\mathcal{R}(p_{\mathcal{C}}, q_{\mathcal{C}})$ is bounded above and below by positive constant multiples of $1/D_{\mathrm{KL}}(q_{\mathcal{C}}\|p_{\mathcal{C}})$ for $q_{\mathcal{C}}$ sufficiently close to $p_{\mathcal{C}}$.

Proof. The numerator of Eq. equation A.4 is continuous in $q_{\mathcal{C}}$. At $q_{\mathcal{C}} = p_{\mathcal{C}}$, it equals

$$\begin{aligned}
&\sum_{v\in\mathcal{C}} p_{\mathcal{C}}(v)\|p_{\mathcal{C}} - e_v\|_2^2 \\
&\quad= \sum_{v\in\mathcal{C}} p_{\mathcal{C}}(v)\left(\|p_{\mathcal{C}}\|_2^2 - 2p_{\mathcal{C}}(v) + 1\right) \\
&\quad= \|p_{\mathcal{C}}\|_2^2 \sum_{v\in\mathcal{C}} p_{\mathcal{C}}(v) - 2\sum_{v\in\mathcal{C}} p_{\mathcal{C}}(v)^2 + \sum_{v\in\mathcal{C}} p_{\mathcal{C}}(v) \\
&\quad= \|p_{\mathcal{C}}\|_2^2 - 2\|p_{\mathcal{C}}\|_2^2 + 1 \\
&\quad= 1 - \|p_{\mathcal{C}}\|_2^2 > 0.
\end{aligned} \tag{A.5}$$

The strict inequality follows from $p_{\mathcal{C}}(v) > 0$ for every $v \in \mathcal{C}$ and $|\mathcal{C}| \geq 2$. By continuity, the numerator remains bounded above and below by positive constants when $q_{\mathcal{C}}$ is sufficiently close to $p_{\mathcal{C}}$.

The probability differences satisfy

$$\sum_{v\in\mathcal{C}}\big(q_{\mathcal{C}}(v) - p_{\mathcal{C}}(v)\big) = \sum_{v\in\mathcal{C}} q_{\mathcal{C}}(v) - \sum_{v\in\mathcal{C}} p_{\mathcal{C}}(v) = 0.$$

A second-order Taylor expansion at $q_{\mathcal{C}} = p_{\mathcal{C}}$ gives

$$\begin{aligned}
&D_{\mathrm{KL}}(q_{\mathcal{C}}\|p_{\mathcal{C}}) \\
&\quad= \sum_{v\in\mathcal{C}} q_{\mathcal{C}}(v)\log\frac{q_{\mathcal{C}}(v)}{p_{\mathcal{C}}(v)} \\
&\quad= \sum_{v\in\mathcal{C}}\left[q_{\mathcal{C}}(v) - p_{\mathcal{C}}(v) + \frac{\big(q_{\mathcal{C}}(v) - p_{\mathcal{C}}(v)\big)^2}{2p_{\mathcal{C}}(v)}\right] + r(p_{\mathcal{C}}, q_{\mathcal{C}}) \\
&\quad= \sum_{v\in\mathcal{C}}\big(q_{\mathcal{C}}(v) - p_{\mathcal{C}}(v)\big) + \frac{1}{2}\sum_{v\in\mathcal{C}}\frac{\big(q_{\mathcal{C}}(v) - p_{\mathcal{C}}(v)\big)^2}{p_{\mathcal{C}}(v)} + r(p_{\mathcal{C}}, q_{\mathcal{C}}) \\
&\quad= \frac{1}{2}\sum_{v\in\mathcal{C}}\frac{\big(q_{\mathcal{C}}(v) - p_{\mathcal{C}}(v)\big)^2}{p_{\mathcal{C}}(v)} + r(p_{\mathcal{C}}, q_{\mathcal{C}}),
\end{aligned} \tag{A.6}$$

where the remainder satisfies

$$\lim_{q_{\mathcal{C}} \to p_{\mathcal{C}}} \frac{|r(p_{\mathcal{C}}, q_{\mathcal{C}})|}{\|q_{\mathcal{C}} - p_{\mathcal{C}}\|_2^2} = 0.$$

Let $p_{\min} = \min_{v \in \mathcal{C}} p_{\mathcal{C}}(v) > 0$ and $p_{\max} = \max_{v \in \mathcal{C}} p_{\mathcal{C}}(v)$. Then

$$\frac{1}{2p_{\max}} \|q_{\mathcal{C}} - p_{\mathcal{C}}\|_2^2 \le \frac{1}{2} \sum_{v \in \mathcal{C}} \frac{\big(q_{\mathcal{C}}(v) - p_{\mathcal{C}}(v)\big)^2}{p_{\mathcal{C}}(v)} \le \frac{1}{2p_{\min}} \|q_{\mathcal{C}} - p_{\mathcal{C}}\|_2^2. \tag{A.7}$$

Together, Eq. equation A.1,equation A.6, and equation A.7 imply

$$D_{\mathrm{KL}}(q_{\mathcal{C}} \| p_{\mathcal{C}}) = \Theta\big(\|q_{\mathcal{C}} - p_{\mathcal{C}}\|_2^2\big).$$

Combining this relation with Eq. equation A.5 yields

$$\begin{aligned} \mathcal{R}(p_{\mathcal{C}}, q_{\mathcal{C}}) &= \frac{\mathbb{E}_{Y \sim p_{\mathcal{C}}} \left[\|\widehat{g}_Y - g_{\mathcal{C}}\|_2^2\right]}{\|p_{\mathcal{C}} - q_{\mathcal{C}}\|_2^2} \\ &= \Theta\left(\frac{1}{\|p_{\mathcal{C}} - q_{\mathcal{C}}\|_2^2}\right) \\ &= \Theta\left(\frac{1}{D_{\mathrm{KL}}(q_{\mathcal{C}} \| p_{\mathcal{C}})}\right). \end{aligned}$$

□

Interpretation. Importance correction makes the sampled gradient unbiased, but its sampling variance converges to a positive constant as the teacher distribution approaches the fixed student distribution. Meanwhile, the squared norm of $g_{\mathcal{C}}$ vanishes, causing the relative mean-squared error to grow inversely with the forward KL.

A.2 Gradient Error under Support Renormalization

Fix a nonempty support $\mathcal{S} \subseteq \mathcal{C}$. For $v \in \mathcal{S}$, the support-normalized distributions satisfy

$$p_{\mathcal{S}}(v) = \frac{p_{\mathcal{C}}(v)}{\sum_{u \in \mathcal{S}} p_{\mathcal{C}}(u)}, \qquad q_{\mathcal{S}}(v) = \frac{q_{\mathcal{C}}(v)}{\sum_{u \in \mathcal{S}} q_{\mathcal{C}}(u)}. \tag{A.8}$$

The student distribution satisfies

$$\begin{aligned} p_{\mathcal{S}}(v) &= \frac{\exp(z_v) / \sum_{w \in \mathcal{C}} \exp(z_w)}{\sum_{u \in \mathcal{S}} \exp(z_u) / \sum_{w \in \mathcal{C}} \exp(z_w)} \\ &= \frac{\exp(z_v)}{\sum_{u \in \mathcal{S}} \exp(z_u)}. \end{aligned} \tag{A.9}$$

Let

$$L_{\mathcal{S}} = D_{\mathrm{KL}}(q_{\mathcal{S}} \| p_{\mathcal{S}}), \qquad h_{\mathcal{S}} = \nabla_{z_{\mathcal{C}}} L_{\mathcal{S}}, \qquad \delta = \delta_{k,t}^{\leftarrow} = D_{\mathrm{KL}}(p_{\mathcal{C}} \| q_{\mathcal{C}}), \tag{A.10}$$

and denote the discarded student and teacher probability masses by

$$\varepsilon_p = \sum_{v \in \mathcal{C} \setminus \mathcal{S}} p_{\mathcal{C}}(v), \qquad \varepsilon_q = \sum_{v \in \mathcal{C} \setminus \mathcal{S}} q_{\mathcal{C}}(v).$$

Proposition 3 (Gradient error under support renormalization)**.** *The gradient with respect to the selected logits is*

$$\nabla_{z_{\mathcal{S}}} L_{\mathcal{S}} = p_{\mathcal{S}} - q_{\mathcal{S}}. \tag{A.11}$$

The gradient with respect to all candidate logits is therefore

$$h_{\mathcal{S}}(v) = \begin{cases} p_{\mathcal{S}}(v) - q_{\mathcal{S}}(v), & v \in \mathcal{S}, \\ 0, & v \in \mathcal{C} \setminus \mathcal{S}. \end{cases} \tag{A.12}$$

Relative to $g_{\mathcal{C}}$ in Eq. equation A.1, it satisfies

$$\|h_{\mathcal{S}} - g_{\mathcal{C}}\|_2 \le 2(\varepsilon_p + \varepsilon_q) \le 2\big(\delta + \mathrm{e}\, \varepsilon_q\big), \tag{A.13}$$

where $\mathrm{e} = \exp(1)$.

Proof. Using Eq. equation A.9,

$$\begin{aligned}
L_{\mathcal{S}} &= \sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\log\frac{q_{\mathcal{S}}(v)}{p_{\mathcal{S}}(v)} \\
&= \sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\log q_{\mathcal{S}}(v) - \sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\left(z_v - \log\sum_{u\in\mathcal{S}}\exp(z_u)\right) \\
&= \sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\log q_{\mathcal{S}}(v) - \sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v) z_v \\
&\quad + \left(\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\right)\log\sum_{u\in\mathcal{S}}\exp(z_u) \\
&= \sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\log q_{\mathcal{S}}(v) - \sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v) z_v + \log\sum_{u\in\mathcal{S}}\exp(z_u).
\end{aligned}$$

For $v \in \mathcal{S}$,

$$\begin{aligned}
\frac{\partial L_{\mathcal{S}}}{\partial z_v} &= -q_{\mathcal{S}}(v) + \frac{1}{\sum_{u\in\mathcal{S}}\exp(z_u)}\frac{\partial}{\partial z_v}\sum_{u\in\mathcal{S}}\exp(z_u) \\
&= -q_{\mathcal{S}}(v) + \frac{\exp(z_v)}{\sum_{u\in\mathcal{S}}\exp(z_u)} \\
&= p_{\mathcal{S}}(v) - q_{\mathcal{S}}(v).
\end{aligned}$$

The loss is independent of z_v for $v \in \mathcal{C}\setminus\mathcal{S}$, which gives Eq. equation A.11 and equation A.12.

For $a_{\mathcal{C}} \in \{p_{\mathcal{C}}, q_{\mathcal{C}}\}$, let $a_{\mathcal{S}}$ be its renormalized restriction to $\mathcal{S}$ and let $\widetilde{a}_{\mathcal{S}}$ denote the extension of $a_{\mathcal{S}}$ by zero to $\mathcal{C}$. Write $\varepsilon_a = \sum_{v\in\mathcal{C}\setminus\mathcal{S}} a_{\mathcal{C}}(v)$. Then

$$\begin{aligned}
\|\widetilde{a}_{\mathcal{S}} - a_{\mathcal{C}}\|_1 &= \sum_{v\in\mathcal{S}}\left|\frac{a_{\mathcal{C}}(v)}{1-\varepsilon_a} - a_{\mathcal{C}}(v)\right| + \sum_{v\in\mathcal{C}\setminus\mathcal{S}} a_{\mathcal{C}}(v) \\
&= \frac{\varepsilon_a}{1-\varepsilon_a}\sum_{v\in\mathcal{S}} a_{\mathcal{C}}(v) + \varepsilon_a \\
&= \frac{\varepsilon_a}{1-\varepsilon_a}(1-\varepsilon_a) + \varepsilon_a \\
&= 2\varepsilon_a.
\end{aligned}$$

Since $h_{\mathcal{S}} = \widetilde{p}_{\mathcal{S}} - \widetilde{q}_{\mathcal{S}}$,

$$\begin{aligned}
\|h_{\mathcal{S}} - g_{\mathcal{C}}\|_2 &\le \|h_{\mathcal{S}} - g_{\mathcal{C}}\|_1 \\
&= \|(\widetilde{p}_{\mathcal{S}} - p_{\mathcal{C}}) - (\widetilde{q}_{\mathcal{S}} - q_{\mathcal{C}})\|_1 \\
&\le \|\widetilde{p}_{\mathcal{S}} - p_{\mathcal{C}}\|_1 + \|\widetilde{q}_{\mathcal{S}} - q_{\mathcal{C}}\|_1 \\
&= 2(\varepsilon_p + \varepsilon_q).
\end{aligned}$$

Let $\mathbb{I}\{\cdot\}$ denote the indicator function. To bound the student tail mass, define

$$r(v) = \frac{q_{\mathcal{C}}(v)\exp(\mathbb{I}\{v\notin\mathcal{S}\})}{Z}, \qquad v\in\mathcal{C},$$

where

$$\begin{aligned}
Z &= \sum_{v\in\mathcal{C}} q_{\mathcal{C}}(v)\exp(\mathbb{I}\{v\notin\mathcal{S}\}) \\
&= \sum_{v\in\mathcal{S}} q_{\mathcal{C}}(v) + \mathrm{e}\sum_{v\in\mathcal{C}\setminus\mathcal{S}} q_{\mathcal{C}}(v) \\
&= 1 - \varepsilon_q + \mathrm{e}\,\varepsilon_q \\
&= 1 + (\mathrm{e}-1)\varepsilon_q.
\end{aligned}$$

Nonnegativity of KL gives

$$\begin{aligned}
0 &\le D_{\mathrm{KL}}(p_{\mathcal{C}} \| r) \\
&= \sum_{v \in \mathcal{C}} p_{\mathcal{C}}(v) \log \frac{p_{\mathcal{C}}(v) Z}{q_{\mathcal{C}}(v) \exp(\mathbb{I}\{v \notin \mathcal{S}\})} \\
&= \sum_{v \in \mathcal{C}} p_{\mathcal{C}}(v) \log \frac{p_{\mathcal{C}}(v)}{q_{\mathcal{C}}(v)} - \sum_{v \in \mathcal{C} \setminus \mathcal{S}} p_{\mathcal{C}}(v) \\
&\quad + \left(\sum_{v \in \mathcal{C}} p_{\mathcal{C}}(v) \right) \log Z \\
&= \delta - \varepsilon_p + \log\bigl(1 + (\mathrm{e} - 1)\varepsilon_q\bigr).
\end{aligned}$$

Consequently, using $\log(1 + x) \le x$ for $x \ge 0$,

$$\begin{aligned}
\varepsilon_p + \varepsilon_q &\le \delta + \log\bigl(1 + (\mathrm{e} - 1)\varepsilon_q\bigr) + \varepsilon_q \\
&\le \delta + (\mathrm{e} - 1)\varepsilon_q + \varepsilon_q \\
&= \delta + \mathrm{e}\, \varepsilon_q.
\end{aligned}$$

Combining these inequalities proves Eq. equation A.13. □

Corollary 2 (Gradient fidelity under the HiSD retention threshold)**.** *Let $\mathcal{S} = \mathcal{S}_{k,t}$ be the HiSD adaptive support selected for $(p_{\mathcal{C}}, q_{\mathcal{C}})$, satisfying*

$$\sum_{v \in \mathcal{S}} q_{\mathcal{C}}(v) \ge \eta_{k,t} = \exp(-\alpha\delta)\,, \qquad \alpha > 0.$$

The gradient error obeys

$$\begin{aligned}
\|h_{\mathcal{S}} - g_{\mathcal{C}}\|_2 &\le 2\,[\delta + \mathrm{e}\,(1 - \exp(-\alpha\delta))] \\
&\le 2(1 + \mathrm{e}\alpha)\delta.
\end{aligned}$$

For $|\mathcal{C}| \ge 2$, fix $p_{\mathcal{C}}$ and α. As $q_{\mathcal{C}} \to p_{\mathcal{C}}$ with $q_{\mathcal{C}} \ne p_{\mathcal{C}}$ and $\mathcal{S}$ reselected by HiSD for each $q_{\mathcal{C}}$, the relative squared error satisfies

$$\mathcal{R}_{\mathcal{S}}(p_{\mathcal{C}}, q_{\mathcal{C}}) := \frac{\|h_{\mathcal{S}} - g_{\mathcal{C}}\|_2^2}{\|g_{\mathcal{C}}\|_2^2} = O(\delta)\,.$$

Here, $O(\delta)$ denotes an upper bound by a positive constant multiple of δ near alignment. For the sampled estimator in Eq. equation A.2, the relative mean-squared error satisfies

$$\mathcal{R}(p_{\mathcal{C}}, q_{\mathcal{C}}) = \Theta\bigl((\delta)^{-1}\bigr)\,.$$

Proof. The retained-mass constraint gives

$$\begin{aligned}
\varepsilon_q &= 1 - \sum_{v \in \mathcal{S}} q_{\mathcal{C}}(v) \\
&\le 1 - \eta_{k,t} \\
&= 1 - \exp(-\alpha\delta) \\
&\le \alpha\delta.
\end{aligned}$$

Substitution into Eq. equation A.13 yields

$$\begin{aligned}
\|h_{\mathcal{S}} - g_{\mathcal{C}}\|_2 &\le 2\bigl(\delta + \mathrm{e}\,\varepsilon_q\bigr) \\
&\le 2\,[\delta + \mathrm{e}\,(1 - \exp(-\alpha\delta))] \\
&\le 2\bigl(\delta + \mathrm{e}\alpha\delta\bigr) \\
&= 2(1 + \mathrm{e}\alpha)\delta.
\end{aligned}$$

Write $\xi = q_{\mathcal{C}} - p_{\mathcal{C}}$, with $\xi_v = q_{\mathcal{C}}(v) - p_{\mathcal{C}}(v)$. Then

$$\sum_{v \in \mathcal{C}} \xi_v = \sum_{v \in \mathcal{C}} q_{\mathcal{C}}(v) - \sum_{v \in \mathcal{C}} p_{\mathcal{C}}(v) = 0.$$

Expanding around $\xi = 0$ gives

$$\begin{aligned}
\delta &= \sum_{v\in\mathcal{C}} p_{\mathcal{C}}(v) \log \frac{p_{\mathcal{C}}(v)}{p_{\mathcal{C}}(v) + \xi_v} \\
&= -\sum_{v\in\mathcal{C}} p_{\mathcal{C}}(v) \log\left(1 + \frac{\xi_v}{p_{\mathcal{C}}(v)}\right) \\
&= -\sum_{v\in\mathcal{C}} p_{\mathcal{C}}(v) \left(\frac{\xi_v}{p_{\mathcal{C}}(v)} - \frac{\xi_v^2}{2p_{\mathcal{C}}(v)^2}\right) + O(\|\xi\|_2^3) \\
&= -\sum_{v\in\mathcal{C}} \xi_v + \frac{1}{2}\sum_{v\in\mathcal{C}} \frac{\xi_v^2}{p_{\mathcal{C}}(v)} + O(\|\xi\|_2^3) \\
&= \frac{1}{2}\sum_{v\in\mathcal{C}} \frac{\xi_v^2}{p_{\mathcal{C}}(v)} + O(\|\xi\|_2^3).
\end{aligned}$$

Applying Eq. equation A.7 to $\xi = q_{\mathcal{C}} - p_{\mathcal{C}}$ yields

$$\frac{\|\xi\|_2^2}{2p_{\max}} \le \frac{1}{2}\sum_{v\in\mathcal{C}} \frac{\xi_v^2}{p_{\mathcal{C}}(v)} \le \frac{\|\xi\|_2^2}{2p_{\min}}.$$

Since $\|\xi\|_2^3/\|\xi\|_2^2 = \|\xi\|_2 \to 0$, the cubic remainder is negligible relative to the quadratic term. Consequently,

$$\delta = \Theta(\|\xi\|_2^2) = \Theta\big(\|p_{\mathcal{C}} - q_{\mathcal{C}}\|_2^2\big) = \Theta(\|g_{\mathcal{C}}\|_2^2). \tag{A.14}$$

Therefore,

$$\begin{aligned}
0 \le \mathcal{R}_{\mathcal{S}}(p_{\mathcal{C}}, q_{\mathcal{C}}) &= \frac{\|h_{\mathcal{S}} - g_{\mathcal{C}}\|_2^2}{\|g_{\mathcal{C}}\|_2^2} \\
&\le 4(1 + \mathrm{e}\alpha)^2 \frac{(\delta)^2}{\|g_{\mathcal{C}}\|_2^2} \\
&= O\left(\frac{(\delta)^2}{\delta}\right) \\
&= O(\delta)\,.
\end{aligned}$$

By Eq. equation A.3 and equation A.5,

$$\mathbb{E}_{Y\sim p_{\mathcal{C}}}\left[\|\widehat{g}_Y - g_{\mathcal{C}}\|_2^2\right] \longrightarrow 1 - \|p_{\mathcal{C}}\|_2^2 > 0.$$

Together with Eq. equation A.14, this gives

$$\begin{aligned}
\mathcal{R}(p_{\mathcal{C}}, q_{\mathcal{C}}) &= \frac{\mathbb{E}_{Y\sim p_{\mathcal{C}}}\left[\|\widehat{g}_Y - g_{\mathcal{C}}\|_2^2\right]}{\|g_{\mathcal{C}}\|_2^2} \\
&= \Theta\left(\frac{1}{\|g_{\mathcal{C}}\|_2^2}\right) \\
&= \Theta\big((\delta)^{-1}\big)\,.
\end{aligned}$$

□

Interpretation. For a fixed context and support, distributional supervision eliminates the single-token sampling variance analyzed in Section A.1. The L_2 error of $h_{\mathcal{S}}$ relative to $g_{\mathcal{C}}$ is bounded by $2(\varepsilon_p + \varepsilon_q)$, so retaining more probability mass in both candidate distributions tightens the gradient approximation bound. Under HiSD's retention threshold, the absolute gradient error is $O(\delta)$. In the fixed-$p_{\mathcal{C}}$ alignment limit, the relative squared error tends to zero, whereas the relative mean-squared error of the sampled estimator diverges.

A.3 Relative Probability Structure Retained by Adaptive Support

Let $\mathcal{S} = \mathcal{S}_{k,t}$ be the adaptive support in Eq. equation 5, and write $m = |\mathcal{S}|$. We use the support-normalized distributions $p_{\mathcal{S}}$ and $q_{\mathcal{S}}$ and the loss $L_{\mathcal{S}}$ defined in Eq. equation A.8 and equation A.10.

Proposition 4 (Relative probability structure retained by adaptive support)**.** *The support $\mathcal{S}$ has minimum cardinality among all subsets of $\mathcal{C}$ whose probability mass under $q_{\mathcal{C}}$ is at least $\eta_{k,t}$. The Hessian of $L_{\mathcal{S}}$ with respect to $z_{\mathcal{S}}$ is*

$$H_{\mathcal{S}} := \nabla^2_{z_{\mathcal{S}}} L_{\mathcal{S}} = \mathrm{Diag}(p_{\mathcal{S}}) - p_{\mathcal{S}} p_{\mathcal{S}}^\top, \tag{A.15}$$

where $\mathrm{Diag}(p_{\mathcal{S}})$ denotes the diagonal matrix with diagonal entries given by $p_{\mathcal{S}}$. It satisfies

$$\ker(H_{\mathcal{S}}) = \mathrm{span}\{\mathbf{1}\}, \qquad \mathrm{rank}(H_{\mathcal{S}}) = m - 1, \tag{A.16}$$

where $\ker(\cdot)$ and $\mathrm{rank}(\cdot)$ denote the null space and rank of a matrix, respectively, and $\mathbf{1} \in \mathbb{R}^m$ denotes the all-ones vector. Thus, the loss has positive curvature along $m-1$ independent relative-logit directions, and common shifts of all selected logits are its only zero-curvature directions.

Proof. Write $q_{\mathcal{C},(1)} \geq \cdots \geq q_{\mathcal{C},(|\mathcal{C}|)}$ for the entries of $q_{\mathcal{C}}$ in nonincreasing order. By construction,

$$m = \min\left\{ j \in \{1, \ldots, |\mathcal{C}|\} : \sum_{i=1}^{j} q_{\mathcal{C},(i)} \geq \eta_{k,t} \right\}, \qquad \sum_{v \in \mathcal{S}} q_{\mathcal{C}}(v) = \sum_{i=1}^{m} q_{\mathcal{C},(i)} \geq \eta_{k,t}.$$

For any $\mathcal{A} \subseteq \mathcal{C}$ with $|\mathcal{A}| = j < m$,

$$\sum_{v \in \mathcal{A}} q_{\mathcal{C}}(v) \leq \sum_{i=1}^{j} q_{\mathcal{C},(i)} < \eta_{k,t}.$$

Hence, no support smaller than $\mathcal{S}$ satisfies the retained-mass constraint.

Let $Z_{\mathcal{S}} = \sum_{w \in \mathcal{S}} \exp(z_w)$. Differentiating Eq. equation A.11, for $u, v \in \mathcal{S}$,

$$\begin{aligned}
\frac{\partial^2 L_{\mathcal{S}}}{\partial z_v \partial z_u} &= \frac{\partial}{\partial z_v}\big(p_{\mathcal{S}}(u) - q_{\mathcal{S}}(u)\big) \\
&= \frac{\partial p_{\mathcal{S}}(u)}{\partial z_v} \\
&= \frac{\mathbb{I}\{u = v\} \exp(z_u) Z_{\mathcal{S}} - \exp(z_u)\exp(z_v)}{Z_{\mathcal{S}}^2} \\
&= \mathbb{I}\{u = v\} \frac{\exp(z_u)}{Z_{\mathcal{S}}} - \frac{\exp(z_u)}{Z_{\mathcal{S}}} \frac{\exp(z_v)}{Z_{\mathcal{S}}} \\
&= p_{\mathcal{S}}(u)\big(\mathbb{I}\{u = v\} - p_{\mathcal{S}}(v)\big).
\end{aligned}$$

This gives Eq. equation A.15. To characterize the null space of $H_{\mathcal{S}}$, expand its quadratic form using $\sum_{v \in \mathcal{S}} p_{\mathcal{S}}(v) = 1$. For any $a = (a_v)_{v \in \mathcal{S}} \in \mathbb{R}^m$,

$$\begin{aligned}
a^\top H_{\mathcal{S}} a &= \sum_{v \in \mathcal{S}} p_{\mathcal{S}}(v) a_v^2 - \left(\sum_{v \in \mathcal{S}} p_{\mathcal{S}}(v) a_v \right)^2 \\
&= \frac{1}{2}\Bigg[\left(\sum_{u \in \mathcal{S}} p_{\mathcal{S}}(u) a_u^2 \right) \left(\sum_{v \in \mathcal{S}} p_{\mathcal{S}}(v) \right) \\
&\qquad + \left(\sum_{v \in \mathcal{S}} p_{\mathcal{S}}(v) a_v^2 \right) \left(\sum_{u \in \mathcal{S}} p_{\mathcal{S}}(u) \right) \\
&\qquad - 2 \left(\sum_{u \in \mathcal{S}} p_{\mathcal{S}}(u) a_u \right) \left(\sum_{v \in \mathcal{S}} p_{\mathcal{S}}(v) a_v \right) \Bigg] \\
&= \frac{1}{2} \sum_{u,v \in \mathcal{S}} p_{\mathcal{S}}(u) p_{\mathcal{S}}(v) \left(a_u^2 + a_v^2 - 2 a_u a_v \right) \\
&= \frac{1}{2} \sum_{u,v \in \mathcal{S}} p_{\mathcal{S}}(u) p_{\mathcal{S}}(v) (a_u - a_v)^2 \geq 0.
\end{aligned}$$

Since $p_{\mathcal{S}}(v) > 0$ for every $v \in \mathcal{S}$, this quadratic form vanishes exactly when $a = c\mathbf{1}$ for some $c \in \mathbb{R}$. Moreover,

$$\begin{aligned} H_{\mathcal{S}}\mathbf{1} &= \left(\mathrm{Diag}(p_{\mathcal{S}}) - p_{\mathcal{S}}p_{\mathcal{S}}^{\top}\right)\mathbf{1} \\ &= p_{\mathcal{S}} - p_{\mathcal{S}}\sum_{v\in\mathcal{S}} p_{\mathcal{S}}(v) = 0. \end{aligned}$$

These identities show that the null space consists precisely of constant vectors. Applying the rank–nullity theorem gives

$$\ker(H_{\mathcal{S}}) = \mathrm{span}\{\mathbf{1}\}, \qquad \mathrm{rank}(H_{\mathcal{S}}) = m - \dim\ker(H_{\mathcal{S}}) = m - 1.$$

□

Corollary 3 (Quadratic penalty on relative logit errors near alignment)**.** *Fix $\mathcal{S}$ and $q_{\mathcal{S}}$, and define the centered logit displacement*

$$\Delta_v = z_v - \log q_{\mathcal{S}}(v) - \sum_{u\in\mathcal{S}} q_{\mathcal{S}}(u)\left[z_u - \log q_{\mathcal{S}}(u)\right], \qquad v \in \mathcal{S}. \tag{A.17}$$

Writing $\Delta = (\Delta_v)_{v\in\mathcal{S}} \in \mathbb{R}^m$, as $\|\Delta\|_2 \to 0$,

$$\begin{aligned} D_{\mathrm{KL}}(q_{\mathcal{S}}\|p_{\mathcal{S}}) = \frac{1}{4}\sum_{u,v\in\mathcal{S}} & q_{\mathcal{S}}(u)q_{\mathcal{S}}(v) \\ & \times \left(\log\frac{p_{\mathcal{S}}(u)}{p_{\mathcal{S}}(v)} - \log\frac{q_{\mathcal{S}}(u)}{q_{\mathcal{S}}(v)}\right)^2 + O(\|\Delta\|_2^3). \end{aligned} \tag{A.18}$$

Proof. Let

$$\mu = \sum_{u\in\mathcal{S}} q_{\mathcal{S}}(u)\left[z_u - \log q_{\mathcal{S}}(u)\right], \qquad z_v^* = \log q_{\mathcal{S}}(v), \quad v \in \mathcal{S},$$

and write $z_{\mathcal{S}}^* = (z_v^*)_{v\in\mathcal{S}}$. The centering in Eq. equation A.17 ensures that Δ has zero mean under $q_{\mathcal{S}}$:

$$\begin{aligned} \sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\Delta_v &= \sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\left[z_v - \log q_{\mathcal{S}}(v) - \mu\right] \\ &= \mu - \mu\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v) = 0. \end{aligned} \tag{A.19}$$

Since $z_v = z_v^* + \Delta_v + \mu$,

$$\begin{aligned} p_{\mathcal{S}}(v) &= \frac{\exp(z_v^* + \Delta_v + \mu)}{\sum_{u\in\mathcal{S}}\exp(z_u^* + \Delta_u + \mu)} \\ &= \frac{\exp(\mu)q_{\mathcal{S}}(v)\exp(\Delta_v)}{\exp(\mu)\sum_{u\in\mathcal{S}} q_{\mathcal{S}}(u)\exp(\Delta_u)} \\ &= \frac{q_{\mathcal{S}}(v)\exp(\Delta_v)}{\sum_{u\in\mathcal{S}} q_{\mathcal{S}}(u)\exp(\Delta_u)} \\ &= \frac{\exp(z_v^* + \Delta_v)}{\sum_{u\in\mathcal{S}}\exp(z_u^* + \Delta_u)}. \end{aligned}$$

The common shift leaves the support-normalized distribution unchanged, so

$$L_{\mathcal{S}}(z_{\mathcal{S}}) = L_{\mathcal{S}}(z_{\mathcal{S}}^* + \Delta).$$

At $z_{\mathcal{S}} = z_{\mathcal{S}}^*$,

$$\begin{aligned} p_{\mathcal{S}}(v) &= \frac{\exp(\log q_{\mathcal{S}}(v))}{\sum_{u\in\mathcal{S}}\exp(\log q_{\mathcal{S}}(u))} \\ &= \frac{q_{\mathcal{S}}(v)}{\sum_{u\in\mathcal{S}} q_{\mathcal{S}}(u)} \\ &= q_{\mathcal{S}}(v). \end{aligned}$$

Substituting $p_{\mathcal{S}} = q_{\mathcal{S}}$ into the loss definition, the gradient in Eq. equation A.11, and the Hessian in Eq. equation A.15 gives

$$\begin{aligned} L_{\mathcal{S}}(z^*_{\mathcal{S}}) &= \sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\log\frac{q_{\mathcal{S}}(v)}{q_{\mathcal{S}}(v)} = 0, \\ \nabla_{z_{\mathcal{S}}} L_{\mathcal{S}}(z^*_{\mathcal{S}}) &= q_{\mathcal{S}} - q_{\mathcal{S}} = 0, \\ H^*_{\mathcal{S}} := \nabla^2_{z_{\mathcal{S}}} L_{\mathcal{S}}(z^*_{\mathcal{S}}) &= \mathrm{Diag}(q_{\mathcal{S}}) - q_{\mathcal{S}} q_{\mathcal{S}}^\top. \end{aligned}$$

The loss and gradient vanish at $z^*_{\mathcal{S}}$, so the second-order Taylor expansion starts with the quadratic term. The centering identity in Eq. equation A.19 further eliminates the squared weighted mean:

$$\begin{aligned} L_{\mathcal{S}}(z_{\mathcal{S}}) &= L_{\mathcal{S}}(z^*_{\mathcal{S}} + \Delta) \\ &= L_{\mathcal{S}}(z^*_{\mathcal{S}}) + \nabla_{z_{\mathcal{S}}} L_{\mathcal{S}}(z^*_{\mathcal{S}})^\top \Delta \\ &\quad + \frac{1}{2}\Delta^\top H^*_{\mathcal{S}}\Delta + O(\|\Delta\|_2^3) \\ &= \frac{1}{2}\Delta^\top \left(\mathrm{Diag}(q_{\mathcal{S}}) - q_{\mathcal{S}} q_{\mathcal{S}}^\top\right)\Delta + O(\|\Delta\|_2^3) \\ &= \frac{1}{2}\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\Delta_v^2 - \frac{1}{2}\left(\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\Delta_v\right)^2 + O(\|\Delta\|_2^3) \\ &= \frac{1}{2}\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\Delta_v^2 + O(\|\Delta\|_2^3). \end{aligned}$$

Using $\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v) = 1$ and the centering identity in Eq. equation A.19, the pairwise differences satisfy

$$\begin{aligned} &\sum_{u,v\in\mathcal{S}} q_{\mathcal{S}}(u)q_{\mathcal{S}}(v)(\Delta_u - \Delta_v)^2 \\ &\quad = \sum_{u,v\in\mathcal{S}} q_{\mathcal{S}}(u)q_{\mathcal{S}}(v)\left(\Delta_u^2 + \Delta_v^2 - 2\Delta_u\Delta_v\right) \\ &\quad = \left(\sum_{u\in\mathcal{S}} q_{\mathcal{S}}(u)\Delta_u^2\right)\left(\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\right) \\ &\qquad + \left(\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\Delta_v^2\right)\left(\sum_{u\in\mathcal{S}} q_{\mathcal{S}}(u)\right) \\ &\qquad - 2\left(\sum_{u\in\mathcal{S}} q_{\mathcal{S}}(u)\Delta_u\right)\left(\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\Delta_v\right) \\ &\quad = 2\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\Delta_v^2 - 2\left(\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\Delta_v\right)^2 \\ &\quad = 2\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\Delta_v^2. \end{aligned}$$

Rearranging this identity gives the pairwise representation of the quadratic term:

$$\frac{1}{2}\sum_{v\in\mathcal{S}} q_{\mathcal{S}}(v)\Delta_v^2 = \frac{1}{4}\sum_{u,v\in\mathcal{S}} q_{\mathcal{S}}(u)q_{\mathcal{S}}(v)(\Delta_u - \Delta_v)^2. \tag{A.20}$$

The shared softmax normalizer cancels in the ratio of two support probabilities:

$$\frac{p_{\mathcal{S}}(u)}{p_{\mathcal{S}}(v)} = \frac{\exp(z_u)/\sum_{w\in\mathcal{S}}\exp(z_w)}{\exp(z_v)/\sum_{w\in\mathcal{S}}\exp(z_w)} = \exp(z_u - z_v).$$

Using this ratio identity and the definition of Δ_v in Eq. equation A.17 gives

$$\begin{aligned} \Delta_u - \Delta_v &= [z_u - \log q_{\mathcal{S}}(u) - \mu] - [z_v - \log q_{\mathcal{S}}(v) - \mu] \\ &= (z_u - z_v) - [\log q_{\mathcal{S}}(u) - \log q_{\mathcal{S}}(v)] \\ &= \log\frac{p_{\mathcal{S}}(u)}{p_{\mathcal{S}}(v)} - \log\frac{q_{\mathcal{S}}(u)}{q_{\mathcal{S}}(v)}. \end{aligned} \tag{A.21}$$

Substituting Eq. equation A.20 and equation A.21 into the Taylor expansion yields Eq. equation A.18. □

Interpretation. A support of size m constrains $m-1$ independent relative-logit directions. Near alignment, the support-normalized KL loss penalizes errors in pairwise log-probability ratios, weighted by the products of the corresponding support-normalized teacher probabilities.

A.4 Gradient Interpretation of L_1 Discrepancy Coverage

To interpret the L_1 discrepancy coverage in Figure 1(d), consider the vocabulary distributions $p_{\mathcal{V}}$ and $q_{\mathcal{V}}$. For $p_{\mathcal{V}} \neq q_{\mathcal{V}}$ and a fixed subset $S \subseteq \mathcal{V}$, define the fraction of the L_1 discrepancy contained in S as

$$C_1(S) = \frac{\sum_{v\in S} |p_{\mathcal{V}}(v) - q_{\mathcal{V}}(v)|}{\sum_{v\in\mathcal{V}} |p_{\mathcal{V}}(v) - q_{\mathcal{V}}(v)|} = \frac{\sum_{v\in S} |p_{\mathcal{V}}(v) - q_{\mathcal{V}}(v)|}{\|p_{\mathcal{V}} - q_{\mathcal{V}}\|_1}. \tag{A.22}$$

Let

$$L_{\mathcal{V}}(z_{\mathcal{V}}) = D_{\mathrm{KL}}(q_{\mathcal{V}} \| p_{\mathcal{V}}), \qquad g_{\mathcal{V}} = \nabla_{z_{\mathcal{V}}} L_{\mathcal{V}}(z_{\mathcal{V}}).$$

Proposition 5 (L_1 discrepancy coverage as gradient coverage)**.** *For $p_{\mathcal{V}} \neq q_{\mathcal{V}}$ and any fixed $S \subseteq \mathcal{V}$,*

$$C_1(S) = \frac{\|g_S\|_1}{\|g_{\mathcal{V}}\|_1}, \tag{A.23}$$

where $g_S = (g_{\mathcal{V}}(v))_{v\in S}$ contains the coordinates of the full-vocabulary forward-KL gradient $g_{\mathcal{V}}$ indexed by S.

Proof. Expanding the forward KL using $p_{\mathcal{V}} = \mathrm{softmax}(z_{\mathcal{V}})$ and $\sum_{v\in\mathcal{V}} q_{\mathcal{V}}(v) = 1$ gives

$$\begin{aligned}
L_{\mathcal{V}}(z_{\mathcal{V}}) &= \sum_{v\in\mathcal{V}} q_{\mathcal{V}}(v) \log \frac{q_{\mathcal{V}}(v)}{p_{\mathcal{V}}(v)} \\
&= \sum_{v\in\mathcal{V}} q_{\mathcal{V}}(v) \log q_{\mathcal{V}}(v) - \sum_{v\in\mathcal{V}} q_{\mathcal{V}}(v) \log p_{\mathcal{V}}(v) \\
&= \sum_{v\in\mathcal{V}} q_{\mathcal{V}}(v) \log q_{\mathcal{V}}(v) - \sum_{v\in\mathcal{V}} q_{\mathcal{V}}(v) \log \frac{\exp(z_v)}{\sum_{w\in\mathcal{V}} \exp(z_w)} \\
&= \sum_{v\in\mathcal{V}} q_{\mathcal{V}}(v) \log q_{\mathcal{V}}(v) - \sum_{v\in\mathcal{V}} q_{\mathcal{V}}(v) \left(z_v - \log \sum_{w\in\mathcal{V}} \exp(z_w) \right) \\
&= \sum_{v\in\mathcal{V}} q_{\mathcal{V}}(v) \log q_{\mathcal{V}}(v) - \sum_{v\in\mathcal{V}} q_{\mathcal{V}}(v) z_v + \log \sum_{w\in\mathcal{V}} \exp(z_w).
\end{aligned}$$

Differentiating this expression with respect to z_v, for $v \in \mathcal{V}$, yields

$$\begin{aligned}
g_{\mathcal{V}}(v) &= \frac{\partial L_{\mathcal{V}}}{\partial z_v} \\
&= -q_{\mathcal{V}}(v) + \frac{1}{\sum_{w\in\mathcal{V}} \exp(z_w)} \frac{\partial}{\partial z_v} \sum_{w\in\mathcal{V}} \exp(z_w) \\
&= -q_{\mathcal{V}}(v) + \frac{\exp(z_v)}{\sum_{w\in\mathcal{V}} \exp(z_w)} \\
&= p_{\mathcal{V}}(v) - q_{\mathcal{V}}(v).
\end{aligned} \tag{A.24}$$

Substituting Eq. equation A.24 into the gradient norm ratio and using Eq. equation A.22 gives

$$\begin{aligned}
\frac{\|g_S\|_1}{\|g_{\mathcal{V}}\|_1} &= \frac{\sum_{v\in S} |g_{\mathcal{V}}(v)|}{\sum_{v\in\mathcal{V}} |g_{\mathcal{V}}(v)|} \\
&= \frac{\sum_{v\in S} |p_{\mathcal{V}}(v) - q_{\mathcal{V}}(v)|}{\sum_{v\in\mathcal{V}} |p_{\mathcal{V}}(v) - q_{\mathcal{V}}(v)|} \\
&= C_1(S).
\end{aligned}$$

□

Corollary 4 (Retained first-order descent). *Under the conditions of Proposition 5, let $u = (u_v)_{v\in\mathcal{V}} \in \mathbb{R}^{|\mathcal{V}|}$ denote a perturbation of the student logits $z_\mathcal{V}$. For a uniform coordinate-wise logit budget $\epsilon > 0$, define the maximal first-order KL decrease as*

$$\mathcal{I}_S(\epsilon) = \max_{\substack{\|u\|_\infty \le \epsilon \\ u_v = 0 \text{ for } v\in\mathcal{V}\setminus S}} -g_\mathcal{V}^\top u. \tag{A.25}$$

The fraction of the maximal first-order decrease retained by S is

$$\frac{\mathcal{I}_S(\epsilon)}{\mathcal{I}_\mathcal{V}(\epsilon)} = C_1(S).$$

Proof. As $u \to 0$, a first-order Taylor expansion at $z_\mathcal{V}$ identifies $-g_\mathcal{V}^\top u$ as the linear term in the KL decrease:

$$L_\mathcal{V}(z_\mathcal{V} + u) = L_\mathcal{V}(z_\mathcal{V}) + g_\mathcal{V}^\top u + O(\|u\|_2^2),$$
$$L_\mathcal{V}(z_\mathcal{V}) - L_\mathcal{V}(z_\mathcal{V} + u) = -g_\mathcal{V}^\top u + O(\|u\|_2^2).$$

The support constraint and coordinate budget in Eq. equation A.25 imply that every feasible perturbation satisfies

$$\begin{aligned} -g_\mathcal{V}^\top u &= -\sum_{v\in\mathcal{V}} g_\mathcal{V}(v)u_v \\ &= -\sum_{v\in S} g_\mathcal{V}(v)u_v \\ &\le \sum_{v\in S} |g_\mathcal{V}(v)|\,|u_v| \\ &\le \epsilon \sum_{v\in S} |g_\mathcal{V}(v)| \\ &= \epsilon\|g_S\|_1. \end{aligned}$$

To attain this bound, choose $u^* = (u_v^*)_{v\in\mathcal{V}}$ with

$$u_v^* = \begin{cases} -\epsilon\,\mathrm{sign}(g_\mathcal{V}(v)), & v \in S, \\ 0, & v \in \mathcal{V}\setminus S, \end{cases}$$

where sign denotes the sign function, with $\mathrm{sign}(0) = 0$. This perturbation satisfies the support constraint and $\|u^*\|_\infty \le \epsilon$, and its first-order decrease is

$$\begin{aligned} -g_\mathcal{V}^\top u^* &= -\sum_{v\in S} g_\mathcal{V}(v)\big(-\epsilon\,\mathrm{sign}(g_\mathcal{V}(v))\big) \\ &= \epsilon \sum_{v\in S} g_\mathcal{V}(v)\,\mathrm{sign}(g_\mathcal{V}(v)) \\ &= \epsilon \sum_{v\in S} |g_\mathcal{V}(v)| \\ &= \epsilon\|g_S\|_1. \end{aligned}$$

The upper bound and the attaining perturbation determine $\mathcal{I}_S(\epsilon)$. Applying the same argument with $S = \mathcal{V}$ and using Eq. equation A.24 gives

$$\mathcal{I}_S(\epsilon) = \epsilon\|g_S\|_1,$$
$$\mathcal{I}_\mathcal{V}(\epsilon) = \epsilon\|g_\mathcal{V}\|_1 = \epsilon\|p_\mathcal{V} - q_\mathcal{V}\|_1 > 0.$$

The strict inequality follows from $\epsilon > 0$ and $p_\mathcal{V} \ne q_\mathcal{V}$. Dividing these optimal values and applying Eq. equation A.23 yields

$$\frac{\mathcal{I}_S(\epsilon)}{\mathcal{I}_\mathcal{V}(\epsilon)} = \frac{\epsilon\|g_S\|_1}{\epsilon\|g_\mathcal{V}\|_1} = \frac{\|g_S\|_1}{\|g_\mathcal{V}\|_1} = C_1(S).$$

□

Interpretation. The L_1 discrepancy coverage exactly measures the fraction of the L_1 norm of $g_{\mathcal{V}}$ contributed by coordinates in S. Under a uniform coordinate-wise logit budget, it also equals the fraction of the maximal first-order KL decrease attainable by modifying only the logits in S. For any coverage threshold $\eta_{\text{cov}} \in [0, 1]$, $C_1(S) \geq \eta_{\text{cov}}$ guarantees that these coordinates retain at least an η_{cov} fraction of this first-order descent capacity.

A.5 Gating Interpretation of Relative Progress

Proposition 6 (Relative progress as a discrepancy gate)**.** *The relative-progress weight λ_k can be interpreted as a sigmoid gate on the difference between remaining suffix discrepancy and accumulated prefix discrepancy:*

$$\lambda_k = \sigma\left(\frac{D_k^{\text{suf}} - D_k^{\text{pre}}}{\kappa}\right), \qquad \sigma(x) = \frac{1}{1+\exp(-x)}.$$

Proof. Dividing the numerator and denominator of the relative-progress weight by $\exp(-D_k^{\text{pre}}/\kappa)$ gives

$$\begin{aligned}
\lambda_k &= \frac{\exp(-D_k^{\text{pre}}/\kappa)}{\exp(-D_k^{\text{pre}}/\kappa) + \exp(-D_k^{\text{suf}}/\kappa)} \\
&= \frac{1}{1+\exp\big((D_k^{\text{pre}} - D_k^{\text{suf}})/\kappa\big)} \\
&= \frac{1}{1+\exp\big(-(D_k^{\text{suf}} - D_k^{\text{pre}})/\kappa\big)} \\
&= \sigma\left(\frac{D_k^{\text{suf}} - D_k^{\text{pre}}}{\kappa}\right).
\end{aligned}$$

□

Interpretation. The difference $D_k^{\text{suf}} - D_k^{\text{pre}}$ measures the balance between remaining and accumulated discrepancy. Holding the other term fixed, larger suffix discrepancy increases the supervision weight, whereas larger prefix discrepancy decreases it.

B Experimental Details

B.1 Benchmarks and Evaluation Details

We evaluate HiSD and various baselines across three multi-turn agent benchmarks covering embodied household environment, retrieval-augmented question answering, and simulated web navigation.

- **ALFWorld** (Shridhar et al., 2020) provides text-based household environments in which an agent executes multi-step object manipulations from natural-language instructions. We evaluate six task types, including Pick (placing one target object), Clean (cleaning and placing an object), Heat (heating and placing an object), Cool (cooling and placing an object), Look (examining an object under a light source), and Pick2 (placing two target objects), with a maximum of 50 interaction steps per episode, and report the success rate for each task type and the overall success rate.
- **Search-QA** follows the Search-R1 (Jin et al., 2025) setting, using NQ (real user queries with answers grounded in Wikipedia) (Kwiatkowski et al., 2019) and HotpotQA (multi-hop questions requiring evidence aggregation across Wikipedia passages) (Yang et al., 2018) as representative single-hop and multi-hop training datasets, respectively, with their held-out splits serving as in-domain evaluation. Out-of-domain evaluation covers the single-hop datasets TriviaQA (trivia questions supported by web evidence) (Joshi et al., 2017) and PopQA (entity-centric questions emphasizing long-tail knowledge) (Mallen et al., 2023), together with the multi-hop datasets 2WikiMultiHopQA (questions annotated with reasoning paths) (Ho et al., 2020), MuSiQue (compositional questions requiring multiple reasoning steps) (Trivedi et al., 2022), and Bamboogle (challenging questions requiring iterative search) (Press et al., 2023). Agents access an E5 retriever for at most four retrieval steps, and performance is measured by Exact Match on each dataset and the weighted average across all seven datasets.

- **WebShop** (Yao et al., 2022a) requires an agent to search for products, navigate item pages, select required options, and complete a purchase according to a natural-language instruction within 15 interaction steps. In addition to Success Rate, which measures the proportion of episodes that complete the purchase with all instruction constraints satisfied, we report Score to quantify partial task completion according to how well the selected product satisfies the requested attributes, options, and price constraints.

B.2 Baseline Implementations.

We primarily compare HiSD with three representative categories of baselines corresponding to distinct forms of agent supervision: prompt-based methods that improve agent behavior through reasoning and reflection without parameter updates, RL-based methods driven by sparse outcome rewards, and OPSD-based methods that leverage privileged teachers to provide additional token- or step-level guidance during policy optimization.

- **Prompt-based Agentic Methods.** ReAct (Yao et al., 2022b) interleaves reasoning, actions, and environment observations throughout each trajectory, while Reflexion (Shinn et al., 2023) further converts task feedback into linguistic self-reflections that are stored in episodic memory and incorporated into subsequent trials. Both methods modify only the inference context without updating model parameters.
- **RL-based Methods.** GRPO (Guo et al., 2025) optimizes the policy from final environment rewards by estimating group-relative advantages over multiple trajectories sampled for each task. Under the same optimization framework, Skill-GRPO augments the policy with task-relevant skills from SkillRL (Xia et al., 2026), which distills past interaction trajectories into reusable general and task-specific behavioral guidance. This baseline evaluates whether directly incorporating such experience-derived skills can improve policy learning under sparse outcome rewards.
- **OPSD-based Methods.** These methods construct a privileged teacher from the student policy and provide dense supervision on student-generated sequences. SSPO (Wu et al., 2026) aggregates teacher–student log-probability gaps across all tokens within each interaction step, converts the resulting step-level signal into a clipped multiplier on the GRPO advantage, and applies this reweighting only to incorrect trajectories while preserving the original advantage for correct trajectories. RLSD (Yang et al., 2026) transforms each sampled token's teacher–student log-probability gap into a clipped, direction-aware weight for its GRPO advantage, so the environment reward determines the update direction while the privileged signal controls only its magnitude. Skill-SD (Wang et al., 2026a) uses retrieved skills as privileged information for the teacher and augments GRPO with an importance-weighted K_3-divergence loss over student-sampled tokens. SDAR (Lu et al., 2026) preserves the original GRPO advantage and maps each sampled token's detached teacher–student log-probability gap through a sigmoid gate, strengthening auxiliary distillation for tokens assigned higher probability by the teacher and attenuating it otherwise. The pseudocodes of these methods are provided in Section E.

B.3 Ablation Implementations

Component Ablations. We construct a GRPO+OPSD baseline with a constant step-level coefficient and a fixed teacher top-256 support, then introduce progress-aware step supervision and support-adaptive token supervision separately. In addition, we replace the step- and token-level strategies with alternative variants to further examine the effects of their specific designs. This setup isolates the individual and complementary contributions of the two components under the same GRPO+OPSD training framework, while providing further evidence for the effectiveness and rationale of our adaptive supervision strategies.

- **GRPO+OPSD**, applies a shared distillation coefficient across interaction steps and distills the teacher distribution over a fixed support comprising its top-256 tokens.
- **w/ Step-wise Adaptation**, replaces the uniform step-level coefficient in GRPO+OPSD with the progress-aware adaptive coefficient in HiSD, while retaining a fixed support comprising all 256 teacher candidates.

- **w/ Token-wise Adaptation**, with the constant distillation coefficient across interaction steps, introduces HiSD's support-adaptive token supervision to adjust the support size within the 256 teacher candidates according to instantaneous teacher–student divergence.

Step-level Coefficient Strategies. We retain HiSD's support-adaptive token supervision across all variants and modify only the coefficient assigned to each interaction step. Relative Position provides a fixed positional heuristic, while Prefix Only and Suffix Only separately retain the prefix- and suffix-divergence terms used in HiSD, providing a finer-grained ablation of the two components underlying its step-level adaptation.

- **Prefix Only**, sets the suffix divergence D_k^{suf} in Eq. 2 to 0, so that the corresponding exponential term becomes 1 and the coefficient further simplifies to a sigmoid function of D_k^{pre}, i.e., $\lambda_k^{\mathrm{pre}} = \frac{\exp(-D_k^{\mathrm{pre}}/\kappa)}{\exp(-D_k^{\mathrm{pre}}/\kappa)+1} = \sigma(-D_k^{\mathrm{pre}}/\kappa)$. The resulting coefficient reflects only the divergence accumulated before the current decision, without accounting for the discrepancy for which teacher guidance remains valuable along the subsequent trajectory.
- **Suffix Only**, removes the prefix contribution from Eq. 2 by setting D_k^{pre} to 0, resulting in the prefix exponential term $\exp(-D_k^{\mathrm{pre}}/\kappa) = 1$ and the simplified coefficient $\lambda_k^{\mathrm{suf}} = \frac{1}{1+\exp(-D_k^{\mathrm{suf}}/\kappa)} = \sigma(D_k^{\mathrm{suf}}/\kappa)$. This formulation allocates supervision solely according to the discrepancy remaining from the current decision onward, without considering the divergence already accumulated through preceding interactions.
- **Relative Position**, represents a simplified heuristic for supervision attenuation, assigning coefficients according to relative step position. For a trajectory containing N interaction steps, the coefficients follow the linear schedule $\lambda_k^{\mathrm{pos}} = (N+1-k)/N$, decreasing from 1 toward 0 along the trajectory. To prevent supervision at the final interaction step from vanishing, the coefficient is set to zero at the virtual $(N+1)$-th step, such that the final interaction step receives $\lambda_N^{\mathrm{pos}} = 1/N$.

Token-level Support Strategies. We maintain HiSD's progress-aware step supervision and vary only the construction of token support for distributional distillation. Random Selection and Top Selection adopt matched fixed support sizes, whereas Student Candidate Support preserves discrepancy-adaptive support but changes the candidate source from teacher to student, enabling controlled comparisons of selection criterion, support adaptation, and candidate source.

- **Random Selection**, constructs a fixed-size support by randomly sampling from the teacher's top-256 candidates for each token prediction, with support sizes approximating the respective HiSD averages (ALFWorld: 6, Search-QA: 11, and WebShop: 4) to maintain a comparable supervision budget.
- **Top Selection**, selects a fixed number of highest-probability tokens from the teacher's top-256 candidates (ALFWorld: 6, Search-QA: 11, and WebShop: 4), permitting a comparison between static top-ranked and discrepancy-adaptive support supervision at comparable distillation cost.
- **Student Candidate Support**, constructs the candidate support from the top-256 tokens of the student distribution rather than the teacher distribution, while retaining HiSD's adaptive mechanism to determine the support size within these candidates, thereby isolating the effect of candidate source on distributional guidance.

B.4 Implementations of Motivation Experiments

We provide detailed settings for the four motivation experiments in Figure 1. The first two experiments analyze progressive distribution shift and teacher recovery across interaction steps, while the latter two examine sampled-token gradient fidelity and teacher–student discrepancy concentration at the token level. Figures 1(a), (b), and (d) are conducted on 128 ALFWorld instances using Qwen2.5-7B-Instruct. The student operates without privileged information, while the teacher is constructed from the same policy by incorporating the corresponding skill as privileged information (Xia et al., 2026). Consistent with their respective roles in HiSD, Figures 1(a), (b) and (d) use reverse KL, $D_{\mathrm{KL}}(\pi_\theta \| \pi_T)$, to analyze the divergence signal underlying step- and token-level adaptation, respectively, whereas Figure 1(c) uses forward KL, $D_{\mathrm{KL}}(\pi_T \| \pi_\theta)$, to examine the gradients induced by distillation objective.

Step-wise Teacher–Student Divergence. We compute the teacher–student divergence at each interaction step using the step-wise reverse KL defined in Section. 3.2. At each step index, we aggregate the resulting divergence over all trajectories reaching that step and report the mean and median together with the P25–P75 and P10–P90 intervals in Figure 1(a). It illustrates that teacher–student divergence progressively increases over interaction steps as preceding decisions and their environment feedback compound distribution shift.

Alignment After Teacher Takeover. For each student-induced trajectory, we slide windows of size $w \in \{1, 2, 4, 8\}$ interaction steps along the trajectory. At each window start, the privileged teacher takes over generation for w consecutive interaction steps, conditioned on the preceding student-induced trajectory prefix. The student subsequently resumes from the environment state reached after these w teacher interactions and continues for another w steps using this extended interaction history. Using the step-wise reverse KL defined in Section 3.2, we compute teacher window KL as the mean over the w teacher-controlled steps and student continuation KL as the mean over the following w student-controlled steps. For each window size separately, we assign student continuations to the High and Low groups based on the upper and lower quartiles of their corresponding teacher window KL, respectively. As shown in Figure 1(b), the High group consistently exhibits lower mean student continuation KL than the Low group across all four window sizes, with the gap widening as the takeover window expands, indicating better subsequent alignment following teacher intervention in higher-divergence windows. Relative to the intervention start, these windows span future interactions, and the student's subsequent alignment reflects the potential value of intervening in them. These observations suggest that greater future teacher–student divergence may indicate higher utility of teacher supervision.

Sampled-Token Gradient Fidelity. To examine how faithfully sampled-token supervision approximates full-distribution gradients at different levels of teacher–student divergence, we conduct a toy experiment. Inspired by the simplified toy setting in EOPD (Jin et al., 2026), we remove autoregressive conditioning and represent each token prediction as categorical distributions over $V = 256$ candidates, corresponding to the candidate space in HiSD. We sample student logits from $z_i \sim \mathcal{N}(0, 1)$ and construct the student and teacher distributions as

$$p = \mathrm{softmax}(z), \qquad q_\beta = \mathrm{softmax}(z + \beta s),$$

where s is an independently sampled and standardized evidence vector that abstracts the token-wise effect of privileged information. Equivalently,

$$q_{\beta,i} = \frac{p_i \exp(\beta s_i)}{\sum_j p_j \exp(\beta s_j)},$$

showing that the teacher preserves the student's underlying policy belief while using privileged information to reweight candidate tokens, with β controlling the resulting distributional shift. We sample 200 student distributions and 10 evidence vectors for each student, and vary β over 25 logarithmically spaced values in $[0.02, 3.0]$, producing approximately 50,000 teacher–student pairs whose divergence is measured by forward KL.

Treating q_β as stop-gradient, full-distribution supervision and importance-corrected student-sampled token supervision induce

$$L_{\mathrm{full}} = D_{\mathrm{KL}}(q_\beta \| p), \qquad g_{\mathrm{full}} = p - q_\beta,$$

and

$$L_i = -\,\mathrm{sg}\left[\frac{q_{\beta,i}}{p_i}\right] \log p_i, \qquad g_i = \mathrm{sg}\left[\frac{q_{\beta,i}}{p_i}\right](p - e_i), \qquad i \sim p,$$

respectively, where e_i is the one-hot vector of the sampled token and the importance ratio is detached during differentiation. The importance-corrected sampled-token gradient is unbiased in expectation, as established in Proposition 1:

$$\mathbb{E}_{i\sim p}[g_i] = g_{\mathrm{full}}.$$

Nevertheless, an individual sampled-token update may still differ substantially from the full-distribution gradient. We therefore enumerate all V candidate tokens and compute their student-weighted directional and magnitude similarities:

$$S_{\mathrm{dir}} = \sum_{i=1}^{V} p_i \cos(g_i, g_{\mathrm{full}}), \qquad S_{\mathrm{mag}} = \sum_{i=1}^{V} p_i \frac{2\|g_i\|_2 \|g_{\mathrm{full}}\|_2}{\|g_i\|_2^2 + \|g_{\mathrm{full}}\|_2^2}.$$

Table 3: Training and hyperparameter configurations for HiSD.

Configuration	ALFWorld	Search-QA	WebShop
Training Configurations			
Training batch size	16	128	16
Validation batch size	64	512	64
Group size	8	8	8
Learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}
Warmup ratio	–	0.1	–
Maximum prompt length	2048	4096	4096
Maximum response length	512	512	512
KL regularization coefficient	0.01	0.001	0.01
Invalid-action penalty coefficient	0.1	0.01	0.1
Maximum interaction steps	50	4	15
Training steps	150	150	150
HiSD-specific Hyperparameters			
Base distillation coefficient λ_{base}	0.01	0.01	0.01
Divergence temperature κ	1.0	1.0	1.0
Discount factor γ	0.9	0.9	0.9
Distortion coefficient α	3.0	3.0	3.0
Teacher candidate support K	256	256	256

Higher values indicate a more faithful approximation of the direction and magnitude of full-distribution supervision. We divide the teacher–student pairs into 8 equal-frequency bins according to their forward KL and report the median direction and magnitude Similarities in Figure 1(c). It shows that when the teacher and student are closely aligned, $\|p - q_\beta\|_2$ becomes small while g_i remains dominated by a single sampled token, resulting in poor directional and magnitude agreement, with both similarities increase as KL divergence grows. We further provide theoretical support for this loss of gradient fidelity near alignment through the relative-error analysis in Corollary 1.

Teacher–Student Discrepancy Concentration. To examine how teacher–student discrepancy is distributed over the vocabulary at different levels of instantaneous divergence, we analyze the student and teacher distributions over the complete vocabulary of $V = 151{,}936$ tokens at each response position. We sort the vocabulary in descending order of teacher probability, with $q_{(j)}$ and $p_{(j)}$ denoting the teacher and student probabilities of the token at rank j, respectively. For each target discrepancy coverage $\eta \in \{0.3, 0.5, 0.9\}$, we define

$$m_\eta = \min\left\{ m : \frac{\sum_{j=1}^{m} \left| q_{(j)} - p_{(j)} \right|}{\sum_{j=1}^{V} \left| q_{(j)} - p_{(j)} \right|} \geq \eta \right\}, \qquad r_\eta = \frac{m_\eta}{V},$$

where the ratio measures the fraction of the full-vocabulary forward-KL gradient's L_1 norm covered by the top-m teacher-ranked tokens, as established in Proposition 5, and m_η and r_η represent the minimum number and ratio of teacher-ranked tokens required to reach the target gradient coverage η. We group response tokens according to their token-wise reverse KL divergence and report the required token ratios for 30%, 50%, and 90% discrepancy coverage in Figure 1(d). It suggests that teacher–student discrepancy becomes increasingly concentrated in the teacher's high-probability region as divergence grows, with fewer teacher-ranked tokens required to cover the same proportion of discrepancy. As the teacher and student become more closely aligned, the required token ratio increases substantially, revealing a more diffuse discrepancy that requires broader support to preserve fine-grained distributional information.

B.5 Training and Hyperparameter Configurations

Table 3 reports the detailed training configurations and the hyperparameters of HiSD.

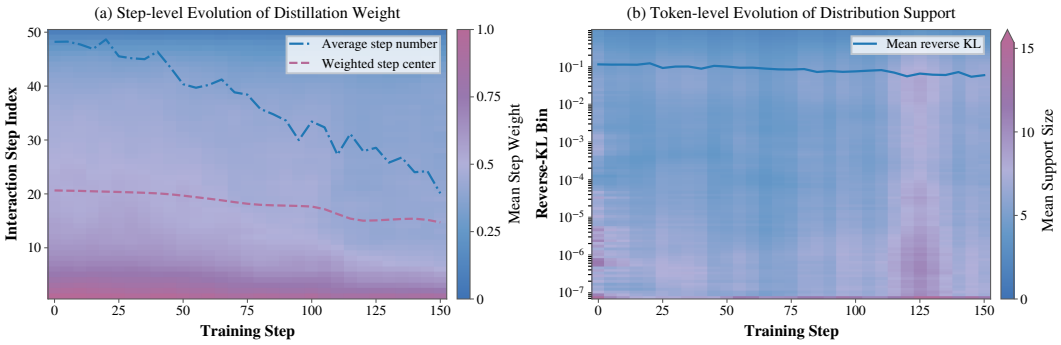


Figure 8: Evolution of HiSD's adaptive supervision during training. (a) Step-level distillation weight shifts with the evolving trajectory length during training. The blue dash-dotted curve denotes the average number of interaction steps per trajectory, while the red dashed curve denotes the distillation-weight center, i.e., the step index k at which $\lambda_k = 0.5$. (b) Token-level divergences are grouped into bins according to their reverse-KL values. The heatmap shows the mean adaptive support size within each divergence bin, and the blue curve denotes the mean reverse KL over training steps.

C Additional Experiments

C.1 Training Dynamics

Figure 8(a) shows that the average trajectory length decreases throughout training, accompanied by a slower decrease in the distillation-weight center. As a result, the weight center gradually moves toward later positions relative to the shortened trajectory, indicating that HiSD shifts its distillation focus from early interactions to more downstream decisions as training progresses. Meanwhile, Figure 8(b) illustrates that the average teacher–student divergence gradually decreases, while support size remains inversely adjusted to instantaneous divergence, enabling broader distributional supervision as the policy becomes more closely aligned with the teacher.

In Figure 9, HiSD maintains sustained improvements in training success rate across all three benchmarks. Compared with GRPO and SDAR, HiSD exhibits smoother entropy evolution and more stable gradient magnitudes, suggesting that adaptive distillation provides effective guidance while preserving stable policy optimization dynamics. The mean step weight follows distinct task-dependent trajectories, increasing during later training on ALFWorld, peaking early on Search-QA, and gradually decreasing on WebShop, indicating that step-level supervision dynamically adapts to the task-specific evolution of teacher–student discrepancy rather than following a predetermined schedule. The adaptive support generally contracts during early-to-middle training and expands toward the end, but operates at substantially different scales across benchmarks, reflecting the changing magnitude of teacher–student discrepancy within each task. These dynamics demonstrate that HiSD continuously adapts both the allocation and granularity of teacher supervision to evolving divergence patterns, maintaining a stable relative scale between reinforcement learning and distillation objectives throughout training.

C.2 Repetition Analysis

Definition of Repeated Exploration. For ALFWorld, each evaluated interaction step consists of an environment action and its returned observation. We employ DeepSeek-V4-Pro Xu et al. (2026) as an automatic annotator and provide it with the task instruction and the complete action–observation trajectory. The annotator jointly evaluates every interaction step against all preceding steps according to the action-specific criteria summarized in Table 4. Following the annotation prompt provided in Prompt 1, we conduct a single-pass trajectory-level annotation, where the complete trajectory is provided to the annotator at once instead of issuing separate queries for different step prefixes. The resulting output contains one record per interaction step, including a binary repetition label. For positive cases, we additionally collect the repetition category, the most relevant preceding step index, and a concise justification to facilitate traceability and verification of the annotations. For Search-QA, we identify repeated exploration using a heuristic rule rather than an LLM annotator.

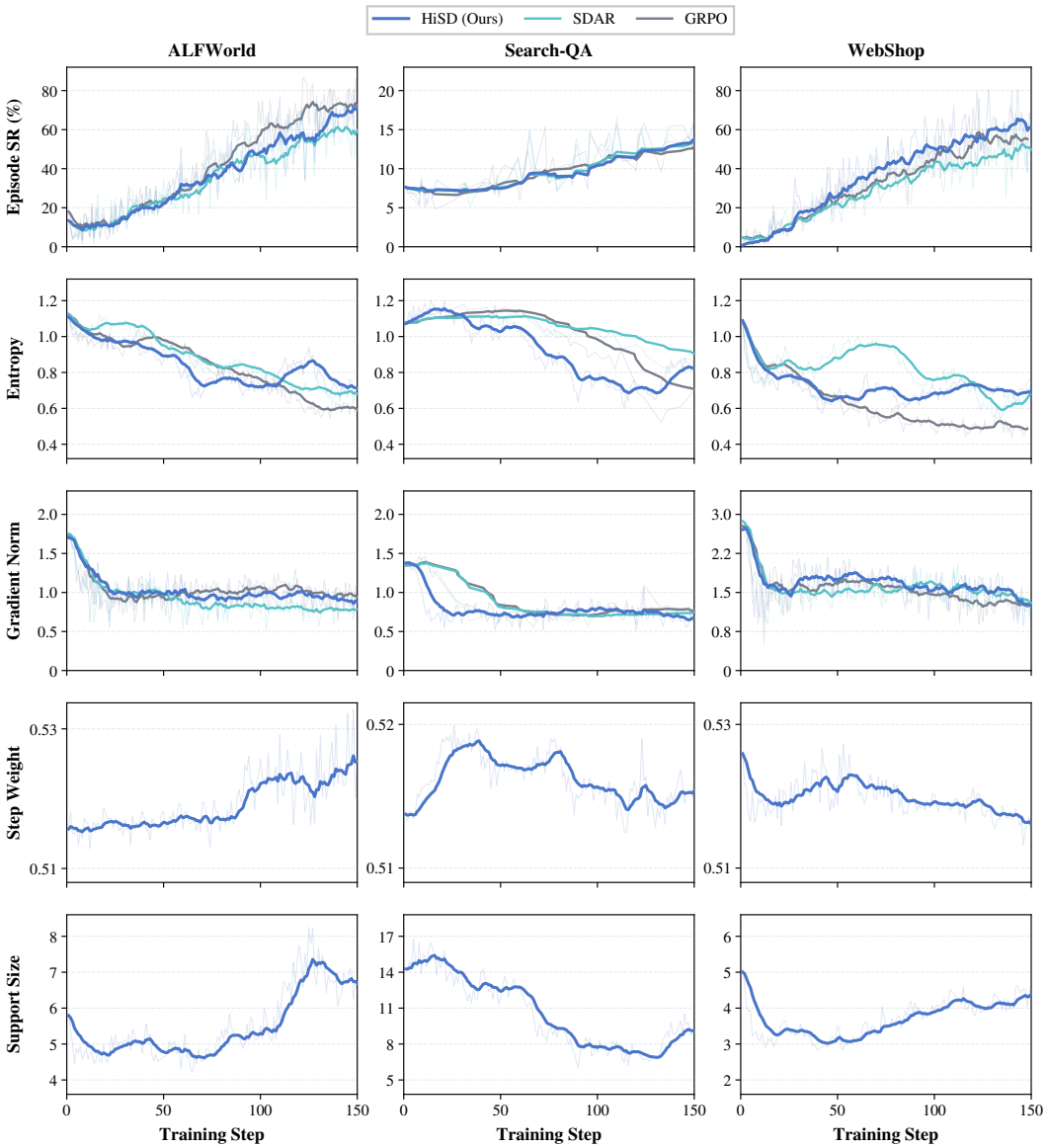


Figure 9: Training dynamics of HiSD, SDAR, and GRPO on ALFWorld, Search-QA, and WebShop using Qwen2.5-1.5B-Instruct.

We evaluate each retrieval step, comprising a retrieval query and its corresponding top-3 passages, and label it as repeated if any preceding retrieval step returns the same top-3 passages.

Human Verification. To assess the reliability of the automatic annotations, we construct a stratified human verification set from ALFWorld evaluation trajectories generated by Qwen2.5-7B-Instruct. Specifically, we randomly sample two successful and two failed trajectories from each of the six task categories, resulting in 24 trajectories for manual verification. Three human annotators independently label whether each interaction step exhibits repeated exploration according to the criteria in Table 4, without access to the automatic annotations. We determine the human reference label for each step by majority voting over the three binary labels. We then compare the DeepSeek-V4-Pro annotations against these human reference labels across all annotated steps, obtaining an F1 score of 92.66% with repeated exploration as the positive class.

Table 4: Criteria for repeated exploration in ALFWorld.

Category and Actions	Criterion	Examples
Repeated State Check `look`, `inventory`, `examine`	Rechecks the state of the same target while the previous observation remains valid after intervening actions or environmental changes.	*Rep:* Executing `inventory` twice without any intervening object transfer. *Non-rep:* Executing `inventory` after taking an object.
Repeated Navigation `go to`	Revisits the same location to repeat prior interactions, although the information or outcomes obtained there remain valid after intervening actions or environmental changes.	*Rep:* Returning to a cabinet already confirmed not to contain the target object without any relevant intervening change. *Non-rep:* Returning to collect a second target object previously found there in a Pick-Two task.
Repeated Receptacle Operation `open`, `close`	Repeats an operation or an open–close cycle on the same receptacle without intervening actions or environmental changes creating a need to repeat it.	*Rep:* Reopening a previously inspected drawer to check its contents again, although the contents have not changed. *Non-rep:* Opening a previously uninspected drawer after unsuccessful searches of other drawers.
Repeated Object Transfer `take`, `move`	Repeats the same object transfer while its previous successful outcome still holds, or retries it without resolving the conditions that caused its failure.	*Rep:* Retrying the same `take` action while the agent still holds another object. *Non-rep:* Retrying the action after moving the held object away and freeing the inventory.
Repeated Property Operation `clean`, `heat`, `cool`	Repeats the same property operation on the same object while its previous effect persists, or retries it without resolving the conditions that caused its failure.	*Rep:* Heating the same object again after a successful heating operation, without subsequent cooling. *Non-rep:* Heating the object for the first time after retrieving it.

Measurement of Repeated Exploration. Given M evaluated trajectories $\{\tau_i\}_{i=1}^{M}$, where N_i is the number of interaction steps in τ_i and $\chi_{i,k}^{\text{rep}} \in \{0,1\}$ indicates whether its k-th step is repeated, the trajectory-level repetition rate R_{traj} and step-level repetition rate R_{step} are defined as

$$R_{\text{traj}} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}\left[\sum_{k=1}^{N_i} \chi_{i,k}^{\text{rep}} > 0\right], \qquad R_{\text{step}} = \frac{\sum_{i=1}^{M}\sum_{k=1}^{N_i} \chi_{i,k}^{\text{rep}}}{\sum_{i=1}^{M} N_i}.$$

Repeated Exploration across Task Difficulty. For ALFWorld, we group instances by the context length accumulated by ReAct, while for Search-QA, TriviaQA and PopQA constitute the single-hop group and 2Wiki, MuSiQue, and Bamboogle constitute the multi-hop group. Figure 10 shows that both repetition rates increase with context length, with the three methods performing similarly on the easiest instances and HiSD consistently exhibiting less repetition as task difficulty increases. Most notably, in the 1.0k–1.5k interval, HiSD reduces the proportion of trajectories containing repeated exploration from 44.4% under either single-level variant to 22.2%, with a similarly pronounced reduction at the step level. Figure 11 further shows that repeated retrieval is particularly prevalent on 2Wiki and MuSiQue, while HiSD consistently achieves the lowest trajectory- and step-level repetition rates across all five QA benchmarks. These results demonstrate that combining step- and token-level adaptation more consistently suppresses redundant exploration and facilitates more efficient task completion.

Repeated Exploration across Task Difficulty. For ALFWorld, we group instances by the context length accumulated by ReAct, while for Search-QA, TriviaQA and PopQA constitute the single-hop group and 2Wiki, MuSiQue, and Bamboogle constitute the multi-hop group. Figure 10 shows that both repetition rates increase with context length, with the three methods performing similarly on the easiest instances and HiSD consistently exhibiting less repetition as task difficulty increases.

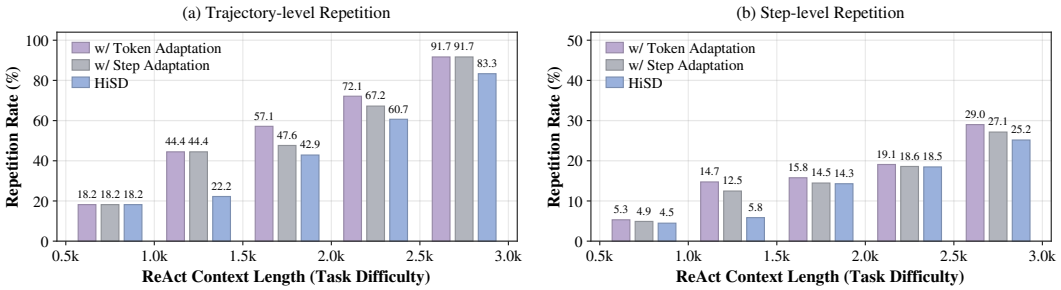


Figure 10: Repeated exploration across ALFWorld task difficulty (grouped by ReAct context length).

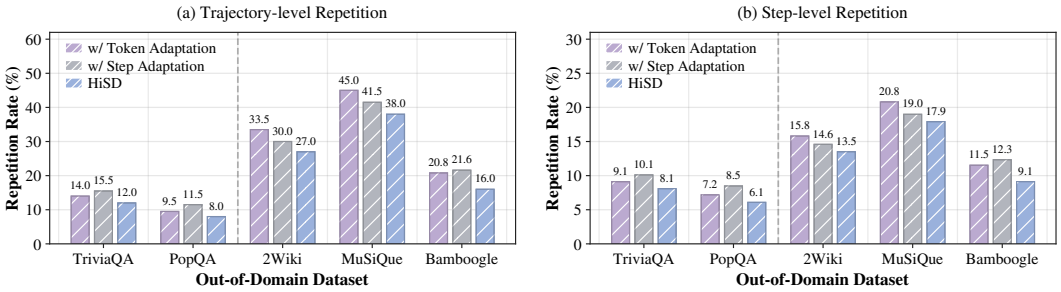


Figure 11: Repeated exploration across out-of-domain Search-QA datasets.

Figure 11 further shows that on single-hop datasets (TriviaQA and PopQA), token-level adaptation consistently achieves lower repetition rates than step-level adaptation. Since these tasks typically require only a few retrieval interactions to identify a single factual relation, redundant exploration is less affected by accumulated interaction-level deviations. In such cases, repeated exploration in these tasks often reflects attempts to reformulate the same retrieval intent with different lexical expressions, such as replacing entity mentions or query phrases to improve retrieval matching. Since these cases are dominated by query-level retrieval mismatch rather than trajectory-level state deviation, token-level supervision provides fine-grained guidance over the local generation distribution. In contrast, step-level adaptation becomes more beneficial on multi-hop datasets (2Wiki and MuSiQue), where the dependency across retrieval steps results in deviations to propagate along the trajectory.

Interestingly, Bamboogle exhibits an opposite trend, where token-level adaptation achieves lower repetition rates than step-level adaptation despite its multi-hop nature. Unlike 2Wiki and MuSiQue, where intermediate entities or answers obtained from earlier retrieval steps determine the subsequent search targets, Bamboogle mainly requires identifying and combining relevant facts that can be retrieved more independently. Therefore, its retrieval process is less sensitive to errors propagated across interaction steps, and redundant exploration is more likely to arise from local query decisions, where token-level adaptation provides more effective guidance.

D Case Study

As shown in Figure 13, HiSD adjusts distillation according to the remaining teacher–student discrepancy along the trajectory. When the student enter the correct route, teacher supervision quickly weakens to avoid unnecessary or even harmful late-stage guidance. In contrast, when the student makes a critical mistake while the teacher remains correct, HiSD keeps strong supervision in the subsequent steps to preserve sufficient corrective signal. At the token level, large supports appear mainly in reasoning spans, especially in the student-failure/teacher-success case, indicating that HiSD allocates finer-grained supervision to reasoning-related tokens. This observation is further consistent with the word-cloud analysis in This observation is further consistent with the word-cloud analysis in Figure 12, where high-support words are dominated by reasoning and discourse expressions (e.g., “therefore” and “further”), whereas low-support words are mainly associated with entities and actions (e.g., “knife” and “put”).

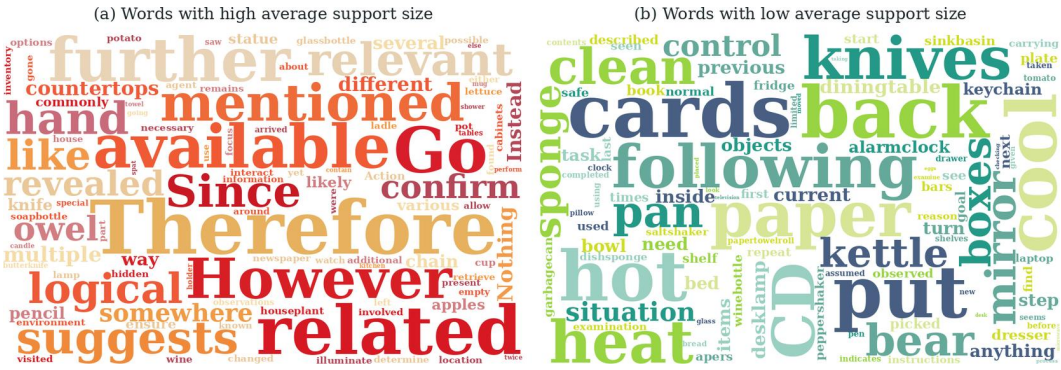


Figure 12: Word-cloud analysis of adaptive token support on ALFWorld. In the high-support panel, larger words indicate larger average support sizes, whereas in the low-support panel, larger words indicate smaller average support sizes.

E Algorithm

Algorithm 1 summarizes the training procedure of HiSD. Relative to GRPO, HiSD retains its on-policy rollout and group-relative policy optimization backbone, while adding a privileged self-teacher forward pass with adaptive distillation through progress-aware step supervision and support-adaptive token supervision. The training procedures of the four OPSD-based baselines, namely SSPO, RLSD, Skill-SD, and SDAR, are presented in Algorithms 2, 3, 4, and 5, respectively.

F Prompts

We provide the interaction prompts for ALFWorld, Search-QA, and WebShop in Prompts 2, 3, and 4, respectively. The student and privileged teacher use the same interaction context and output format, while the teacher additionally receives task-relevant skill context as privileged information. At inference time, the agent operates without access to privileged information.



Figure 13: Case study of HiSD on ALFWorld. The left and right panels show critical steps from successful and failed student trajectories, respectively, on samples where the teacher exhibits the opposite outcome.

Algorithm 1 HiSD: Hierarchical Self-Distillation for Agentic Reinforcement Learning

Require: policy π_θ, environment $\mathcal{E}$, group size G; privileged information r; candidate size K, decay γ, temperature κ, distortion coefficient α, distillation coefficient λ_{base}

for each training iteration **do**
  sample a batch of tasks $\{x\}$
  for each task x **do**
    // Step 1: GRPO optimization
    sample G trajectories $\{\tau^{(1)}, \dots, \tau^{(G)}\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ ▷ on-policy rollout
    obtain rewards $\{R^{(i)}\}_{i=1}^{G}$ and advantages $\{\widehat{A}^{(i)}\}_{i=1}^{G}$
    compute the clipped GRPO loss $\mathcal{L}_{\text{GRPO}}$
    // Step 2: Hierarchical self-distillation
    for $i = 1, \dots, G$ **do**
      for $k = 1, \dots, N_i$ **do**
        for $t = 1, \dots, T_{i,k}$ **do**
          $c_{i,k,t} \leftarrow (H_{i,k}, y_{i,k,<t})$
          $\pi_T(\cdot \mid c_{i,k,t}, r) \leftarrow \text{sg}[\pi_\theta(\cdot \mid c_{i,k,t}, r)]$ ▷ privileged teacher
          $\mathcal{C}_{i,k,t} \leftarrow \text{TopK}_K(\pi_T(\cdot \mid c_{i,k,t}, r))$
          normalize π_T and π_θ over $\mathcal{C}_{i,k,t}$ to obtain $\bar{\pi}_T$ and $\bar{\pi}_\theta$
          $\delta^{\leftarrow}_{i,k,t} \leftarrow \text{sg}[D_{\text{KL}}(\bar{\pi}_\theta \,\|\, \bar{\pi}_T)]$
        end for
        $d_{i,k} \leftarrow \frac{1}{T_{i,k}} \sum_{t=1}^{T_{i,k}} \delta^{\leftarrow}_{i,k,t}$
      end for
      // Step 3: Progress-aware step supervision
      for $k = 1, \dots, N_i$ **do**
        $D^{\text{pre}}_{i,k} \leftarrow \sum_{j=1}^{k-1} \gamma^{k-1-j} d_{i,j}$
        $D^{\text{suf}}_{i,k} \leftarrow \sum_{j=k}^{N_i} \gamma^{j-k} d_{i,j}$
        $\lambda_{i,k} \leftarrow \frac{\exp(-D^{\text{pre}}_{i,k}/\kappa)}{\exp(-D^{\text{pre}}_{i,k}/\kappa)+\exp(-D^{\text{suf}}_{i,k}/\kappa)}$ ▷ progress-aware step weight
      end for
      // Step 4: Support-adaptive token supervision
      for $k = 1, \dots, N_i$ **do**
        for $t = 1, \dots, T_{i,k}$ **do**
          $\eta_{i,k,t} \leftarrow \exp(-\alpha \delta^{\leftarrow}_{i,k,t})$
          sort $\mathcal{C}_{i,k,t}$ by $\bar{\pi}_T$ in descending order as $\{v^{(1)}, \dots, v^{(K)}\}$
          $m_{i,k,t} \leftarrow \min\left\{m : \sum_{j=1}^{m} \bar{\pi}_T(v^{(j)}) \geq \eta_{i,k,t}\right\}$
          $\mathcal{S}_{i,k,t} \leftarrow \{v^{(1)}, \dots, v^{(m_{i,k,t})}\}$ ▷ adaptive support
          renormalize $\bar{\pi}_T$ and $\bar{\pi}_\theta$ over $\mathcal{S}_{i,k,t}$ to obtain $\widetilde{\pi}_T$ and $\widetilde{\pi}_\theta$
          $\widetilde{\delta}^{\rightarrow}_{i,k,t} \leftarrow D_{\text{KL}}(\widetilde{\pi}_T \,\|\, \widetilde{\pi}_\theta)$
        end for
      end for
      $\mathcal{L}^{(i)}_{\text{HiSD}} \leftarrow \frac{1}{N_i} \sum_{k=1}^{N_i} \lambda_{i,k} \left(\frac{1}{T_{i,k}} \sum_{t=1}^{T_{i,k}} \widetilde{\delta}^{\rightarrow}_{i,k,t}\right)$
    end for
  end for
  $\mathcal{L}_{\text{HiSD}} \leftarrow \text{mean}_{x,i}\left[\mathcal{L}^{(i)}_{\text{HiSD}}\right]$
  // Step 5: Joint policy update
  update θ by minimizing $\mathcal{L}_{\text{GRPO}} + \lambda_{\text{base}} \mathcal{L}_{\text{HiSD}}$ ▷ policy update
end for

Algorithm 2 SSPO: Step-Level Self-Distilled Policy Optimization

Require: policy π_θ, environment $\mathcal{E}$, group size G; privileged information r, weight cap ϵ_w

for each training iteration **do**
  sample a batch of tasks $\{x\}$
  for each task x **do**
    // Step 1: On-policy rollout and reward estimation
    sample G trajectories $\{\tau^{(1)}, \dots, \tau^{(G)}\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ ▷ on-policy rollout
    obtain rewards $\{R^{(i)}\}_{i=1}^{G}$, success indicators $\{z^{(i)}\}_{i=1}^{G}$, and advantages $\{\widehat{A}^{(i)}\}_{i=1}^{G}$
    // Step 2: Step-level privileged-information gain
    for $i = 1, \dots, G$ **do**
      for $k = 1, \dots, N_i$ **do**
        if $z^{(i)} = 0$ **then**
          for $t = 1, \dots, T_{i,k}$ **do**
            $c_{i,k,t} \leftarrow (H_{i,k}, y_{i,k,<t})$
            $\pi_T(\cdot \mid c_{i,k,t}, r) \leftarrow \text{sg}[\pi_\theta(\cdot \mid c_{i,k,t}, r)]$ ▷ privileged teacher
          end for
          $\Delta_{i,k}^{\text{step}} \leftarrow \text{sg}\Big[\sum_{t=1}^{T_{i,k}} \left(\log \pi_T(y_{i,k,t} \mid c_{i,k,t}, r) - \log \pi_{\theta_{\text{old}}}(y_{i,k,t} \mid c_{i,k,t})\right)\Big]$
          $w_{i,k} \leftarrow \min\Big\{\exp\Big(\text{sign}(\widehat{A}^{(i)})\Delta_{i,k}^{\text{step}}\Big), 1 + \epsilon_w\Big\}$
        else
          $w_{i,k} \leftarrow 1$ ▷ leave successful trajectories unchanged
        end if
        for $t = 1, \dots, T_{i,k}$ **do**
          $\widetilde{A}_{i,k,t} \leftarrow w_{i,k}\widehat{A}^{(i)}$
        end for
      end for
    end for
    // Step 3: Step-reweighted policy optimization
    compute the clipped SSPO loss $\mathcal{L}_{\text{SSPO}}(x)$ using $\{\widetilde{A}_{i,k,t}\}$
  end for
  $\mathcal{L}_{\text{SSPO}} \leftarrow \text{mean}_x[\mathcal{L}_{\text{SSPO}}(x)]$
  // Step 4: Policy update
  update θ by minimizing $\mathcal{L}_{\text{SSPO}}$ ▷ policy update
end for

Algorithm 3 RLSD: RLVR with Self-Distillation

Require: policy π_θ, environment $\mathcal{E}$, group size G; privileged information r, mixing coefficient λ, weight clip ϵ_w

for each training iteration **do**
  sample a batch of tasks $\{x\}$
  for each task x **do**
    // Step 1: On-policy rollout and reward estimation
    sample G trajectories $\{\tau^{(1)}, \dots, \tau^{(G)}\} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ ▷ on-policy rollout
    obtain rewards $\{R^{(i)}\}_{i=1}^{G}$ and advantages $\{\widehat{A}^{(i)}\}_{i=1}^{G}$
    // Step 2: Token-level self-distilled reweighting
    for $i = 1, \dots, G$ **do**
      for $k = 1, \dots, N_i$ **do**
        for $t = 1, \dots, T_{i,k}$ **do**
          $c_{i,k,t} \leftarrow (H_{i,k}, y_{i,k,<t})$
          $\pi_T(\cdot \mid c_{i,k,t}, r) \leftarrow \text{sg}[\pi_\theta(\cdot \mid c_{i,k,t}, r)]$ ▷ privileged teacher
          $\Delta_{i,k,t} \leftarrow \text{sg}[\log \pi_T(y_{i,k,t} \mid c_{i,k,t}, r) - \log \pi_{\theta_{\text{old}}}(y_{i,k,t} \mid c_{i,k,t})]$
          $w_{i,k,t} \leftarrow \text{clip}\Big(\exp\Big(\text{sign}(\widehat{A}^{(i)})\Delta_{i,k,t}\Big), 1-\epsilon_w, 1+\epsilon_w\Big)$
          $\widetilde{A}_{i,k,t} \leftarrow \widehat{A}^{(i)}\left[(1-\lambda) + \lambda w_{i,k,t}\right]$ ▷ direction-preserving credit
        end for
      end for
    end for
    // Step 3: Token-reweighted policy optimization
    compute the clipped RLSD loss $\mathcal{L}_{\text{RLSD}}(x)$ using $\{\widetilde{A}_{i,k,t}\}$
  end for
  $\mathcal{L}_{\text{RLSD}} \leftarrow \text{mean}_x[\mathcal{L}_{\text{RLSD}}(x)]$
  // Step 4: Policy update
  update θ by minimizing $\mathcal{L}_{\text{RLSD}}$ ▷ policy update
end for

Algorithm 4 Skill-SD: Skill-Conditioned Self-Distillation

Require: policy π_θ, environment $\mathcal{E}$, group size G; privileged skill r, distillation coefficient λ_{SD}

for each training iteration **do**
  sample a batch of tasks $\{x\}$
  for each task x **do**
    // Step 1: GRPO optimization
    sample G trajectories $\{\tau^{(1)}, \dots, \tau^{(G)}\} \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid x)$ ▷ on-policy rollout
    obtain rewards $\{R^{(i)}\}_{i=1}^{G}$ and advantages $\{\widehat{A}^{(i)}\}_{i=1}^{G}$
    compute the clipped GRPO loss $\mathcal{L}_{\mathrm{GRPO}}(x)$
    // Step 2: Skill-conditioned self-distillation
    for $i = 1, \dots, G$ **do**
      $M_i \leftarrow \sum_{k=1}^{N_i} T_{i,k}$
      for $k = 1, \dots, N_i$ **do**
        for $t = 1, \dots, T_{i,k}$ **do**
          $c_{i,k,t} \leftarrow (H_{i,k}, y_{i,k,<t})$
          $\pi_T(\cdot \mid c_{i,k,t}, r) \leftarrow \mathrm{sg}[\pi_\theta(\cdot \mid c_{i,k,t}, r)]$ ▷ skill-conditioned teacher
          $d_{i,k,t} \leftarrow \log \pi_\theta(y_{i,k,t} \mid c_{i,k,t}) - \log \pi_T(y_{i,k,t} \mid c_{i,k,t}, r)$
          $q^{(3)}_{i,k,t} \leftarrow \exp(-d_{i,k,t}) - 1 + d_{i,k,t}$ ▷ K_3 estimator
          $\rho^{\mathrm{SD}}_{i,k,t} \leftarrow \exp(\log \pi_\theta(y_{i,k,t} \mid c_{i,k,t}) - \mathrm{sg}[\log \pi_{\theta_{\mathrm{old}}}(y_{i,k,t} \mid c_{i,k,t})])$
          $\ell^{\mathrm{SD}}_{i,k,t} \leftarrow \rho^{\mathrm{SD}}_{i,k,t} q^{(3)}_{i,k,t}$
        end for
      end for
      $\mathcal{L}^{(i)}_{\text{Skill-SD}} \leftarrow \frac{1}{M_i} \sum_{k=1}^{N_i} \sum_{t=1}^{T_{i,k}} \ell^{\mathrm{SD}}_{i,k,t}$
    end for
  end for
  $\mathcal{L}_{\mathrm{GRPO}} \leftarrow \mathrm{mean}_x[\mathcal{L}_{\mathrm{GRPO}}(x)]$
  $\mathcal{L}_{\text{Skill-SD}} \leftarrow \mathrm{mean}_{x,i}\left[\mathcal{L}^{(i)}_{\text{Skill-SD}}\right]$
  // Step 3: Joint policy update
  update θ by minimizing $\mathcal{L}_{\mathrm{GRPO}} + \lambda_{\mathrm{SD}} \mathcal{L}_{\text{Skill-SD}}$ ▷ policy update
end for

Algorithm 5 SDAR: Self-Distilled Agentic Reinforcement Learning

Require: policy π_θ, environment $\mathcal{E}$, group size G; privileged information r, gate sharpness β, distillation coefficient λ_{SDAR}

for each training iteration **do**
  sample a batch of tasks $\{x\}$
  for each task x **do**
    // Step 1: GRPO optimization
    sample G trajectories $\{\tau^{(1)}, \ldots, \tau^{(G)}\} \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid x)$ ▷ on-policy rollout
    obtain rewards $\{R^{(i)}\}_{i=1}^{G}$ and advantages $\{\widehat{A}^{(i)}\}_{i=1}^{G}$
    compute the clipped GRPO loss $\mathcal{L}_{\mathrm{GRPO}}(x)$
    // Step 2: Token-level gated self-distillation
    for $i = 1, \ldots, G$ **do**
      $M_i \leftarrow \sum_{k=1}^{N_i} T_{i,k}$
      for $k = 1, \ldots, N_i$ **do**
        for $t = 1, \ldots, T_{i,k}$ **do**
          $c_{i,k,t} \leftarrow (H_{i,k}, y_{i,k,<t})$
          $\pi_T(\cdot \mid c_{i,k,t}, r) \leftarrow \mathrm{sg}[\pi_\theta(\cdot \mid c_{i,k,t}, r)]$ ▷ privileged teacher
          $\Delta_{i,k,t} \leftarrow \mathrm{sg}[\log \pi_T(y_{i,k,t} \mid c_{i,k,t}, r) - \log \pi_\theta(y_{i,k,t} \mid c_{i,k,t})]$
          $g_{i,k,t} \leftarrow \sigma(\beta \Delta_{i,k,t})$ ▷ distillation gate
          $\ell_{i,k,t}^{\mathrm{SDAR}} \leftarrow g_{i,k,t} \left(\log \pi_T(y_{i,k,t} \mid c_{i,k,t}, r) - \log \pi_\theta(y_{i,k,t} \mid c_{i,k,t})\right)$
        end for
      end for
      $\mathcal{L}_{\mathrm{SDAR}}^{(i)} \leftarrow \frac{1}{M_i} \sum_{k=1}^{N_i} \sum_{t=1}^{T_{i,k}} \ell_{i,k,t}^{\mathrm{SDAR}}$
    end for
  end for
  $\mathcal{L}_{\mathrm{GRPO}} \leftarrow \mathrm{mean}_x[\mathcal{L}_{\mathrm{GRPO}}(x)]$
  $\mathcal{L}_{\mathrm{SDAR}} \leftarrow \mathrm{mean}_{x,i}\left[\mathcal{L}_{\mathrm{SDAR}}^{(i)}\right]$
  // Step 3: Joint policy update
  update θ by minimizing $\mathcal{L}_{\mathrm{GRPO}} + \lambda_{\mathrm{SDAR}} \mathcal{L}_{\mathrm{SDAR}}$ ▷ policy update
end for

Prompt 1: Repeated Exploration Annotation Prompt for ALFWorld

You are an expert evaluator for identifying repeated exploration behaviors in ALFWorld agent trajectories.

Given a task instruction and a complete trajectory consisting of interleaved `<action>` and `<environment>` interactions:

{instruction}

{trajectory}

Your goal is to annotate whether each interaction step contains repeated exploration. The evaluation should be performed over the entire trajectory in a single pass. For each step, compare the current action and its resulting observation against all preceding steps to determine whether the action repeats previously explored behavior.

The criteria for repeated exploration are summarized as follows:

{criteria}

Return the annotation results strictly in JSON format. The output should contain one record for each interaction step in the original trajectory order:

```
{
  "step_id":  <integer>,
  "repeated":  <true/false>,
  "category":  <string or null>,
  "preceding_step":  <integer or null>,
  "justification":  <string or null>
}
```

For non-repeated steps, set *category*, *preceding_step*, and *justification* to null. For repeated steps, provide the corresponding repetition category, the most relevant preceding interaction step index, and a concise justification to ensure the annotation is traceable and verifiable.

Prompt 2: Interaction Prompt for ALFWorld

You are an expert agent operating in the ALFRED Embodied Environment. Your task is to: {task_description}

{skill_context}

Prior to this step, you have already taken {step_count} step(s). Below are the most recent {history_length} observations and the corresponding actions you took: {action_history}

You are now at step {current_step} and your current observation is: {current_observation}

Your admissible actions of the current situation are: [{admissible_actions}].

Now it's your turn to take an action.

You should first reason step-by-step about the current situation. This reasoning process MUST be enclosed within `<think>` `</think>` tags.

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

Prompt 3: Interaction Prompt for Search-QA

You are an expert agent tasked with answering the given question step-by-step.

{skill_context}

Your question: {task_description}

Prior to this step, you have already taken {step_count} step(s). Below is the interaction history where <search> </search> wrapped your past search queries and <information> </information> wrapped the corresponding search results returned by the external search engine. History:

{memory_context}

Now it's your turn to respond for the current step.

You should first conduct reasoning process. This process MUST be enclosed within <think> </think> tags.

After completing your reasoning, choose only one of the following actions (do not perform both):

(1) If you find you lack some knowledge, you can call a search engine to get more external information using format: <search> your query </search>.

(2) If you have enough knowledge to answer the question confidently, provide your final answer within <answer> </answer> tags, without detailed illustrations. For example, <answer> Beijing </answer>.

Prompt 4: Interaction Prompt for WebShop

You are an expert autonomous agent operating in the WebShop e-commerce environment.

{skill_context}

Your task is to: {task_description}.

Prior to this step, you have already taken {step_count} step(s). Below are the most recent {history_length} observations and the corresponding actions you took: {action_history}

You are now at step {current_step} and your current observation is: {current_observation}.

Your admissible actions of the current situation are: [{admissible_actions}].

Now it's your turn to take one action for the current step.

You should first reason step-by-step about the current situation, then think carefully which admissible action best advances the shopping goal. This reasoning process MUST be enclosed within <think> </think> tags.

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