arXiv:2608.08960v1 [cs.AI] 9 Aug 2026

Reading is not Reasoning: Bridging the Agentic Policy Gap in Vision–Text Compression

Cheng Fan[1], Junyi Zhou[1], Tingzhang Luo[1], RongJian Xu[1], Qiyanhui Lu[1], Mingjian Zhu[2], Hanting Chen[2], Jianyuan Guo[1]*

[1]City University of Hong Kong
[2]Huawei Technologies Ltd.
chengfan2-c@my.cityu.edu.hk, jianyguo@cityu.edu.hk

Abstract

Multi-step language-model agents repeatedly process growing interaction histories, leading to substantial context costs. Vision–text compression reduces these costs by rendering history as images, but the resulting modality shift creates a marked capability gap. Through controlled evaluations of history recovery, matched-state decisions, and complete trajectories, we show that this gap cannot be explained by OCR quality alone. Visual-history agents exhibit systematic drift in action selection, query formulation, stopping, and evidence use, revealing an agentic policy gap. We introduce **CAPS**, a two-stage **C**ross-modal **A**gentic **P**olicy **S**elf-distillation framework that uses the same model's stronger text-history policy to supervise its visual-history counterpart. Offline trajectory self-distillation transfers successful text-policy behavior to visual-history inputs, while online policy self-distillation provides dense supervision on states visited by the visual-history policy during reinforcement learning. On SearchQA, CAPS improves over AgentOCR by 5.0% and 3.4% with 3B and 7B backbones, respectively. On full-history ALFWorld, the corresponding gains are 15.6% and 14.5%. Across settings, CAPS reduces average memory-context cost by up to 63.3% and peak cost by up to 83.4% relative to matched text-history policies. These results show that explicit cross-modal policy self-distillation can preserve agent capability under vision–text compression. Our code will be made publicly available in a future release.

Introduction

"Reading furnishes the mind only with materials of knowledge; it is thinking that makes what we read ours."

— John Locke, *Of the Conduct of the Understanding*

Large language model agents solve complex tasks by interleaving reasoning and action over multiple rounds of interaction with an environment (Nakano et al. 2021; Yao et al. 2022; Schick et al. 2023). Because each decision is conditioned on preceding interactions, new observations and actions are typically appended to the context. As tasks require more retrieval, planning, and tool-use steps, repeatedly processing this growing history incurs increasing token, memory, and latency costs (Kang et al. 2025; Lu et al. 2026a; Chhikara et al. 2025; Xu et al. 2026; Zhou et al. 2025). Practical multi-step agents therefore need a compact history representation that preserves the information required for future decisions.

*Corresponding author.

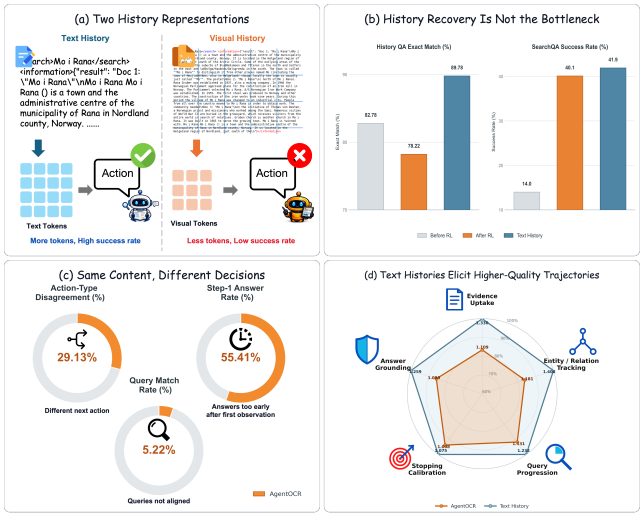

Figure 1: **Textual-history vs. visual-history policy. (a)** Textual and visual history representations under the same interaction trajectory. **(b)** Task improvement is not explained by improved history recovery. **(c)** Visual-history policies make different decisions from text-history policies under matched states. **(d)** Text-history policies produce higher-quality trajectories across multiple dimensions.

Vision–text compression (VTC) offers a promising approach to this trade-off. Rather than discarding or summarizing content, VTC renders text into information-dense images that represent the same content with substantially fewer visual tokens (Xing et al. 2026; Wei, Sun, and Li 2025; Cheng et al. 2026; Wang et al. 2024). AgentOCR (Feng et al. 2026b) extends this paradigm to multi-step agents by rendering accumulated observations and actions as visual history and training the agent to act directly from the compressed interface. However, this modality shift incurs a substantial capability loss: despite receiving the same interaction content, the visual-history policy remains markedly weaker than its text-history counterpart, as shown in Fig. 1 (a). Context efficiency, therefore, does not necessarily imply capability preservation. This observation raises a fundamental question: *does this gap arise because the agent cannot reliably read the compressed history, or because it fails to reason and act on what it reads?*

To answer this question, we analyze the gap progressively, from history recovery, through individual decisions, to com-

plete trajectories (detailed in Sec.). On History QA constructed from real trajectories, reinforcement learning greatly improves task success without improving recovery accuracy, indicating that its gains do not come from stronger OCR. Under matched semantic states, changing only the history modality alters action types, search queries, and stopping decisions. At the trajectory level, the visual-history policy also exhibits weaker evidence use, query progression, stopping calibration, and answer grounding. Together, these results reveal an *agentic policy gap* rather than a purely perceptual gap: visual-history agents may read the relevant content yet still fail to reason and act effectively upon it.

Viewing the problem as a policy gap suggests a direct source of supervision: the agent's stronger text-history policy. Each interaction state provides paired contexts containing the same observations and actions, differing only in whether the history is represented as text or a rendered image. Because the two policies share the same base model and action space, this correspondence enables direct cross-modal behavior transfer. We therefore introduce **CAPS**, a two-stage **C**ross-modal **A**gentic **P**olicy **S**elf-distillation framework that transfers reasoning and decision behavior from text-history to visual-history policies.

CAPS addresses two complementary state distributions. In the first stage, *offline trajectory self-distillation* collects successful rollouts from the text-history policy, renders each history prefix into its visual form, and trains the visual-history policy to reproduce the corresponding complete response, including its reasoning and action. This stage transfers successful behavior, but its supervision is restricted to states visited by the text-history policy. The second stage therefore performs *online policy self-distillation* via reinforcement learning. On states induced by the visual-history policy's own actions, a frozen text-history policy receives the paired textual history and provides full-vocabulary supervision on the student-generated response prefixes. CAPS thereby extends cross-modal transfer from successful teacher trajectories to the target policy's own interaction distribution, while retaining environmental reward as the primary learning signal.

Experiments on SearchQA and ALFWorld demonstrate that CAPS improves the capability–efficiency trade-off across both search and embodied agent tasks. On SearchQA, CAPS improves over AgentOCR by 5.0% and 3.4% with 3B and 7B backbones, respectively. On full-history ALFWorld, the corresponding gains are 15.6% and 14.5%. These improvements are achieved while substantially reducing context cost relative to matched text-history policies. Further diagnostics show that CAPS increases action and query agreement with the text-history policy and improves trajectory quality across all evaluated dimensions.

Our contributions are threefold. **(i)** We provide a controlled diagnosis of the capability loss induced by vision–text compression. The results identify an agentic policy gap that cannot be explained by OCR quality alone. **(ii)** We propose CAPS, a two-stage cross-modal policy self-distillation framework that transfers text-history behavior to visual-history policies through offline and online distillation. **(iii)** Extensive evaluations demonstrate that our propsoed CAPS improves the capability–efficiency trade-off for LLM agents.

Related Work

Multi-step agents. LLM agents solve open-ended tasks through repeated reasoning, action, and environment feedback across software, embodied, and search domains (Jimenez et al. 2024; Wang et al. 2023; Shridhar et al. 2020; Jin et al. 2025). WebGPT and ReAct established prompting and tool-use paradigms (Nakano et al. 2021; Yao et al. 2022), while later methods directly optimize agent policies with reinforcement learning (Rafailov et al. 2023; Sheng et al. 2024; Wang et al. 2025). GRPO uses critic-free group-relative updates, and GiGPO adds episode- and step-level credit assignment for long-horizon tasks (Shao et al. 2024; Feng et al. 2026a). To control growing interaction contexts, ACON and LongSeeker use learned compression or orchestration (Kang et al. 2025; Lu et al. 2026a), whereas SKILL0 internalizes training-time skills to reduce inference-time retrieval.

Vision–text compression. Vision–text compression represents text as images to exploit the information density of visual tokens (Rust et al. 2023; Xing et al. 2025; Wang et al. 2026; Shi et al. 2026). VIST, DeepSeek-OCR, and Glyph develop visual representations for long-context compression (Xing et al. 2026; Wei, Sun, and Li 2025; Cheng et al. 2026). AgentOCR extends this approach to multi-step agents with visual histories, segment optical caching, and agent-controlled compression (Feng et al. 2026b), while SKILL0 jointly renders skills and interaction histories (Lu et al. 2026b). Reading, Not Thinking attributes the text-to-image modality gap partly to degraded reasoning rather than text recognition alone (Sun et al. 2026).

Distillation. Knowledge distillation transfers teacher behavior beyond hard outcome labels (Gu et al. 2024; Xu et al. 2025). On-policy distillation instead supervises student-sampled sequences, reducing the distribution mismatch of fixed off-policy data (Agarwal et al. 2024; Yang et al. 2025; Xiao et al. 2026). OPD further requires compatible reasoning patterns and teacher information not already available to the student (Li et al. 2026). OPSD conditions one model on privileged teacher and standard student contexts, matching their token distributions on student-generated trajectories (Zhao et al. 2026). SDAR extends this formulation to multi-turn agents by combining reinforcement learning with bounded, gated self-distillation (Zhao et al. 2026).

Diagnosing the Visual-History Policy Gap

Visual-history agents remain substantially weaker than text-history agents despite using the same backbone architecture and parameter scale. This gap may arise because rendering makes task-relevant text difficult to recover, or because the visual interface alters how the model uses accessible information to select subsequent actions. We distinguish these two possibilities by examining history recovery, decisions under matched interaction states, and trajectory-level decision quality.

Problem Formulation. We consider a finite-horizon interaction between an agent and an environment $\mathcal{E}$. Let $\mathcal{I}$ denote the task instruction and o_t the observation available at step t. The interaction history at step t is

$$h_t = (o_1, a_1, o_2, a_2, \ldots, o_t), \tag{1}$$

where a_i is the environment action executed at step i. Given $(\mathcal{I}, h_t)$, the policy samples a textual response

$$y_t \sim \pi_\theta(\cdot \mid \mathcal{I}, h_t), \qquad a_t = \Gamma(y_t), \tag{2}$$

where y_t contains intermediate reasoning and an executable action, and the task-specific parser Γ extracts the action. The environment executes a_t and returns o_{t+1}, continuing until the task is completed or the interaction limit is reached.

Following AgentOCR (Feng et al. 2026b), the text history is serialized into ordered segments and rendered as images:

$$\begin{aligned}(\ell_{t,1}, \ldots, \ell_{t,K_t}) &= \mathrm{Split}(h_t),\\ I_t &= \mathrm{Stack}_{k=1}^{K_t}\, \mathcal{R}(\ell_{t,k}; \psi),\end{aligned} \tag{3}$$

where $\mathcal{R}$ is a deterministic renderer and ψ specifies the font, color scheme, padding, and image bounds. Rendered segments are cached by their content and reused when available. The resulting I_t is encoded as visual tokens and replaces the text tokens corresponding to h_t.

History Recovery Is Not the Primary Bottleneck. Reinforcement learning substantially improves the visual-history policy. We first test whether this gain is explained by improved recovery of text from the visual history. Using real AgentOCR trajectories on SearchQA, we construct a History QA benchmark of $4{,}000$ balanced visual questions. Each question asks the model to extract information explicitly recorded in the history, including the latest query, a retrieved document title, keyword presence, or a year. The benchmark requires visual localization and text extraction but no complex sequential decision making. Full construction details are provided in the appendix.

As shown in Fig. 1(b), GRPO raises the average SearchQA EM of the 7B visual-history agent from 14.0% to 40.1%, while image-based History QA decreases from 82.78% to 78.22%. When the same diagnostic histories are provided as text, the corresponding scores are 90.03% and 89.78%. Visual history recovery is therefore imperfect, but the large task improvement cannot be attributed to improved OCR or text extraction.

Matched Semantic States Reveal a Decision Gap. We next compare the two policies under matched interaction states. We sample $10{,}000$ decision prefixes from real SearchQA trajectories generated by the text-history policy. For each prefix, the policies receive the same question and interaction content; only the history representation differs. Further sampling and evaluation details are provided in the appendix.

Figure 1(c) shows a substantial behavioral shift. AgentOCR does not match the text policy's next action type on 29.13% of the sampled states. At the first decision after receiving a retrieval observation, AgentOCR answers in 55.41% of cases, compared with 29.80% for the text policy. Even when both policies choose Search, only 5.22% of query pairs reach a token Jaccard similarity of at least 0.8. Thus, replacing text history with its visual representation changes whether the agent continues searching, when it stops, and what it searches for.

Text Histories Elicit Higher-Quality Trajectories. To assess whether this behavioral difference has a consistent direction in quality, we construct a blind trajectory rubric from real SearchQA validation trajectories generated by AgentOCR and the text-history policy. We draw a stratified sample of $2{,}800$ instances and obtain $2{,}795$ valid judgments from DeepSeek-V4-Flash. Complete trajectories are anonymized and scored from 0 to 2 along five dimensions: evidence uptake, entity/relation tracking, query progression, stopping calibration, and answer grounding. We use inverse-probability weights to account for the sampling design. Detailed criteria of the judge are provided in the appendix.

As shown in Fig. 1(d), the text-history policy performs better in all five dimensions. It scores 1.330 versus 1.109 in evidence uptake, 1.408 versus 1.181 in entity/relation tracking, 1.232 versus 1.131 in query progression, 1.075 versus 1.008 in stopping calibration, and 1.259 versus 1.080 in answer grounding. The visual-history policy therefore differs from the text-history policy in both its decisions and the quality of its reasoning trajectories. We identify this decision-level discrepancy, rather than OCR alone, as a primary source of the remaining performance gap on agentic tasks.

Method

Vision–text compression renders an agent's accumulated text history as images, reducing token usage but changing the interface through which the model accesses its interaction history. As shown in Sec. , this change induces a substantial policy shift and degrades performance on agentic tasks. Since the same model produces stronger trajectories when conditioned on text histories, we use its text-history policy as a teacher for the visual-history policy. **CAPS** transfers policy capability across the two history interfaces through two-stage cross-modal agentic self-distillation. Offline trajectory self-distillation first initializes the model to reason and act from visual histories. During reinforcement learning, online policy self-distillation further transfers the text-history policy to states visited by the visual-history policy itself. This approach preserves the token savings of visual history while reducing the performance gap relative to the text-history policy.

Source and Target Policies

For each canonical interaction state $(\mathcal{I}, h_t)$, the text history and its deterministic rendering define a pair of semantically corresponding conditioning contexts:

$$x_t^T = (\mathcal{I}, h_t), \qquad x_t^I = (\mathcal{I}, I_t). \tag{4}$$

We use a frozen text-history policy π_ϕ^T as the source policy and a trainable visual-history policy π_θ^I as the target policy:

$$\pi_\phi^T(\cdot \mid x_t^T), \qquad \pi_\theta^I(\cdot \mid x_t^I). \tag{5}$$

The two policies share the same model architecture and action space and differ only in their history representations. We seek to minimize their distributional discrepancy over semantically corresponding states while keeping the source policy fixed:

$$\begin{aligned}\mathcal{J}_{\mathrm{tr}}(\theta; \phi) = \mathbb{E}_{(\mathcal{I}, h_t) \sim \rho_{\mathrm{tr}}}\Big[& \\ & D\big(\pi_\phi^T(\cdot \mid x_t^T) \,\|\, \pi_\theta^I(\cdot \mid x_t^I)\big)\Big],\\ \theta^\star = \arg\min_\theta \mathcal{J}_{\mathrm{tr}}(\theta; \phi), & \qquad \phi \text{ is fixed},\end{aligned} \tag{6}$$

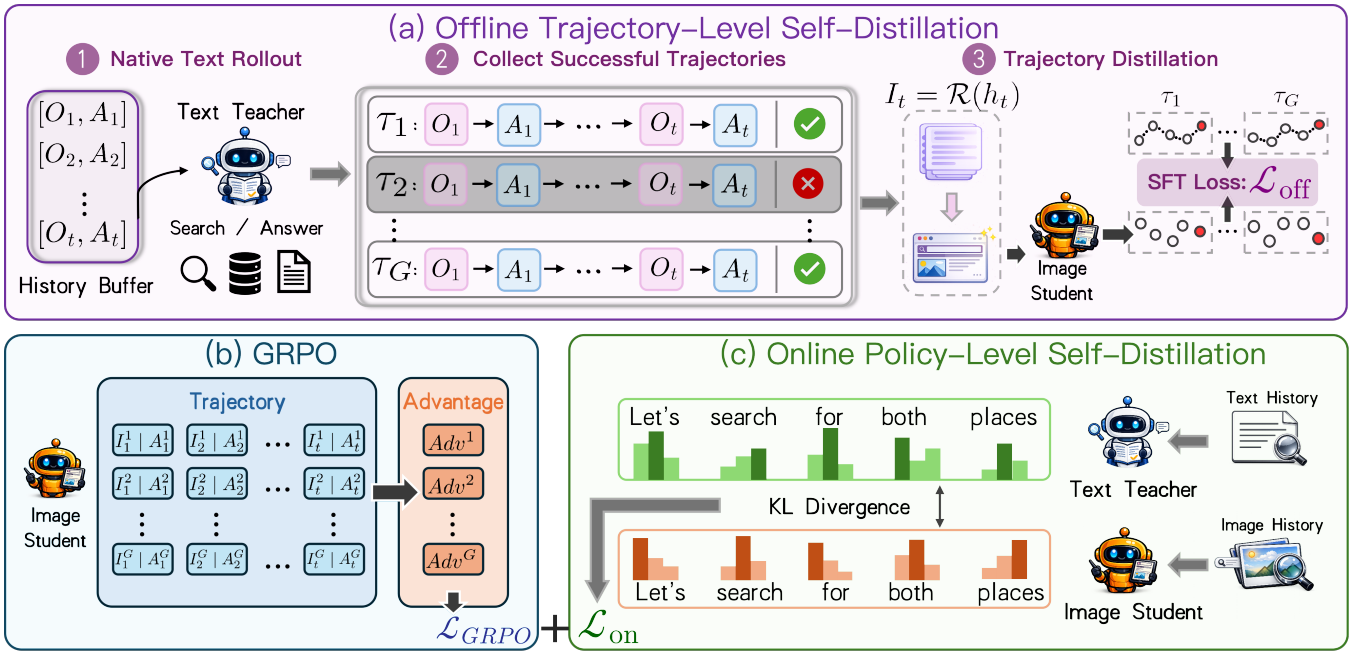

Figure 2: **Overview of CAPS.** Offline distillation transfers successful text-policy behavior to visual histories, GRPO optimizes task rewards, and online distillation supervises student-visited states.

where $D(\cdot\|\cdot)$ denotes a discrepancy between response distributions and ρ_{tr} is the interaction-state distribution used for cross-modal self-distillation.

Offline Trajectory Self-Distillation

Trajectory collection and filtering. Following AgentOCR (Feng et al. 2026b), we train a text-history agent with GRPO and use the resulting policy as the source policy π_ϕ^T. We then collect its interaction trajectories:

$$\tau_i = \left(\mathcal{I}_i, \{h_{i,t}, y_{i,t}^T, a_{i,t}\}_{t=1}^{L_i}\right). \tag{7}$$

We retain trajectories that successfully complete the task, contain parsable actions, and have responses that are not truncated by the length limit. When multiple valid trajectories are available for the same task instance, a deterministic ranking selects one representative trajectory, prioritizing fewer interaction steps, fewer repeated actions, shorter responses, and shorter input contexts. Full data-processing details are provided in the appendix.

For each decision step t in a retained trajectory, $y_{i,t}^T$ denotes the complete source-policy response, including its reasoning and action. We keep the task instruction, interaction content, and response unchanged, and replace only the history representation:

$$(\mathcal{I}_i, h_{i,t}, y_{i,t}^T) \longmapsto (\mathcal{I}_i, I_{i,t}, y_{i,t}^T). \tag{8}$$

The resulting offline dataset is

$$\mathcal{D}_{\text{off}} = \{(\mathcal{I}_i, I_{i,t}, y_{i,t}^T)\}_{i,t}. \tag{9}$$

Training objective. The visual-history policy learns the source responses using next-token prediction objective:

$$\mathcal{L}_{\text{off}}(\theta) = -\mathbb{E}_{(\mathcal{I},I,y^T)\sim\mathcal{D}_{\text{off}}}\left[\frac{1}{|y^T|}\sum_{k=1}^{|y^T|}\log \pi_\theta^I\left(y_k^T \mid \mathcal{I}, I, y_{<k}^T\right)\right]. \tag{10}$$

Supervising the complete response distills both the intermediate reasoning and the action-generation behavior of the source policy.

Visual-History Reinforcement Learning

The checkpoint obtained from offline trajectory self-distillation initializes the visual-history policy, which is then optimized with GRPO (Feng et al. 2026b; Lu et al. 2026b). For a task $\mathcal{I}$, the policy samples a group of G trajectories $\{\tau_i\}_{i=1}^G$ with environmental returns $\{R_i\}_{i=1}^G$. The relative advantage $\widehat{A}_i$ of each trajectory is estimated by normalizing its return within the group. The GRPO objective is

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}\left[\frac{1}{\sum_{i=1}^G|\tau_i|}\sum_{i=1}^G\sum_{t=1}^{|\tau_i|}\min\begin{pmatrix} r_{i,t}(\theta)\widehat{A}_i, \\ \text{clip}(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)\,\widehat{A}_i\end{pmatrix}\right], \tag{11}$$

where the importance sampling ratio is

$$r_{i,t}(\theta) = \frac{\pi_\theta^I(y_{i,t} \mid \mathcal{I}, \tau_{i,<t})}{\pi_{\theta_{\text{old}}}^S(y_{i,t} \mid \mathcal{I}, \tau_{i,<t})}. \tag{12}$$

Online Policy Self-Distillation

Offline trajectory self-distillation covers only states visited by the source policy. During reinforcement learning and deployment, however, the visual policy visits history states induced

by its own previous actions. We therefore use online policy distillation (Zhao et al. 2026) as an auxiliary objective during GRPO, distilling the source policy on the target policy's state distribution. For every state visited by the visual policy during rollout, the system retains the canonical text history formed by the same student actions and environment observations. The student and teacher receive the paired contexts x_t^I and x_t^T from Eq. (4), respectively. The source policy π_ϕ^T is a frozen text-history policy, whereas the target policy π_θ^I is the visual-history policy being optimized. The teacher receives no ground-truth solution and instead reads the text-history representation of the same student-visited interaction history. The student first samples a response $\widehat{y}$ from $\pi_\theta^I(\cdot \mid x_t^I)$. The teacher and student then compute full-vocabulary next-token distributions on the same student-generated prefix $\widehat{y}_{<k}$:

$$\begin{aligned} p_k^S(v) &= \pi_\theta^I(v \mid x_t^I, \widehat{y}_{<k}), \quad v \in \mathcal{V}, \\ p_k^T(v) &= \pi_\phi^T(v \mid x_t^T, \widehat{y}_{<k}), \quad v \in \mathcal{V}. \end{aligned} \tag{13}$$

Full-vocabulary forward KL. For sample b in an update batch $\mathcal{B}$, the teacher-to-student forward KL at response position k is

$$D_{b,k}^{\mathrm{FKL}} = \sum_{v \in \mathcal{V}} p_{b,k}^T(v) \log \frac{p_{b,k}^T(v)}{p_{b,k}^S(v)}. \tag{14}$$

To prevent unusually large position-level KL values from dominating the distillation gradient, we cap each value at τ. Let $m_{b,k} \in \{0, 1\}$ denote the intersection of the student and teacher response masks. If the complete response contains no valid environment action, all of its positions are set to zero. The online distillation objective is

$$\mathcal{L}_{\mathrm{on}}(\theta) = \frac{\sum_{b \in \mathcal{B}} \sum_k m_{b,k} \min\left(D_{b,k}^{\mathrm{FKL}}, \tau\right)}{\max\left(1, \sum_{b \in \mathcal{B}} \sum_k m_{b,k}\right)}. \tag{15}$$

Here, τ is the clipping threshold.

Joint training objective. Environmental reward remains the primary learning signal, while online distillation provides dense policy supervision on student-visited states. The joint objective is

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{GRPO}}(\theta) + \lambda \mathcal{L}_{\mathrm{on}}(\theta), \tag{16}$$

where λ controls the strength of online policy self-distillation.

Experiments

Experimental Setup

Benchmarks. We evaluate CAPS on two multi-turn agent benchmarks. SearchQA follows the evaluation suite used by AgentOCR and SKILL0 (Feng et al. 2026b; Lu et al. 2026b), covering three single-hop datasets (NQ (Kwiatkowski et al. 2019), TriviaQA (Joshi et al. 2017), and PopQA (Mallen et al. 2023)) and four multi-hop datasets (HotpotQA (Yang et al. 2018), 2Wiki (Ho et al. 2020), MuSiQue (Trivedi et al. 2022), and Bamboogle (Press et al. 2023)). NQ and HotpotQA are used for training, and the remaining datasets measure out-of-domain generalization. The agent queries an E5 retriever that returns the top three passages and interacts with the environment for at most four steps. ALFWorld (Shridhar et al. 2020) contains six categories of embodied household tasks and requires substantially longer interaction. Following AgentOCR and SKILL0, our main comparison retains the full interaction history ($H = 50$). We also report $H = 2$, which follows the default configuration of GiGPO.

Baselines. We compare CAPS with text-history and visual-history baselines. *Text (w/o RL)* and *OCR (w/o RL)* use the original text history and the rendered visual history, respectively, without RL. *Text + GRPO (Qwen2.5)* is the text-history policy reported by AgentOCR (Qwen et al. 2025). We additionally train *Text + GRPO (Qwen2.5-VL)*, which uses the same VLM family as CAPS and serves as its source policy. The visual-history agent baselines are AgentOCR and SKILL0, both of which use vision–text compression. We evaluate 3B and 7B models throughout.

Training details. CAPS uses Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct as the target backbones (Bai et al. 2025), with a size-matched Text + GRPO (Qwen2.5-VL) policy as the teacher. Unless otherwise stated, the core training configuration and hyperparameters follow AgentOCR. Offline trajectory self-distillation uses successful teacher trajectories collected from the same training split used by subsequent GRPO; it introduces no additional task instances. Full hyperparameters are provided in the appendix.

Main Results

SearchQA. Table 1 reports the SearchQA results. CAPS obtains the best visual-history result on all six datasets other than Bamboogle at both model scales. With the 3B backbone, it reaches 39.2 average EM, improving over AgentOCR by 5.0 points. With the 7B backbone, it reaches 43.5, a 3.4-point gain over AgentOCR and 0.5 points above its text-history teacher. SKILL0 remains first in the overall average because of its substantially higher Bamboogle score; CAPS ranks second overall while leading on the other six datasets. Relative to the matched text-history teacher, CAPS reduces average and peak memory-context cost by 34.8% and 83.4% for 3B, and by 51.3% and 54.6% for 7B. Thus, the visual-history policy approaches or exceeds its text-history teacher while processing substantially fewer context tokens.

ALFWorld. Table 2 shows consistent gains on longer embodied tasks. Under the full-history setting ($H = 50$), CAPS reaches 93.8% success with the 3B backbone and 95.7% with the 7B backbone. These results exceed SKILL0 by 5.9 points and AgentOCR by 15.6 and 14.5 points, respectively. CAPS also outperforms its text-history teacher by 2.8 points for 3B and 3.1 points for 7B. The 7B policy reduces average and peak memory-context cost by 63.3% and 70.0%, while the 3B policy reduces them by 62.5% and 57.6%. With $H = 2$, CAPS obtains 74.6% and 84.8% success with 3B and 7B, respectively. Without modifying the visual-history renderer or explicitly optimizing image compression, CAPS uses fewer average context tokens than AgentOCR and SKILL0 at both model scales. We attribute this gain to better policy optimization: more effective reasoning and action selection complete tasks in fewer steps, yielding shorter histories.

Table 1: **SearchQA performance and memory-context cost.** We report EM (%) and the average and peak memory-context token cost per step. †Results reported by AgentOCR (Feng et al. 2026b). ‡Results reported by SKILL0 (Lu et al. 2026b). Bold and underlined values denote the best and second-best performance, respectively.

Method	Single-Hop			Multi-Hop				Overall	Tokens/Step	
	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.↑	Avg.↓	Max.↓
Qwen2.5-(VL)-3B-Instruct										
Text (w/o RL)†	9.4	31.3	19.8	15.0	14.8	4.7	16.8	15.9	0.48k	7.34k
Text + GRPO (Qwen2.5)†	39.3	60.6	41.1	37.4	34.6	15.4	26.4	36.4	0.61k	9.55k
Text + GRPO (Qwen2.5-VL)	45.5	62.3	47.7	42.4	42.5	16.0	35.2	41.7	0.57k	10.63k
OCR (w/o RL)†	10.2	27.7	10.9	9.1	12.2	3.7	15.2	12.7	0.15k	1.33k
AgentOCR†	38.6	56.5	41.7	33.6	30.7	14.6	24.0	34.2	0.26k	2.50k
SKILL0‡	39.8	57.5	42.3	35.1	33.7	13.3	**63.7**	**40.8**	**0.18k**	–
CAPS	**42.5**	**60.1**	**44.4**	**40.1**	**37.8**	**15.4**	34.4	39.2	0.37k(34.8%↓)	**1.76k**(83.4%↓)
Qwen2.5-(VL)-7B-Instruct										
Text (w/o RL)†	10.4	32.4	22.3	15.8	15.4	7.2	19.2	17.5	0.70k	10.96k
Text + GRPO (Qwen2.5)†	45.1	63.7	44.0	43.6	43.2	16.8	37.6	41.9	0.73k	13.84k
Text + GRPO (Qwen2.5-VL)	46.9	65.5	48.5	44.6	38.1	18.7	38.4	43.0	0.60k	3.83k
OCR (w/o RL)†	6.9	30.4	12.0	10.5	9.1	5.5	24.0	14.0	0.26k	2.21k
AgentOCR†	43.1	61.0	45.4	40.8	38.3	15.7	36.8	40.1	0.36k	2.65k
SKILL0‡	42.7	61.1	45.3	40.0	38.3	16.4	**66.9**	**44.4**	0.34k	–
CAPS	**44.4**	**64.2**	**47.1**	**42.9**	**42.3**	**18.3**	45.6	43.5	**0.29k**(51.3%↓)	**1.74k**(54.6%↓)

Ablation Studies

Offline and online self-distillation. Table 3 isolates the two stages of CAPS. Using only offline trajectory self-distillation reaches 37.4 and 42.9 average EM for 3B and 7B, while using only online policy self-distillation reaches 37.1 and 41.4. Combining them improves the averages to 39.2 and 43.5, showing that the two stages provide complementary supervision. We also replace the visual-history input in offline distillation with the original text trajectory. This variant falls to 36.3 and 40.7 average EM. For 7B, its result is close to AgentOCR without distillation (40.1), despite using successful teacher trajectories. The comparison shows that offline distillation must initialize the target policy with the ability to reason and make decisions from visual history; exposure to successful text trajectories alone is not sufficient.

Image compression factor. Following AgentOCR, the compression factor $c \geq 1$ controls additional downsampling of the rendered history image. Before minimum-size clipping, an image of width W and height H is resized to

$$W_c = \left\lfloor \frac{W}{\sqrt{c}} \right\rfloor, \qquad H_c = \left\lfloor \frac{H}{\sqrt{c}} \right\rfloor. \tag{17}$$

Its area is therefore reduced by approximately a factor of c. We use $c = 1$ by default. As shown in Fig. 3, stronger compression causes a rapid performance drop: increasing c from 1.0 to 2.5 reduces SearchQA success from 43.54% to 35.50%. Lower resolution makes visual history harder to recover, causing the policy to perform more redundant searches during reasoning and decision making; the average number of search calls increases from 1.83 to 2.15. These additional interactions lengthen the history and offset part of the token savings expected from lower image resolution. We therefore do not reduce token cost through additional image downsampling in the default setting. Instead, CAPS lowers context cost through a better decision policy that completes tasks with fewer interactions.

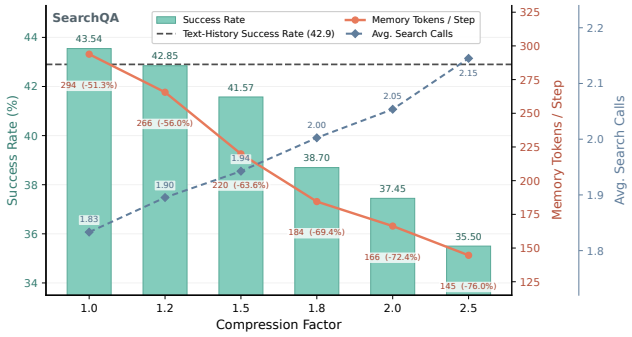

Figure 3: **SearchQA performance, memory-context cost, and search behavior under different image compression factors.** All points evaluate the same 7B CAPS checkpoint.

Policy-Gap Analysis

We repeat the diagnostic evaluations from Sec. to test whether CAPS improves the mechanisms identified there.

History recovery. Table 4 compares generic history recovery with end-task performance. CAPS reaches 43.5% on SearchQA, compared with 40.1% for AgentOCR, but its image-history QA score is lower (73.70 versus 78.22). The base VLM obtains the highest image-history QA score of 82.78 while reaching only 14.0% on SearchQA. CAPS therefore does not improve performance by strengthening generic OCR or history transcription. Instead, the gain comes from using the recovered information more effectively for agent

Table 2: **ALFWorld performance and memory-context cost with Qwen2.5-(VL)-3B/7B-Instruct.** We report success rate (%), the history window H, and the average and peak memory-context token cost per step.

Method	H	Pick	Look	Clean	Heat	Cool	Pick2	Avg.↑	Tokens/Step	
									Avg.↓	Max.↓
Qwen2.5-(VL)-3B-Instruct										
Text (w/o RL)†	50	34.7	18.4	12.7	7.3	14.5	10.4	16.3	1.09k	3.04k
Text + GRPO (Qwen2.5)†	50	92.6	85.7	70.6	86.6	79.3	65.0	79.9	1.02k	3.13k
Text + GRPO (Qwen2.5-VL)	50	95.5	91.6	96.6	93.8	80.9	84.5	91.0	0.54k	2.25k
Text + GRPO (Qwen2.5-VL)	2	91.1	86.2	92.8	77.1	56.6	75.2	80.5	0.09k	0.46k
OCR (w/o RL)†	50	42.8	21.8	10.1	6.2	6.2	9.9	16.2	0.49k	1.63k
AgentOCR†	50	91.9	81.8	76.0	73.3	76.1	70.0	78.2	0.38k	1.14k
SKILL0‡	50	**95.6**	80.4	**100.0**	86.7	78.7	75.2	87.9	0.38k	–
CAPS	50	94.4	**91.7**	97.2	**100.0**	**93.0**	**83.8**	**93.8**	**0.20k**(62.5%↓)	**0.95k**(57.6%↓)
CAPS	2	93.9	71.7	75.0	53.1	46.9	79.5	74.6	**0.03k**(60.5%↓)	**0.11k**(75.5%↓)
Qwen2.5-(VL)-7B-Instruct										
Text (w/o RL)†	50	67.6	35.4	19.3	31.3	30.1	4.4	31.3	1.08k	3.36k
Text + GRPO (Qwen2.5)†	50	92.6	93.8	85.2	80.0	82.7	56.5	81.8	0.95k	2.81k
Text + GRPO (Qwen2.5-VL)	50	100.0	100.0	93.2	96.9	83.4	82.0	92.6	0.58k	2.90k
Text + GRPO (Qwen2.5-VL)	2	95.6	64.1	86.9	96.9	67.7	70.8	83.2	0.09k	0.69k
OCR (w/o RL)†	50	61.0	33.2	17.2	11.6	12.5	16.5	25.3	0.47k	1.36k
AgentOCR†	50	95.6	96.2	78.1	73.2	72.4	72.0	81.2	0.43k	1.22k
SKILL0‡	50	**100.0**	85.8	94.6	81.9	85.7	80.1	89.8	0.41k	–
CAPS	50	96.9	**96.7**	**100.0**	**92.3**	**100.0**	**87.7**	**95.7**	**0.21k**(63.3%↓)	**0.87k**(70.0%↓)
CAPS	2	94.8	42.6	98.3	85.6	79.0	88.0	84.8	**0.03k**(63.5%↓)	**0.13k**(81.6%↓)

Table 3: **SearchQA ablation of our two-stage CAPS.** We report EM (%) and memory-context token cost per step. **Off-SD** denotes Offline Trajectory Self-Distillation performed before GRPO, whereas **On-SD** denotes Online Policy Self-Distillation applied during GRPO.

Variant	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.↑	Tok. Avg.↓	Tok. Max.↓
Qwen2.5-VL-3B-Instruct										
CAPS	**42.5**	**60.1**	**44.4**	**40.1**	**37.8**	**15.4**	**34.4**	**39.2**	0.37k	**1.76k**
Off-SD → GRPO	40.6	59.5	43.2	39.6	36.2	14.0	28.8	37.4	0.27k	3.38k
GRPO + On-SD	41.5	56.8	43.2	36.6	36.4	13.0	32.0	37.1	**0.26k**	3.32k
Text-Traj. SFT → GRPO	42.0	59.2	43.6	37.3	35.5	11.8	24.8	36.3	0.32k	3.48k
Qwen2.5-VL-7B-Instruct										
CAPS	44.4	64.2	47.1	42.9	**42.3**	18.3	**45.6**	**43.5**	0.29k	1.74k
Off-SD → GRPO	**45.7**	**64.7**	**47.8**	**43.7**	41.1	**18.7**	38.4	42.9	0.35k	1.68k
GRPO + On-SD	43.7	62.5	46.3	41.8	40.0	17.8	37.6	41.4	0.32k	2.26k
Text-Traj. SFT → GRPO	45.2	63.6	44.6	39.3	38.7	15.7	37.6	40.7	**0.21k**	**0.86k**

decisions. The lower history-QA score may reflect specialization toward the policy task rather than general visual question answering.

Matched-state decision alignment. Table 5 evaluates the visual-history and text-history policies under matched semantic states. CAPS raises action-type agreement with the text-history policy from 70.87% to 79.53%. More importantly, cases in which the text policy searches but the visual policy answers fall from 51.67% to 32.42%, a reduction of 19.25 points. Search content match also increases from 5.22% to 23.53%. These changes directly address the decision drift observed in Sec. : after distillation, the visual-history policy is less likely to stop before collecting the evidence requested by the text-history policy.

Trajectory quality. Figure 4 reports blind trajectory scores using the same five-dimensional rubric as Sec. . CAPS improves every dimension over AgentOCR, raising the total score from 5.510 to 5.977 and narrowing the gap to the text-history policy (6.303). Gains in evidence use, query progression, stopping, and answer grounding show that CAPS transfers reasoning and decision behavior rather than improving visual text recognition.

Table 4: **History recovery vs. SearchQA Acc.** We report exact match (%) by question type on the strict History QA benchmark.

Metric	Base VLM OCR w/o RL	AgentOCR	CAPS
SearchQA success	14.0	40.1	**43.5**
Document title	**78.2**	71.2	65.6
Keyword presence	**99.3**	99.2	98.3
Last search query	**86.4**	77.6	64.3
Year/number	**67.2**	64.9	66.6
Image overall	**82.78**	78.22	73.70
Text overall	**90.03**	89.78	89.35

Table 5: **Decision alignment under matched semantic states.** Search content match is query token-set Jaccard ≥ 0.8 when both choose Search; answer content match is normalized EM when both choose Answer.

Metric(%)	AgentOCR	CAPS	Gap
Valid action rate	99.96	**99.97**	+0.01
Action agreement	70.87	**79.53**	+8.66
Text Search / Visual Answer	51.67	**32.42**	-19.25
Text Answer / Visual Search	**5.37**	7.85	+2.48
Search content match	5.22	**23.53**	+18.31
Answer content match	71.82	**72.89**	+1.07

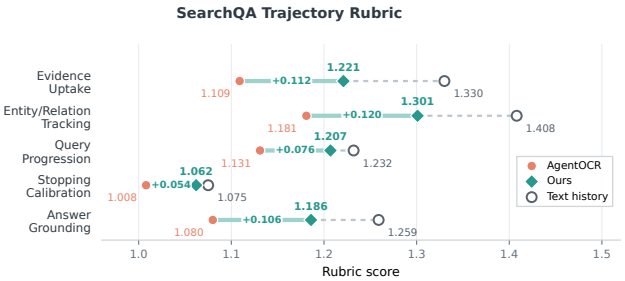

Figure 4: **Blind trajectory-rubric scores on SearchQA.** Each dimension is scored from 0 to 2, and higher is better.

Conclusion

In this paper, we identify an agentic policy gap when multi-step histories are represented as images: visual-history agents can recover content yet use it less effectively for reasoning and decisions. Therefore, we propose CAPS to combine offline trajectory and online policy self-distillation to transfer text-history behavior and improves performance while preserving the token efficiency of vision–text compression. Our results show that effective visual-history compression requires policy alignment, rather than merely preserving recoverable content.

Table 6: Offline distillation data after filtering and one-best-trajectory selection. "Samples" counts step-level SFT examples.

Benchmark	Scale	Scanned	Selected IDs	Samples	Train	Val.
SearchQA	3B	103,424	7,734	18,005	17,599	406
SearchQA	7B	103,424	7,899	18,920	18,487	433
ALFWorld	3B	12,928	1,499	15,011	14,780	231
ALFWorld	7B	12,928	1,460	15,005	14,679	326

A Experimental and Implementation Details

A.1 Training data construction

Offline trajectory self-distillation uses successful rollouts produced by a size-matched text-history policy. These are the actual rollouts saved during text-history policy training. The text-history and visual-history RL policies use the same benchmark training split, so every question used to construct the SFT trajectories already belongs to the task set used by the subsequent visual policy GRPO stage. Offline distillation therefore adds trajectory supervision but introduces no additional task instances. For SearchQA, we read trajectories saved at training step 50 or later. A trajectory is retained only if it succeeds with episode reward 1, all actions are valid, its normalized step indices are contiguous, it contains at most four steps, and every response is no longer than 511 tokens. We additionally reject responses with malformed, empty, conflicting, truncated, or environment-mismatched `<search>` and `<answer>` blocks. The last step must contain a valid final answer.

Several valid rollouts can exist for the same training question. We retain one trajectory per question by lexicographically preferring fewer interaction steps, fewer repeated searches, fewer response tokens, fewer prompt tokens, and finally a later teacher-training step. Every retained trajectory is then expanded into one SFT example per decision step. The input contains the task and the rendered prefix before that decision. Specifically, the original textual interaction history from the text-history teacher trajectory is rendered into a visual history using the same rendering pipeline as in RL; the student therefore receives the rendered image history rather than the original serialized text history. The target is the complete teacher response, including its reasoning and executable action.

ALFWorld uses the same construction with task-specific validity checks. A retained trajectory must succeed with reward 10, finish in a won/done state, contain a contiguous sequence of no more than 30 steps, and have valid `<action>` blocks that match the environment actions. When several trajectories are available for one task instance, we prefer fewer steps, shorter responses, shorter prompts, and a later training step. Both benchmarks use a deterministic 98/2 split by task identifier, so decision steps from one task never cross the SFT train and validation splits.

Table 6 summarizes the resulting filtered datasets.

SearchQA data contain only NQ (Kwiatkowski et al. 2019) and HotpotQA (Yang et al. 2018) training questions. The 3B data contain 3,215 NQ and 4,519 HotpotQA identifiers, while the 7B data contain 3,299 and 4,600, respectively. The remaining five SearchQA datasets are not used to construct offline distillation examples. For ALFWorld, the data cover all six task families.

A.2 Rendering, optimization, and runtime configuration

For both benchmarks, we follow the benchmark and rendering configurations released with AgentOCR (Feng et al. 2026b). Its ALFWorld (Shridhar et al. 2020) setup follows the default GiGPO configuration (Feng et al. 2026a): optical and text prompts are limited to 2,048 and 5,120 tokens, respectively; responses are limited to 512 tokens; each update uses 16 questions with eight rollouts per question; the episode horizon is 50 steps with full history; successful episodes receive reward 10; rollout and validation temperatures are 1.0 and 0.4. Following AgentOCR and SKILL0 (Feng et al. 2026b; Lu et al. 2026b), our main ALFWorld results use $H = 50$, and we additionally evaluate $H = 2$, the GiGPO default history window.

The SearchQA setup follows AgentOCR's Search-R1-style configuration (Feng et al. 2026b; Jin et al. 2025). Optical and text prompts are limited to 4,096 and 14,000 tokens, respectively; responses are limited to 512 tokens; each update uses 128 questions with eight rollouts per question; the episode horizon is four steps with full history; correct answers receive reward 1; rollout temperature is 1.0 and validation is greedy. Both benchmarks use actor learning rate 10^{-6}, PPO mini-batch size 256, and no reference-policy KL penalty in the GRPO objective (Shao et al. 2024).

Rendering also follows AgentOCR (Feng et al. 2026b). Text uses a monospace font and line spacing 1.2. SearchQA uses font size 12, maximum width 560 pixels, and five pixels of final padding; Search tags are blue and Information tags are red. ALFWorld uses font size 10, maximum width 392 pixels, and the same padding; Observations are blue and Actions are red. Step zero receives the same minimally padded blank image used by the runtime pipeline.

Runtime caching follows AgentOCR (Feng et al. 2026b). Unchanged history content is divided into reusable rendered units, cached once, and assembled in chronological order for the current policy input. This avoids repeatedly rendering identical earlier history without changing the information presented to the model.

All unqualified values in Table 7 are shared by 3B and 7B. Scale-dependent settings are reported explicitly in 3B/7B order and are limited to GPU allocation. "Rollout filter" means that online distillation is applied only when the student response contains a valid parsed action; invalid-action responses receive no distillation loss. "Supervised span" means that the divergence is computed over the complete generated response, including both reasoning tokens and the executable action, rather than only the action span.

LoRA targets all matching linear layers, including matching layers in the visual stack; the vision encoder is therefore not frozen. In the notation of the main-paper joint objective, λ is the coefficient of the online distillation loss, while τ is the per-token KL clipping threshold in the online loss. The online teacher is frozen, evaluated in bf16, and scores

Table 7: Training hyperparameters for offline trajectory self-distillation, visual-history GRPO, and online policy self-distillation. Values are shared by 3B and 7B unless a row explicitly reports them as 3B/7B.

Parameter	SearchQA	ALFWorld
Offline trajectory self-distillation		
Global / per-GPU micro batch	32 / 1	16 / 1
Maximum sequence length	8,192	2,560
LoRA rank / alpha / target	32 / 64 / all-linear	32 / 64 / all-linear
Optimizer / β_1, β_2	AdamW / 0.9, 0.95	AdamW / 0.9, 0.95
LR / warmup / weight decay	2×10^{-4} / 0.03 / 0	2×10^{-4} / 0.03 / 0
Scheduler / grad clip / precision	cosine / 1.0 / bf16	cosine / 1.0 / bf16
Configured training budget	2 epochs	2 epochs
Training GPUs (3B / 7B)	2 / 4	4 / 4
Visual-history GRPO		
Questions/update / rollouts/question	128 / 8	16 / 8
Maximum environment steps	4	50
Prompt limits (image / teacher text)	4,096 / 14,000	2,048 / 5,120
Maximum response length	512	512
Actor learning rate / PPO epochs	10^{-6} / 1	10^{-6} / 1
PPO mini / per-GPU micro batch	256 / 8	256 / 8
PPO clip / dual clip / entropy coefficient	0.2 / 3.0 / 0.001	0.2 / 3.0 / 0.001
Success / failure reward	1 / 0	10 / 0
Reference-policy KL coefficient	0	0
Rollout / validation temperature	1.0 / 0.0	1.0 / 0.4
Maximum optimization steps	150	150
Training GPUs (3B / 7B)	2 / 4	4 / 4
Online policy self-distillation		
Teacher temperature	1.1	1.1
Divergence / vocabulary support	forward KL / full	forward KL / full
Distillation weight λ / per-token clip τ	0.05 / 0.05	0.01 / 0.05
Rollout filter / supervised span	valid actions / full response	valid actions / full response
Teacher precision	bf16	bf16

the complete student-generated response under the corresponding text-history prompt. It is removed for validation and inference.

Experiments ran on NVIDIA H200 GPUs. SearchQA uses two GPUs for 3B and four for 7B RL, while ALFWorld uses four GPUs for both scales. All two-epoch SFT runs complete within approximately three hours. For either benchmark, a complete RL run for one model scale typically takes approximately one to two days.

B Diagnostic Evaluation Details and Full Results

B.1 History recovery

The strict History QA set is built from real trajectories saved during SearchQA training. For every noninitial decision, we reconstruct all prior interaction history that was visible to the model when making that decision. Because SearchQA permits at most four decision steps, these examples contain the Search actions and returned results from the preceding one, two, or three steps. We retain histories containing a search action and returned information, and generate four extraction question types: document title, keyword presence, most recent search query, and year/number. Questions whose task identifier appeared in the selected offline SFT data are excluded from the strict split.

The construction scanned 64,578 trajectories. After excluding task identifiers used for offline trajectory self-distillation and removing decisions without a non-empty Search–Information history, balanced sampling produced 4,000 questions from 2,104 task identifiers, with exactly 1,000 questions per type. The set contains 2,258 HotpotQA and 1,742 NQ examples; 1,268 examples come from successful trajectories and 3,768 from trajectories whose actions were all valid. The step distribution is 2,860/959/181 for steps 1/2/3. Image and text modes use the same question and history content, changing only whether the history is rendered.

Table 8: Overall results on the strict History QA benchmark. The EM gap is Text EM minus Image EM. Base VLM denotes `OCR w/o RL`.

Metric	Base VLM	AgentOCR	CAPS
Image EM	82.78	78.22	73.70
Image F1	87.30	84.77	83.49
Text EM	90.03	89.78	89.35
Text F1	91.34	91.24	90.65
Text–Image EM gap	7.25	11.55	15.65

"Image" evaluates each model from the rendered history, whereas "Text" provides the same interaction history directly as text, without an image input. Although CAPSimproves downstream SearchQA, it does not improve generic history extraction, so stronger OCR cannot explain the task gains.

Table 9: History QA EM by source dataset. Base VLM denotes `OCR w/o RL`.

Source Dataset	Base VLM	AgentOCR	CAPS
Image History			
HotpotQA	79.19	74.40	71.04
NQ	87.43	83.18	77.15
Text History			
HotpotQA	87.47	87.29	86.40
NQ	93.34	93.00	93.17

B.2 Matched-state decisions

We extract all nonempty decision prefixes from the text-history validation trajectories. From 51,713 trajectories, this yields 99,969 prefixes; six with invalid text actions are removed. We then draw a 10,000-prefix sample stratified by dataset, step, text action type, and text-trajectory success. The sample contains 5,123 text Search states and 4,877 text Answer states. The two visual policies receive the same question, canonical history content, rendered image, and native SearchQA prompt. Both produce all 10,000 outputs with no missing, duplicate, extra, or input-mismatched prefixes. Generation is deterministic with temperature 0, top-p 1, top-$k = -1$, and at most 512 response tokens.

Valid action is the fraction of visual outputs parsed as Search or Answer; action-type agreement is the fraction matching the text policy's action type. The two cross-action rates are $P(\text{Visual Answer} \mid \text{Text Search})$ and $P(\text{Visual Search} \mid \text{Text Answer})$; they are mismatch proxies, not claims that the visual action is necessarily wrong. Queries are lowercased and split into alphanumeric tokens; exact match compares normalized strings, and token-set Jaccard compares token sets when both policies Search. Search joint alignment counts Jaccard ≥ 0.8 over all text-Search states. Answers use normalized EM. Answer EM is conditional on both policies answering, while answer joint alignment uses all text-Answer states. Hard alignment requires the same action type and either query Jaccard ≥ 0.8 or answer EM; soft alignment averages query Jaccard or answer EM over all states, assigning zero to action mismatches. Gap is CAPS minus AgentOCR, in percentage points for rates. In the breakdowns, N counts sampled prefixes, AO denotes AgentOCR, Answer columns report answer rates, Agree is action-type agreement with text, and Agree Gain is CAPS minus AO Agree.

Repeat rates are computed over each visual policy's Search actions. An exact repeat matches a previous normalized query; a fuzzy repeat has token-set Jaccard at least 0.8. Response statistics cover all 10,000 outputs and count the complete response, including reasoning and the executable action.

The alignment gain is largest at the first decision after observing retrieved evidence and is positive on every dataset. The main change is fewer visual Answer decisions when the text policy chooses Search. Answer-string agreement changes little, while text-Answer/visual-Search mismatch increases by 2.48 points. The result therefore supports closer retrieval and stopping alignment, not uniform copying of every text-policy decision.

Table 10: Complete matched-state action and content alignment.

Metric (%)	AgentOCR	CAPS	Gap
Valid action	99.96	**99.97**	+0.01
Action-type agreement	70.87	**79.53**	+8.66
Text Search / Visual Answer	51.67	**32.42**	-19.25
Text Answer / Visual Search	**5.37**	7.85	+2.48
Query exact match | both Search	1.82	**4.88**	+3.06
Query Jaccard ≥ 0.8 | both Search	5.22	**23.53**	+18.31
Search joint alignment	2.52	**15.89**	+13.37
Answer EM | both Answer	71.82	**72.89**	+1.07
Answer joint alignment	**67.95**	67.15	-0.79
Hard action-content alignment	34.43	**40.89**	+6.46
Soft action-content alignment	42.66	**50.65**	+8.00

Table 11: Answer rates and action-type agreement gain by interaction step.

Step	N	Text Ans.	AO Ans.	CAPS Ans.	Agree Gain
1	5,172	29.80	55.41	39.98	+11.60
2	3,627	66.94	89.85	81.42	+6.26
3	1,201	75.60	94.59	94.34	+3.25

B.3 Trajectory rubric

The trajectory evaluation uses 2,800 questions sampled from 162 strata defined by correctness pattern, dataset, and cross-policy tool-call differences. Rare strata are oversampled. Weighted aggregate statistics use inverse-probability weights whose total corresponds to the 50,505 aligned-question population; the error-tag table reports raw counts.

For each question, the complete trajectories are assigned balanced anonymous labels. The judge does not receive policy names, modality labels, gold answers, or success indicators. It sees the canonical retrieval text and scores evidence uptake, entity/relation tracking, query progression, stopping calibration, and answer grounding from 0 to 2. The primary judge is DeepSeek-V4-Flash (DeepSeek-AI 2026) with thinking enabled, temperature 0, seed 42, and balanced label order. The immutable finalization contains 2,795 valid core judgments (99.82% coverage) and 2,667 judgments with valid step-level metadata (95.25% coverage). Five provider content-filter failures are excluded. The tables report the three policies used in the main diagnostic comparison.

Total sums all five dimensions and ranges from 0 to 10. Local Semantic sums Evidence, Entity, and Grounding and ranges from 0 to 6; Global Control sums Query and Stop and ranges from 0 to 4. Aggregates use unrounded scores.

The rubric results place CAPSbetween AgentOCR and the text-history policy on all five dimensions. The deterministic statistics further show fewer exact or fuzzy repeated searches and fewer response tokens than AgentOCR, consistent with the matched-state analysis of improved retrieval and stopping decisions.

Table 12: Answer rates and action-type agreement by SearchQA dataset.

Dataset	N	Text Answer	AO Answer	CAPS Answer	AO Agree	CAPS Agree
2WikiMultihopQA	3,141	32.79	63.42	54.22	64.63	**75.77**
Bamboogle	25	44.00	56.00	56.00	80.00	**88.00**
HotpotQA	1,541	46.98	68.33	54.25	73.13	**82.80**
MuSiQue	610	34.26	58.36	39.34	71.48	**87.38**
NQ	531	67.98	81.17	79.10	76.27	**77.40**
PopQA	2,387	59.20	82.20	70.26	72.22	**79.64**
TriviaQA	1,765	63.97	82.32	71.61	76.20	**81.02**

Table 13: Repeated search and response-length statistics on matched states.

Metric	AgentOCR	CAPS	Gap
Exact repeated query (%)	5.70	**0.31**	-5.39
Fuzzy repeated query (%)	10.16	**3.83**	-6.33
Mean response tokens	167.66	**68.70**	-98.96
Median response tokens	159	**64**	-95
P95 response tokens	302	**119**	-183
Hit 512-token limit (%)	0.05	**0.03**	-0.02

Table 14: Complete primary trajectory-rubric results. Each base dimension ranges from 0 to 2.

Metric	AgentOCR	CAPS	Text History
Evidence	1.109	1.221	**1.330**
Entity	1.181	1.301	**1.408**
Query	1.131	1.207	**1.232**
Stop	1.008	1.062	**1.075**
Grounding	1.080	1.186	**1.259**
Total	5.510	5.977	**6.303**
Local Semantic	3.371	3.708	**3.996**
Global Control	2.139	2.268	**2.307**

C Prompts and Qualitative Cases

C.1 Agent and diagnostic prompts

Following AgentOCR (Feng et al. 2026b), we report the visual- and text-history prompts separately. Offline SFT uses the visual prompts, while the online teacher uses the corresponding text-history prompt at the student-visited state.

The primary rubric request contains three messages: the rubric system instruction, a dynamically serialized anonymous-trajectory payload, and the output-format instruction. The template below omits real questions and trajectory content. In the actual request, history and retrieval fields are truncated to at most 6,000 characters per field, and model reasoning/output is truncated to at most 2,500 characters.

Figure 5 shows representative visual inputs at two interaction depths using the exact runtime rendering settings.

C.2 Real trajectory cases

Table 17 reports representative examples selected from the finalized rubric set. We show the complete action sequence and the decisive retrieval excerpts rather than full passages; the judge evaluated the complete trajectories. Blue marks searches, gold marks retrieved evidence, red marks an incorrect transition or answer, and green marks the correct answer.

The first case isolates entity binding: both initial retrievals contain the correct mother, but AgentOCR follows a nearby Astor-family distractor, whereas CAPS maintains the relation to Madeleine Force. The second isolates evidence-conditioned stopping: Off-SD continues after retrieving the exact date and drifts to the preceding generation, while CAPS answers as soon as the evidence is sufficient. The improvements therefore reflect how the visual-history policy uses retrieved evidence, rather than access to different facts.

D Limitations

First, rendering choices are important because font size, width, color, and padding affect the readability and token cost of the visual history. We use the original AgentOCR rendering parameters directly and do not conduct a comprehensive ablation over renderer configurations.

Second, offline distillation depends on successful trajectories produced by the text-history policy. A weaker source policy or insufficient coverage of successful states may provide less useful initialization, especially for tasks whose failures are not represented in the retained trajectory set.

Finally, the LLM judge introduces model dependence. The primary rubric has high coverage and uses blind labels with inverse-probability weighting, but human evaluation would further strengthen the diagnostic conclusions.

Table 15: Deterministic statistics on the same weighted 2,800-question sample. Success is weighted trajectory success (%), not the main macro score; Calls/Length are mean Search calls/steps per trajectory; Repeat is the mean earlier-query match count at Jaccard ≥ 0.8, multiplied by 100; Prompt/Resp. are mean tokens per step; Tok./Traj. is their trajectory total; Entity Carry is follow-up searches reusing a retrieved entity (%).

Policy	Success	Calls	Length	Repeat	Prompt/Step	Resp./Step	Tok./Traj.	Entity Carry
AgentOCR	44.71	1.554	2.554	9.80	427.3	161.4	1,576.32	28.53
CAPS	47.38	1.720	2.720	**3.94**	453.8	**67.8**	**1,492.74**	37.12
Text history	47.83	1.920	2.920	7.39	703.7	68.2	2,407.65	43.05

Table 16: Unweighted primary-judge error-tag counts. Tags are not mutually exclusive.

Error tag	AgentOCR	CAPS	Text
No relevant evidence retrieved	191	**120**	156
Evidence present but not used	545	500	**432**
Entity drift	913	764	**611**
Relation composition error	55	58	**52**
Redundant confirmation	**433**	560	832
Premature answer	946	823	**645**
Late stop	25	**19**	60
Answer extraction error	722	723	**613**
Contradicts evidence	401	335	**216**
Invalid action	3	**2**	**2**

Table 17: Qualitative cases from real SearchQA validation trajectories.

Question and behavior	Comparison policy	CAPS
Corrected entity tracking. What is the date of birth of John Jacob Astor VI's mother? *Required relation:* John Jacob Astor VI → mother → birth date.	**Search 1:** "Who is John Jacob Astor VI and when was his mother born?" **Evidence:** the results identify John Jacob Astor VI as the son of **Madeleine Talmage Force**, but also contain a distractor about John Jacob Astor VII and Nancy Astor. **Wrong turn:** AgentOCR follows the distractor and searches "When was Viscountess Astor born?" **Answer: "15 May 1879."** Entity tracking fails despite the relevant mother being present in the first retrieval. Rubric total: 0.	**Search 1:** asks directly for John Jacob Astor VI's mother's birth date. **Evidence:** identifies his mother as **Madeleine Talmage Force**. **Search 2:** "Madeleine Force (mother of John Jacob Astor VI)." **Evidence:** "Madeleine Talmage Force was born on **June 19, 1893**." **Answer: "June 19, 1893."** CAPS preserves the entity relation and retrieves the missing attribute. Rubric total: 10.
Reduced redundant search. When did Otto II, Prince of Anhalt-Aschersleben's father die? *Required relation:* Otto II → father Otto I → death date.	**Search 1:** asks when Otto II's father died. **Sufficient evidence:** Otto II was the son of Otto I, who "died **25 June 1304**." **Redundant continuation:** Off-SD performs two more searches, first for Otto I and then for Henry II, Otto I's father. **Answer: "1266."** It substitutes Henry II's death year for Otto I's after the correct answer was already available. Rubric total: 0.	**Search 1:** "Otto II, Prince of Anhalt-Aschersleben's father's death date." **Sufficient evidence:** the result identifies the father as Otto I and states that he "died **25 June 1304**." **Answer: "25 June 1304."** CAPS answers at the next step, without redundant retrieval or relation drift. Rubric total: 10.

Listing 1: SearchQA visual-history prompt after the first step.

```
<image>

You are an expert agent tasked with answering the given question step-by-step.
Your question: {task_description}

Prior to this step, you have already taken {step_count} step(s).
The image contains the full history:
- Past queries are inside <search>...</search>
- Past results are inside <information>...</information>

Now it's your turn to respond for the current step.
You should first conduct a reasoning process. After completing your reasoning, choose
    only one of the following actions (do not perform both):
(1) If any required knowledge is missing or uncertain, you MUST call a search engine to
    get more external information using format: <search> your query </search>.
(2) Only if the image/history already provides sufficient, reliable information to
    answer with high confidence, provide your final answer within <answer> </answer>
    tags.
```

Listing 2: SearchQA text-history prompt after the first step. The serialized history is inserted at the highlighted memory-context placeholder.

```
You are an expert agent tasked with answering the given question step-by-step.
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. History:
{memory_context}

Now it's your turn to respond for the current step.
You should first conduct a reasoning process. After completing your reasoning, choose
    only one of the following actions (do not perform both):
(1) If any required knowledge is missing or uncertain, you MUST call a search engine to
    get more external information using format: <search> your query </search>.
(2) Only if you have sufficient information to answer the question with high confidence,
     provide your final answer within <answer> </answer> tags.
```

Listing 3: ALFWorld visual-history prompt after the first step.

```
<image>

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

Prior to this step, you have already taken {step_count} step(s). The provided image
    shows the most recent {history_length} observations and the corresponding actions
    you took.

You are now at step {current_step} and your current textual 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.
```

Listing 4: ALFWorld text-history prompt after the first step. The serialized history is inserted at the highlighted action-history placeholder.

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

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

You are now at step {current_step} and your current observation is: {current_observation
    }
Your admissible actions of the current situation are: [{admissible_actions}].

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

Listing 5: History QA visual prompt.

```
<image>

Answer the question using only the interaction history image.
Return only the short answer. Do not explain.
Question: {qa_question}
```

Listing 6: Primary trajectory-rubric system prompt.

```
You are evaluating multi-step search-agent trajectories.
The retrieval text shown after each search is the canonical semantic content that was available in that trajectory's
     history. Do not evaluate OCR transcription. Evaluate whether each anonymous agent used the available
    information correctly.

Score every anonymous trajectory on five dimensions from 0 to 2:
- evidence_uptake: 0 ignores or contradicts evidence; 1 partial; 2 correctly uses key facts.
- entity_relation_tracking: 0 drifts or confuses; 1 mixed; 2 preserves entities and relations.
- query_progression: 0 redundant or irrelevant; 1 some information gain; 2 targets the precise missing fact.
- stopping_calibration: 0 premature answer or needless searches; 1 defensible but inefficient; 2 searches and
    answers at the right time.
- answer_grounding: 0 unsupported or contradicted; 1 partially supported; 2 fully supported by retrieved evidence.

For each step, identify whether accumulated evidence before the action was none, partial, or sufficient. Mark useful
     and redundant search steps. Do not infer model identity, use success labels, or judge final-answer correctness
    against outside knowledge. Return a complete ranking of all anonymous trajectories. Derived efficiency and
    modality-gap metrics are computed locally and must not alter these five scores.
```

Listing 7: Primary trajectory-rubric user-message template.

```
{
  "question": "{question}",
  "anonymous_labels": ["A", "B", "C", "D"],
  "trajectories": {
    "{anonymous_label}": {
      "steps": [
        {
          "step_index": {zero_based_step_index},
          "history_before_action": "{canonical_history_before_this_action}",
          "model_reasoning_and_output": "{model_reasoning_and_raw_output}",
          "action": "{postprocessed_search_or_answer_action}",
          "retrieval_text_after_action": "{canonical_retrieval_text_or_empty_string}"
        }
      ],
      "final_answer": "{parsed_final_answer}"
    }
  },
  "allowed_error_tags": [
    "NO_RELEVANT_EVIDENCE_RETRIEVED",
    "EVIDENCE_PRESENT_NOT_USED",
    "ENTITY_DRIFT",
    "RELATION_COMPOSITION_ERROR",
    "REDUNDANT_CONFIRMATION",
    "PREMATURE_ANSWER",
    "LATE_STOP",
    "ANSWER_EXTRACTION_ERROR",
    "CONTRADICTS_EVIDENCE",
    "INVALID_ACTION",
    "UNJUDGEABLE"
  ]
}

The object under "trajectories" is repeated once for every anonymous label.
For Answer actions, "retrieval_text_after_action" is an empty string.
```

Listing 8: Primary trajectory-rubric output-format instruction.

```
Output-format clarification only. Do not change the semantic rubric, its five dimensions, their meanings, or how the
      trajectories are compared.
Return exactly one JSON object and no prose outside it.
For every anonymous trajectory:
- scores contains exactly evidence_uptake, entity_relation_tracking, query_progression, stopping_calibration, and
     answer_grounding; each score is a JSON number from 0 through 2.
- history_utilization_score is the exact numeric sum of those five scores.
- evidence_state_by_step is a JSON array aligned with trajectory steps and uses none, partial, or sufficient.
- first_sufficient_step is a zero-based integer step index or null.
- useful_search_steps and redundant_search_steps are JSON arrays of zero-based integer step indices.
- error_tags is a JSON array using only the documented error tags.
- primary_failure is a string or null; concise_rationale is a non-empty string.
- confidence is a JSON number from 0.0 through 1.0, never a word or percentage.
pairwise.best_history_user is exactly one anonymous label. pairwise.ranking contains every anonymous label exactly
     once. pairwise.rationale is a non-empty string.
Check all field types, score arithmetic, anonymous labels, and ranking completeness before responding.

Additional step-level schema requirements:
- evidence_state_by_step has exactly one entry per trajectory step and every entry is none, partial, or sufficient.
- first_sufficient_step is null when no step is sufficient; otherwise it is the zero-based index of the first
     sufficient entry.
- useful_search_steps and redundant_search_steps contain unique zero-based integer indices, are disjoint, reference
     only actual search actions, and together classify every search action exactly once.
```

<search>Who is the paternal grandfather of Duke Xuan Of Qin?</search>
<information>{"result": "Doc 1: "Duke Xuan of Qin"\nDuke Xuan of Qin Duke Xuan of Qin (, died 664 BC) was from 675 to 664 BC the twelfth ruler of the Zhou Dynasty state of Qin that eventually united China to become the Qin Dynasty. His ancestral name was Ying (), and Duke Xuan was his posthumous title. Duke Xuan was the eldest of the three sons of his father Duke De of Qin, and succeeded his father as ruler of Qin when Duke De died in 676 BC, aged 34. He reigned for 12 years and died in 664 BC. Although Duke Xuan had nine sons, he\nDoc 2: "Duke Xuan of Qin"\npassed the throne to his younger brother Duke Cheng of Qin, who would in turn pass the throne to the third brother Duke Mu of Qin. Duke Xuan of Qin Duke Xuan of Qin (, died 664 BC) was from 675 to 664 BC the twelfth ruler of the Zhou Dynasty state of Qin that eventually united China to become the Qin Dynasty. His ancestral name was Ying (), and Duke Xuan was his posthumous title. Duke Xuan was the eldest of the three sons of his father Duke De of Qin, and succeeded his father as ruler of Qin\nDoc 3: "Duke Xian of Qin (424\u2013362 BC)"\nDuke Xian of Qin (424\u2013362 BC) Duke Xian of Qin (, 424\u2013362 BC) was from 384 to 362 BC the 29th ruler of the Zhou Dynasty state of Qin that eventually united China to become the Qin Dynasty. His ancestral name was Ying (\u5b34), and Duke Xian was his posthumous title. His given name was Shixi (\u5e08\u96b0) or Lian (\u9023). Duke Xian was the son of Duke Ling of Qin, the 25th ruler of the state of Qin. However, when Duke Ling died in 415 BC, the throne was passed to Duke Ling's uncle Duke Jian, instead of his son.\n"}</information>
<search>who was Duke De of Qin?</search> <information>{"result": "Doc 1: "Duke De of Qin"\nDuke De of Qin Duke De of Qin (, 710\u2013676 BC) was from 677 to 676 BC the eleventh ruler of the Zhou Dynasty state of Qin that eventually united China to become the Qin Dynasty. His ancestral name was Ying (), and Duke De was his posthumous title. Duke De was the second of the three sons of his father Duke Xian of Qin. His younger half-brother Chuzi was the first to succeed Duke Xian in 704 BC, but was killed six years later. His older brother Duke Wu of Qin then ascended the throne and reigned for 20\nDoc 2: "Duke De of Qin"\nthrone as Duke Xuan, Duke Cheng, and Duke Mu, respectively. Duke De of Qin Duke De of Qin (, 710\u2013676 BC) was from 677 to 676 BC the eleventh ruler of the Zhou Dynasty state of Qin that eventually united China to become the Qin Dynasty. His ancestral name was Ying (), and Duke De was his posthumous title. Duke De was the second of the three sons of his father Duke Xian of Qin. His younger half-brother Chuzi was the first to succeed Duke Xian in 704 BC, but was killed six years later. His older brother Duke Wu\nDoc 3: "Duke Xuan of Qin"\nDuke Xuan of Qin Duke Xuan of Qin (, died 664 BC) was from 675 to 664 BC the twelfth ruler of the Zhou Dynasty state of Qin that eventually united China to become the Qin Dynasty. His ancestral name was Ying (), and Duke Xuan was his posthumous title. Duke Xuan was the eldest of the three sons of his father Duke De of Qin, and succeeded his father as ruler of Qin when Duke De died in 676 BC, aged 34. He reigned for 12 years and died in 664 BC. Although Duke Xuan had nine sons, he\n"}</information>

(a) SearchQA: two Search–Information pairs before the third decision.

[Observation]: -= Welcome to TextWorld, ALFRED! =- You are in the middle of a room. Looking quickly around you, you see a bed 1, a desk 1, a drawer 6, a drawer 5, a drawer 4, a drawer 3, a drawer 2, a drawer 1, a garbagecan 1, a shelf 2, a shelf 1, and a sidetable 1. Your task is to: put a book in desk. [Action]: go to drawer 1
[Observation]: You arrive at drawer 1. The drawer 1 is closed. [Action]: open drawer 1
[Observation]: You open the drawer 1. The drawer 1 is open. In it, you see a cellphone 3. [Action]: go to drawer 2
[Observation]: You arrive at drawer 2. The drawer 2 is closed. [Action]: examine drawer 2
[Observation]: The drawer 2 is closed. [Action]: go to shelf 1
[Observation]: You arrive at shelf 1. On the shelf 1, you see a alarmclock 1, a box 1, a creditcard 2, a houseplant 1, and a pen 1. [Action]: go to desk 1
[Observation]: You arrive at desk 1. On the desk 1, you see a cellphone 2, a creditcard 1, a keychain 2, a laptop 1, a mug 3, and a mug 2. [Action]: go to shelf 2
[Observation]: You arrive at shelf 2. On the shelf 2, you see a pencil 1, and a tissuebox 1. [Action]: go to bed 1
[Observation]: You arrive at bed 1. On the bed 1, you see a book 1, a pillow 1, and a teddybear 1. [Action]: take book 1 from bed 1
[Observation]: You pick up the book 1 from the bed 1. [Action]: go to desk 1

(b) ALFWorld: ten Observation–Action pairs before the eleventh decision.

Figure 5: Actual visual histories reconstructed from saved successful trajectories with the runtime renderer and no additional image compression.

References

Agarwal, R.; Vieillard, N.; Zhou, Y.; Stanczyk, P.; Ramos Garea, S.; Geist, M.; and Bachem, O. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, 21246–21263.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Cheng, J.; Liu, Y.; Zhang, X.; Fei, Y.; Hong, W.; Lyu, R.; Wang, W.; Su, Z.; Gu, X.; Liu, X.; et al. 2026. Glyph: Scaling context windows via visual-text compression. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 37145–37158.

Chhikara, P.; Khant, D.; Aryan, S.; Singh, T.; and Yadav, D. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

DeepSeek-AI. 2026. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence.

Feng, L.; Xue, Z.; Liu, T.; and An, B. 2026a. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38: 46375–46408.

Feng, L.; Yang, F.; Chen, F.; Cheng, X.; Xu, H.; Wan, Z.; Yan, M.; and An, B. 2026b. AgentOCR: Reimagining Agent History via Optical Self-Compression. *arXiv preprint arXiv:2601.04786*.

Gu, Y.; Dong, L.; Wei, F.; and Huang, M. 2024. Minillm: Knowledge distillation of large language models. In *The twelfth international conference on learning representations*.

Ho, X.; Nguyen, A.-K. D.; Sugawara, S.; and Aizawa, A. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, 6609–6625.

Jimenez, C. E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; and Narasimhan, K. 2024. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations*, volume 2024, 54107–54157.

Jin, B.; Zeng, H.; Yue, Z.; Yoon, J.; Arik, S.; Wang, D.; Zamani, H.; and Han, J. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. 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)*, 1601–1611.

Kang, M.; Chen, W.-N.; Han, D.; Inan, H. A.; Wutschitz, L.; Chen, Y.; Sim, R.; and Rajmohan, S. 2025. Acon: Optimizing context compression for long-horizon llm agents. *arXiv preprint arXiv:2510.00615*.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Li, Y.; Zuo, Y.; He, B.; Zhang, J.; Xiao, C.; Qian, C.; Yu, T.; Gao, H.-a.; Yang, W.; Liu, Z.; et al. 2026. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*.

Lu, Y.; Ye, R.; Du, Y.; Wang, J.; Liu, S.; and Chen, S. 2026a. LongSeeker: Elastic Context Orchestration for Long-Horizon Search Agents. *arXiv preprint arXiv:2605.05191*.

Lu, Z.; Yao, Z.; Wu, J.; Han, C.; Gu, Q.; Cai, X.; Lu, W.; Xiao, J.; Zhuang, Y.; and Shen, Y. 2026b. Skill0: In-context agentic reinforcement learning for skill internalization. *arXiv preprint arXiv:2604.02268*.

Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; and Hajishirzi, H. 2023. 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)*, 9802–9822.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N. A.; and Lewis, M. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5687–5711.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Rust, P.; Lotz, J. F.; Bugliarello, E.; Salesky, E.; Lhoneux, M. d.; and Elliott, D. 2023. Language modelling with pixels. In *The Eleventh International Conference on Learning Representations*, 1–32. OpenReview.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36: 68539–68551.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sheng, G.; Zhang, C.; Ye, Z.; Wu, X.; Zhang, W.; Zhang, R.; Peng, Y.; Lin, H.; and Wu, C. 2024. Hybridflow:

A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*.

Shi, Y.; Liu, S.; Yang, Y.; Mao, W.; Chen, Y.; Gu, Q.; Su, H.; Cai, X.; Wang, X.; and Zhang, A. 2026. MemOCR: Layout-Aware Visual Memory for Efficient Long-Horizon Reasoning. *arXiv preprint arXiv:2601.21468*.

Shridhar, M.; Yuan, X.; Côté, M.-A.; Bisk, Y.; Trischler, A.; and Hausknecht, M. 2020. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*.

Sun, K.; Yuan, X.; Liu, H.; Zhao, C.; Zhang, C.; Dredze, M.; and Bai, F. 2026. Reading, not thinking: Understanding and bridging the modality gap when text becomes pixels in multimodal llms. *arXiv preprint arXiv:2603.09095*.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics*, 10: 539–554.

Wang, A. J.; Li, L.; Lin, Y.; Li, M.; Wang, L.; and Shou, M. Z. 2024. Leveraging visual tokens for extended text contexts in multi-modal learning. *Advances in Neural Information Processing Systems*, 37: 14325–14348.

Wang, G.; Xie, Y.; Jiang, Y.; Mandlekar, A.; Xiao, C.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Wang, W.; Xiong, S.; Chen, G.; Gao, W.; Guo, S.; He, Y.; Huang, J.; Liu, J.; Li, Z.; Li, X.; et al. 2025. Reinforcement learning optimization for large-scale learning: An efficient and user-friendly scaling library. *arXiv preprint arXiv:2506.06122*.

Wang, X.; Cui, Y.; Wang, J.; Zhang, F.; Wang, Y.; Zhang, X.; Luo, Z.; Sun, Q.; Li, Z.; Wang, Y.; et al. 2026. Multimodal learning with next-token prediction for large multimodal models. *Nature*, 1–7.

Wei, H.; Sun, Y.; and Li, Y. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Xiao, B.; Xia, B.; Yang, B.; Gao, B.; Shen, B.; Zhang, C.; He, C.; Lou, C.; Luo, F.; Wang, G.; et al. 2026. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*.

Xing, L.; Wang, A. J.; Yan, R.; Shu, X.; and Tang, J. 2026. Vision-centric token compression in large language model. *Advances in Neural Information Processing Systems*, 38: 33080–33110.

Xing, L.; Yan, R.; Wang, A. J.; Li, Z.; and Tang, J. 2025. See the text: From tokenization to visual reading. *arXiv preprint arXiv:2510.18840*.

Xu, W.; Han, R.; Wang, Z.; Le, L.; Madeka, D.; Li, L.; Wang, W.; Agarwal, R.; Lee, C.-Y.; and Pfister, T. 2025. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *International Conference on Learning Representations*, volume 2025, 64616–64646.

Xu, W.; Liang, Z.; Mei, K.; Gao, H.; Tan, J.; and Zhang, Y. 2026. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38: 17577–17604.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2369–2380.

Yao, S.; Zhao, J.; Yu, D.; Shafran, I.; Narasimhan, K. R.; and Cao, Y. 2022. React: Synergizing reasoning and acting in language models. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

Zhao, S.; Xie, Z.; Liu, M.; Huang, J.; Pang, G.; Chen, F.; and Grover, A. 2026. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. *arXiv preprint arXiv:2601.18734*.

Zhou, Z.; Qu, A.; Wu, Z.; Kim, S.; Prakash, A.; Rus, D.; Zhao, J.; Low, B. K. H.; and Liang, P. P. 2025. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*.