

AgentOCR: Reimagining Agent History via Optical Self-Compression

Lang Feng^{1,*}, Fuchao Yang^{1,*}, Feng Chen¹, Xin Cheng¹, Haiyang Xu²,
Zhenglin Wan¹, Ming Yan², Bo An^{1,†}

¹Nanyang Technological University, Singapore

²Tongyi Lab, Alibaba Group

{lang005, fuchao001}@e.ntu.edu.sg, boan@ntu.edu.sg

Abstract

Recent advances in large language models (LLMs) enable agentic systems trained with reinforcement learning (RL) over multi-turn interaction, but practical deployment is bottlenecked by rapidly growing textual histories that inflate token and memory costs. We introduce AgentOCR, a framework that exploits visual tokens’ superior information density by representing the accumulated observation-action history as a compact rendered image. To make multi-turn rollouts scalable, AgentOCR proposes segment optical caching. By decomposing history into hashable segments and maintaining a visual cache, this mechanism eliminates redundant re-rendering. Beyond fixed rendering, AgentOCR introduces agentic self-compression, where the agent actively emits a compression rate and is trained with compression-aware reward to adaptively balance task success and token efficiency. We conduct extensive experiments on challenging agentic benchmarks, ALFWorld and search-based QA. Remarkably, AgentOCR preserves over 95% of text-based agent performance while substantially reducing token consumption (>50%), yielding consistent token and memory efficiency. Further analysis validates a 20× rendering speedup from optical caching and effective self-compression balancing. Our code is available at <https://github.com/langfengQ/AgentOCR>.

1 Introduction

Recent advancements in large language models (LLMs) (Achiam et al., 2023; Guo et al., 2025; Yang et al., 2025; Comanici et al., 2025) have enabled agentic systems (Xu et al., 2025b; Chang et al., 2025; Zhang et al., 2025c; Xie et al., 2025) that tightly couple perception, deliberation, and action, thereby reducing the need for continuous human supervision (Yao et al., 2023; Zhang et al.,

2025b). These successes increasingly motivate framing agent improvement as reinforcement learning (RL) over long-horizon trajectories-optimizing tool use, planning, and control policies end-to-end from interaction feedback (Feng et al., 2025b; Wang et al., 2025b).

However, RL training for LLM agents remain difficult. A major challenge arises from the burden of *long-context processing* (Lu et al., 2025; Kang et al., 2025). As agents interact with environments through *multi-turn* decision loops, they must buffer a comprehensive trajectory of past observations and action sequences. As shown in Fig. 1(a), this historical data accumulates relentlessly, causing the input context to swell to a massive volume of tokens within a single trajectory. Such rapid expansion not only exhausts the finite token budget of current LLMs but also incurs prohibitive inference latency and compute cost due to expensive attention prefill and KV-cache management (Shah et al., 2024; Jiang et al., 2024).

Recent breakthroughs in vision language models (VLMs) (Bai et al., 2025b,a; Chen et al., 2024b) and optical character recognition (OCR) (Wei et al., 2025; Cui et al., 2025a; Xing et al., 2025) suggest a promising solution: *visual information density*. Notably, DeepSeek-OCR (Wei et al., 2025) shows that the visual modality can serve as a far more compact carrier of information than text. By rendering textual content into images, the token footprint can be compressed by approximately 10× compared to raw text tokens, substantially reducing the number of tokens processed during model inference.

Building on this, we propose *AgentOCR*, a visually-grounded method that reimagines agent history not as a string of text, but as a dynamic sequence of images (Fig. 1(b)). Specifically, AgentOCR represents the accumulated observation-action history as a compact image and conditions the agent’s policy on this visual history for the multi-turn decision making. To ensure scalability

*Equal Contribution

†Corresponding author

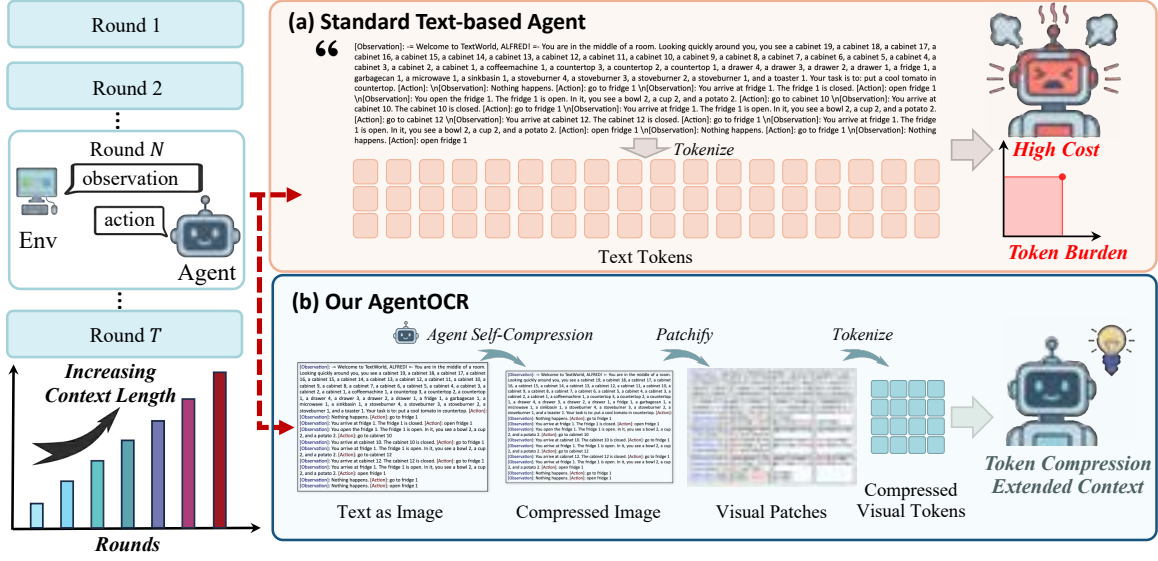


Figure 1: Comparison of text agent and AgentOCR. (a) Text agent accumulates a heavy token burden from raw text history. (b) Our AgentOCR requires significantly fewer visual tokens via optical self-compression.

ity and efficiency in long rollouts, AgentOCR introduces a *segment optical caching*. This mechanism decomposes the history context into segments and maintains a hash-based optical cache, allowing the agent to reuse previously rendered content and avoid redundant processing as the history expands. Beyond static image, AgentOCR features *agentic self-compression*, which empowers the agent to actively modulate its own visual fidelity at each step. By learning via RL, AgentOCR adaptively selects the most appropriate compression factor to save token costs, thereby achieving a favorable balance between task success and efficiency.

We evaluate AgentOCR on two challenging agentic benchmarks: ALFWorld (Shridhar et al., 2021), which features long-horizon decision-making, and search-based QA (Jin et al., 2025), characterized by highly text-dense interactions. Our results demonstrate that AgentOCR preserves over 95% of the task performance of strong text-based agent pipelines, while reducing token consumption by over 50% (up to 80% in peak tokens), leading to substantially lower overhead. Furthermore, our analysis validates that segment optical caching accelerates rendering by over 20 $\times$, and self-compression mechanism effectively optimizes the trade-off between information density and cost.

2 Related Work

2.1 Reinforcement Learning for LLM Agents

RL (Sutton and Barto, 2018) has become a commonly adopted paradigm for aligning agents with

human preferences (Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023) and improving their behaviors (Sheng et al., 2024; Wang et al., 2025a) in complex scenarios, such as PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), Dr. GRPO (Liu et al., 2025b), Clip-Cov (Cui et al., 2025b), GSPO (Zheng et al., 2025), DAPO (Yu et al., 2025b), GiGPO (Feng et al., 2025b), and HGPO (He et al., 2026). These RL-trained agents have been widely deployed in dynamic and open-ended environments, spanning interactive games (Narasimhan et al., 2015; Brockman et al., 2016), GUI control (Rawles et al., 2023; Ye et al., 2025; Liu et al., 2025a), embodied tasks (Shridhar et al., 2021), as well as web (Zhou and Zanette, 2024; Putta et al., 2024; Feng et al., 2025a) and tool-enhanced environments (Qian et al., 2025; Sun et al., 2025; Dong et al., 2025; Xue et al., 2025).

2.2 Optical Character Recognition

OCR converts textual information in images into computer-readable text and has been widely applied in document digitization (Doermann, 1998; Smith, 2007; Arlazarov et al., 2022), image text extraction (Jaderberg et al., 2016; Baek et al., 2019; Long et al., 2018), and document parsing (Katti et al., 2018; Xu et al., 2020; Rausch et al., 2021). With the advent of deep learning, OCR systems have shifted toward end-to-end frameworks, including Nougat (Blecher et al., 2023), Donut (Kim et al., 2022), TrOCR (Li et al., 2023), Pix2Struct (Lee et al., 2023) and GOT-OCR2.0 (Wei et al., 2024),

enabling unified image-to-text modeling for complex scenarios. Recently, OCR has begun to be explored as a vision-text compression mechanism, such as DeepSeek-OCR (Wei et al., 2025), VIST (Xing et al., 2025), and Glyph (Cheng et al., 2025). These approaches offer a novel solution for processing extremely long contexts, however, research in this direction remains in its early stages.

2.3 Agent Memory

Long-horizon interaction with LLM-based agents requires persistent memory, but naively appending full histories quickly exceeds fixed context windows (Zhang et al., 2025d,b; Hu et al., 2025b). Accordingly, long-context language modeling focuses on efficiency and length generalization, including sparse or hierarchical attention (Beltagy et al., 2020; Fu et al., 2024; Xiao et al., 2024), recency-biased positional schemes (Ding et al., 2024; Su et al., 2024; Xiong et al., 2024), and prompt compression (Ge et al., 2023; Yoon et al., 2024; Zhang et al., 2024). In parallel, retrieval-based methods treat external stores as non-parametric memory and fetch relevant information on demand (Ge et al., 2023; Chen et al., 2024a). Beyond context optimization, recent agent frameworks introduce explicit memory modules to support long-term behavior, ranging from virtualized context management and scalable backends (Packer et al., 2023; Chhikara et al., 2025) to structured or hierarchical representations for extended and multi-agent tasks (Xu et al., 2025a; Anokhin et al., 2024; Hu et al., 2025a; Zhang et al., 2025a). More recent work explores learning-based memory control, where agents adaptively write, retain, and retrieve information, treating memory as an active component (Yan et al., 2025; Zhou et al., 2025; Yu et al., 2025a).

3 Preliminaries

3.1 Problem Setup

We formulate the interaction between an LLM agent and an environment (e.g., a physical simulator or an external tool API) as a sequential decision-making process over a finite horizon $T \in \mathbb{N}$. The agent is instantiated as an LLM parameterized by θ , and its behavior is modeled as a stochastic policy π_θ . At each step $t \in \{1, \dots, T\}$, the agent receives an observation $\mathbf{o}_t \in \mathcal{O}$ (e.g., API outputs) and has the interaction history up to step t as

$$\mathbf{h}_t = (\mathbf{o}_1, \mathbf{a}_1, \mathbf{o}_2, \mathbf{a}_2, \dots, \mathbf{o}_t). \quad (1)$$

The agent then samples a textual action $\mathbf{a}_t \sim \pi_\theta(\cdot \mid \mathcal{I}, \mathbf{h}_t)$, where $\mathcal{I}$ denotes the task instruction, $\mathbf{a}_t \in \mathcal{V}^n$ is a token sequence drawn from the vocabulary $\mathcal{V}$ with maximum length $n \in \mathbb{N}$. The action $\mathbf{a}_t$ is flexible and may explicitly include intermediate reasoning (e.g., chain-of-thought (Wei et al., 2022)) or tool invocations. After executing $\mathbf{a}_t$, the environment returns a scalar reward r_t and the next observation $\mathbf{o}_{t+1}$. Notably, $\mathbf{h}_t$ often becomes extensively long due to extended horizons or verbose environmental observations, posing significant challenges for context processing.

3.2 Agentic Reinforcement Learning

RL has become a standard post-training paradigm for enhancing LLM agents. While our work is algorithm-agnostic and compatible with various agentic RL algorithms, we consider Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as the representative algorithm due to its simplicity and efficiency.

Specifically, GRPO samples a group of trajectories $\{\tau_i\}_{i=1}^G$ for each input and estimates advantages $\hat{A}_i$ by normalizing rewards within the group. To ensure training stability, we optimize the clipped surrogate objective:

$$J(\theta) = \mathbb{E} \left[\frac{1}{GT} \sum_{i=1}^G \sum_{t=1}^T \min \left(\rho_{t,i} \hat{A}_i, \text{clip}(\rho_{t,i}, 1 \pm \epsilon) \hat{A}_i \right) \right], \quad (2)$$

where $\rho_{t,i} = \frac{\pi_\theta(\mathbf{a}_{t,i} \mid \mathcal{I}, \mathbf{h}_{t,i})}{\pi_{\theta_{\text{old}}}(\mathbf{a}_{t,i} \mid \mathcal{I}, \mathbf{h}_{t,i})}$ is the importance sampling ratio, ϵ is the clipping hyperparameter. Here, we omit the KL-divergence regularization for notational brevity. Crucially, optimizing Eq. (2) necessitates computing gradients over the entire cumulative history $\mathbf{h}_t$. In realistic agent scenarios, $\mathbf{h}_t$ rapidly accumulates thousands of tokens (e.g., >10k tokens in multi-turn search tasks (Jin et al., 2025)). Since the computational complexity and memory footprint scale with token count, processing such lengthy textual trajectories becomes prohibitively expensive.

4 AgentOCR

Agentic tasks require the policy to condition on an *ever-growing multi-turn interaction history*. This growth creates a severe bottleneck in practice. Not only does it reach the model’s context window limits, but it also drives up computational costs as transformer computation scales superlinearly with sequence length. Hence, efficient token compression is imperative for deploying LLM agents (Kong

et al., 2025).

In this section, we introduce AgentOCR that addresses this bottleneck by reimagining interaction history as an *optical memory*. Instead of processing raw textual logs, we render the accumulated history into a compact visual representation. By leveraging the superior information density of visual tokens compared to text, this approach substantially reduces the token footprint while maintaining full access to historical details. To ensure scalability and dynamic adaptability in long-horizon rollouts, AgentOCR incorporates two innovations: (1) *segment optical caching* (Fig. 2(a)), which eliminates redundant rendering overhead by systematically reusing cached visual segments, and (2) *agentic self-compression* (Fig. 2(b)), a mechanism that empowers the agent to actively modulate the compression rate, thereby optimizing the trade-off between information density and token cost.

In the remainder of this section, we detail optical encoding in Sec. 4.1 for visual representation, followed by segment optical caching in Sec. 4.2 and self-compression and RL training in Sec. 4.3.

4.1 Optical Memory Encoding

Memory buffer. AgentOCR maintains an external memory buffer $\mathcal{M}_t$ that stores the interaction records up to step t . Each record contains an observation-action pair $(\mathbf{o}_t, \mathbf{a}_t)$ (or task-specific equivalents such as tool queries and results). The memory module serializes the entire interaction history into a textual form $\mathbf{h}_t = \text{Fetch}(\mathcal{M}_{t-1})$.

Memory rendering. We define a deterministic renderer $\mathcal{R}$ that maps the textual interaction history to an RGB image $\mathbf{I}_t = \mathcal{R}(\mathbf{h}_t; \psi)$, where ψ denotes rendering hyperparameters (e.g., font family and size, colors, padding, and bounds on image width and height). At step t , AgentOCR constructs a multimodal policy input by combining $\mathcal{I}$ and the rendered history image $\mathbf{I}_t$, and samples an action from a vision-language policy: $\mathbf{a}_t \sim \pi_\theta(\cdot \mid \mathcal{I}, \mathbf{I}_t)$. The sampled action $\mathbf{a}_t$ is then applied to the environment to obtain the next observation $\mathbf{o}_{t+1}$, and the memory buffer is updated accordingly.

4.2 Segment Optical Caching

Rendering the entire history $\mathbf{h}_t$ from scratch at every step is wasteful and becomes a major latency bottleneck in multi-turn rollouts. A naive alternative is to render only the newly appended context and append it to the previously rendered history

image. This yields near-constant per-step rendering overhead, but it cannot reuse recurring content, and its memory still grows with the accumulated rendered pixels.

AgentOCR instead performs caching at the granularity of segments. The core idea is to decompose the full history context into independent segments and memory rendered segments in a dictionary keyed by segment content. At each step, we assemble the history image by stacking cached segment images in order, rendering only segments that have not been seen before. This cache naturally accelerates both recurring boilerplate and repeated tool outputs, and can also reuse newly arrived observations/actions whenever they match previously observed segments.

Segment representation. We split the history context into segments. Let $\text{Split}(\mathbf{h}) = (\ell_1, \dots, \ell_K)$ denote this operation, where each ℓ_i is a text segment. We use a deterministic segment renderer $\mathcal{R}(\ell; \psi)$ that maps a single segment to an RGB image under the same rendering hyperparameters.

Segment cache. For each environment instance e , AgentOCR maintains a per-episode cache

$$\mathcal{C}^{(e)} = \{(k(\ell), \mathbf{I}(\ell))\}, \quad (3)$$

where $k(\ell)$ is a fast content key (e.g., a hash of the normalized segment text, optionally including style metadata) and $\mathbf{I}(\ell)$ is the rendered image of segment ℓ . Unlike naive cache, which incrementally renders the newly appended context, this cache stores each unique segment at most once and reuses it whenever the same segment reappears. In our implementation, $\mathcal{C}^{(e)}$ persists within an episode and is reset at episode boundaries.

Cache lookup and assembly. At step t , we obtain the full history $\mathbf{h}_t$ and split it into segments $\text{Split}(\mathbf{h}_t) = (\ell_{t,1}, \dots, \ell_{t,K_t})$. For each segment $\ell_{t,i}$, we first query the cache. On a miss, we render and insert it:

$$\mathbf{I}(\ell_{t,i}) = \begin{cases} \mathcal{C}^{(e)}[k(\ell_{t,i})], & \text{if } k(\ell_{t,i}) \in \mathcal{C}^{(e)}, \\ \mathcal{R}(\ell_{t,i}; \psi), & \text{otherwise.} \end{cases} \quad (4)$$

$$\text{if miss: } \mathcal{C}^{(e)}[k(\ell_{t,i})] \leftarrow \mathbf{I}(\ell_{t,i}). \quad (5)$$

The full optical memory image is then constructed by vertically stacking segment images in order:

$$\mathbf{I}_t = \text{Stack}\left(\mathbf{I}(\ell_{t,i})\right)_{i=1}^{K_t}. \quad (6)$$

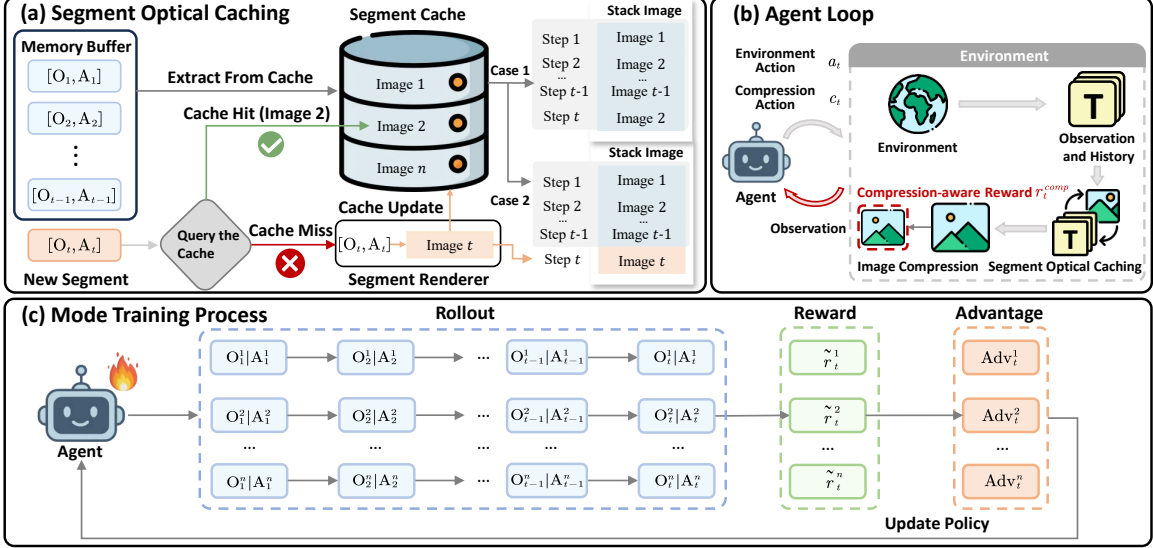


Figure 2: Overview of AgentOCR. (a) Segment optical caching decomposes the history context into segments, reuses cached renderings via content keys, and assembles the optical memory by stacking segment images. (b) The agent receives the optical observation and history, selects an environment action, and a compression rate. (c) The agent is trained with RL, jointly optimizing task performance and token efficiency.

Because newly appended observations or actions are processed as just additional segments, they can be reused without re-rendering whenever they match cached content (e.g., repeated queries and tool responses).

Complexity. Let U_t be the number of cache-miss segments among $\{\ell_{t,i}\}_{i=1}^{K_t}$. The per-step rendering cost becomes $O(U_t)$ segment renders, while cache hits require only dictionary lookup and image stacking. In many agent workloads, interaction logs contain substantial repetition, so typically $U_t \ll K_t$ and rendering overhead is significantly reduced. For spatial complexity, the cache stores one image per unique segment per episode, yielding $O(|C^{(e)}|)$ images rather than $O(T)$ full-history images, thereby avoiding heavy duplication across environment timesteps.

4.3 Agentic Self-Compression

Compression decision. Instead of treating the optical renderer $\mathcal{R}$ (defined in Sec. 4.1) as a static background process, AgentOCR exposes it as an executable tool. We conceptualize this interaction as a parameterized invocation alongside environment actions, the policy generates a structured call via $\langle \text{compression} \rangle c_t \langle / \text{compression} \rangle$, where the compression factor $c_t \geq 1$ dynamically modulates the rendering fidelity. This design aligns with standard tool-use paradigms, allowing the agent to explicitly query its interaction history with variable preci-

sion. Upon receiving the call, the system executes the renderer with the specified compression factor. Formally, this produces a scaled image $\mathbf{I}_{t+1}$:

$$\text{size}(\mathbf{I}_{t+1}) = \left(\left\lfloor \frac{H_{t+1}}{\sqrt{c_t}} \right\rfloor, \left\lfloor \frac{W_{t+1}}{\sqrt{c_t}} \right\rfloor \right). \quad (7)$$

This spatial downsampling operation effectively reduces the number of visual tokens. Consequently, the agent can strategically modulate the compression rate based on specific task characteristics, thereby optimizing the trade-off between token cost and information density.

Compression-aware reward. To incentivize the agent to identify suitable compression without compromising task success, we introduce a compression-aware reward term for RL training. This reward is strictly conditioned on episode success, ensuring that the agent treats compression as a secondary cost-optimization objective rather than a primary goal. Then, we employ a logarithmic reward formulation that reflects the diminishing returns of information density. Formally, let $\mathbb{I}_{\text{succ}}(\tau) \in \{0, 1\}$ be the success indicator. The compression reward at step t is defined as:

$$r_t^{\text{comp}} = \begin{cases} \ln(c_t), & \text{if } \mathbb{I}_{\text{succ}}(\tau) = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

The total reward used for RL optimization is given by $\tilde{r}_t = r_t + \lambda r_t^{\text{comp}}$, where $\lambda \geq 0$ is the weight parameter of compression reward, which controls the

trade-off between task performance and compression efficiency. This scalar reward is used by the agentic RL optimizer (e.g., GRPO) to update the policy π_θ through the objective in Eq. (2). However, applying the compression reward r_t^{comp} at every training iteration can induce overly greedy behavior, where the agent aggressively increases compression to maximize immediate reward. To mitigate this effect, we adopt an intermittent reinforcement schedule, injecting the compression reward only at intervals of K training iterations. This schedule introduces the efficiency signal periodically while maintaining the primary optimization pressure on task completion. Under this design, AgentOCR learns to allocate vision tokens adaptively, achieving further token reduction while maintaining strong task performance.

5 Experiment

In this section, we present comprehensive empirical evaluations of AgentOCR across two representative multi-turn agent benchmarks. Specifically, our experiments aim to investigate the following key aspects: (1) the comparative performance and token efficiency of AgentOCR relative to text-based agents; (2) the quantitative analysis of the vision-text compression ratio of AgentOCR; (3) the computational efficiency of the segment optical caching mechanism; (4) the ablation study on the effectiveness of the self-compression mechanism.

5.1 Experimental Setup

Benchmarks. We evaluate AgentOCR on ALFWorld (Shridhar et al., 2021) and search-based QA (Jin et al., 2025). Both benchmarks exhibit sustained context growth but with different interaction profiles. ALFWorld contains the embodied tasks, requiring the agent to manipulate objects within a simulated household environment. Search-based QA, conversely, focuses on multi-turn tool use and information retrieval. The agent must actively interact with a search engine to query external knowledge, requiring it to deal with denser, web-style textual traces.

Baselines. We compare AgentOCR against both text-based and optical-history variants across prompting and RL regimes. Specifically, we evaluate: (1) *Text (w/o RL)*, which feeds raw textual history to a text-only model; (2) *OCR (w/o RL)*, which conditions the model on rendered optical memory images without RL; and (3) *Text + GRPO*,

a strong baseline that applies RL directly to the raw textual context. In contrast, our AgentOCR applies GRPO to the optical-memory agent, enabling efficient optimization over compact visual histories.

Training details. We use the Qwen2.5-VL (Bai et al., 2025b) family as the backbone models. Text-only variants use Qwen2.5-3B/7B-Instruct, while optical-history variants use Qwen2.5-VL-3B/7B-Instruct. We keep all training settings and hyperparameters identical across methods to ensure controlled comparisons. For extra hyperparameters of AgentOCR, we set $\lambda = 0.01$ and $K = 5$. More details are provided in Appendix B.

5.2 Main Results

We first evaluate the overall performance of all methods, with results reported in Tab. 1 and Tab. 2. A direct comparison of the inference-only baselines (“Text” vs. “OCR”) reveals the inherent efficiency advantage of the visual modality. Across both benchmarks, optical history drastically reduces token consumption, cutting average usage by approximately 55% on ALFWorld and 70% on search tasks. However, this compression initially comes at a cost. The off-the-shelf VLM struggles to ground the condensed visual history effectively, resulting in a substantial performance drop relative to their text-based counterparts.

AgentOCR effectively bridges this gap through RL training, aligning the policy to the visual modality to attain task performance comparable to text-based baselines across model scales. On ALFWorld, AgentOCR with 3B and 7B models achieves 78.2% and 81.2% respectively, virtually matching the text agents (within a $\sim 1\%$ margin). This trend holds for search tasks, where AgentOCR retains over 95% of the performance of the Text+GRPO baselines (e.g., achieving 40.1% vs. 41.9% on the 7B model). Crucially, these results underscore that AgentOCR offers a highly favorable trade-off between token cost and task success. Rather than merely matching the baseline, our method fundamentally alleviates the inference bottleneck by slashing token consumption by $>50\%$ (up to 80.9% in peak contexts), proving that high-density visual representations can support rigorous agentic reasoning with significantly reduced overhead.

5.3 Vision-Text Compression Analysis

In this part, we investigate the trade-off between token compression efficiency and task performance

Method	Pick & Place	Look	ALFWorld				Avg.	Tokens/Step	
			Clean	Heat	Cool	Pick2 & Place		Avg.	Max.
Qwen2.5-(VL)-3B-Instruct									
Text (w/o RL)	34.7	18.4	12.7	7.3	14.5	10.4	16.3	1.09k	3.04k
OCR (w/o RL)	42.8	21.8	10.1	6.2	6.2	9.9	16.2	0.49k	1.63k
Text + GRPO	92.6	85.7	70.6	86.6	79.3	65.0	79.9	1.02k	3.13k
AgentOCR	91.9	81.8	76.0	73.3	76.1	70.0	78.2	0.38k(61.7%↓)	1.14k(63.6%↓)
Qwen2.5-(VL)-7B-Instruct									
Text (w/o RL)	67.6	35.4	19.3	31.3	30.1	4.4	31.3	1.08k	3.36k
OCR (w/o RL)	61.0	33.2	17.2	11.6	12.5	16.5	25.3	0.47k	1.36k
Text + GRPO	92.6	93.8	85.2	80.0	82.7	56.5	81.8	0.95k	2.81k
AgentOCR	95.6	96.2	78.1	73.2	72.4	72.0	81.2	0.43k(54.7%↓)	1.22k(56.6%↓)

Table 1: Performance on ALFWorld tasks. We report the success rate (%) and the average and peak memory context token cost per step.

Method	Single-Hop				Multi-Hop				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
OCR (w/o RL)	10.2	27.7	10.9	9.1	12.2	3.7	15.2	12.7	0.15k	1.33k
Text + GRPO	39.3	60.6	41.1	37.4	34.6	15.4	26.4	36.4	0.61k	9.55k
AgentOCR	38.6	56.5	41.7	33.6	30.7	14.6	24.0	34.2	0.26k _(57.4%↓)	2.50k _(73.8%↓)
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
OCR (w/o RL)	6.9	30.4	12.0	10.5	9.1	5.5	24.0	14.0	0.26k	2.21k
Text + GRPO	45.1	63.7	44.0	43.6	43.2	16.8	37.6	41.9	0.73k	13.84k
AgentOCR	43.1	61.0	45.4	40.8	38.3	15.7	36.8	40.1	0.36k _(50.7%↓)	2.65k _(80.9%↓)

Table 2: Performance on search-based QA tasks. We report the exact matching score (%) and the average and peak memory context token cost per step. [†] and ^{*} denote in-domain and out-of-domain respectively.

using the trained AgentOCR (7B) across varying fixed compression factors ($c_t \in [1.0, 2.0]$). As illustrated in Fig. 3, increasing the compression factor yields substantial gains in token efficiency but incurs a performance penalty. Notably, we identify a robust compression zone up to $\sim 55\%$ token savings (at $c_t = 1.2$), where the model successfully maintains over 95% of the text-based performance (99.5% for ALFWorld and 95.0% for Search). However, surpassing this efficiency threshold triggers an accelerated performance decay. For instance, as savings increase further to $\sim 67\%$ ($c_t = 2.0$), the average performance drops significantly, highlighting the non-linear tension between information density and reasoning accuracy.

We further observe a distinct divergence in robustness beyond this threshold. ALFWorld demonstrates high resilience, retaining 87.2% of performance even at $c_t = 2.0$, likely due to its reliance on coarse-grained scene understanding. In contrast, the text-dense search task is highly sensitive, with performance plummeting to 66.8% as aggres-

sive downscaling blurs critical textual cues. These findings underscore the necessity of our *agentic self-compression* mechanism. While a static image offers a safe baseline, a dynamic policy is required to exploit higher compression rates in robust steps while reverting to high fidelity for sensitive reasoning, thereby breaking the static trade-off ceiling.

5.4 Analysis of Cache

To evaluate the scalability of segment optical caching, we compare it against two alternatives: *no cache*, which re-renders the full history $\mathbf{h}_t$ at every step, and *naive cache*, which incrementally renders the newly appended context at step t and appends it to a growing optical-memory image.

Tab. 3 shows that *no cache* suffers from large latency and strong growth over time (3509.39 ms on average and 115.43 ms/step), reflecting the redundant cost of repeatedly rendering an increasingly long history. *Naive cache* removes this time growth and achieves a $17.28\times$ speedup, since each step renders only the appended suffix, yielding an

Method	Avg (ms)	Render Time		Speedup $\uparrow$	Cache Mem		Mem Save $\uparrow$
		Grow (ms/step)			Peak (MB)	Grow (MB/step)	
No Cache	3509.39	115.43	1.00 $\times$	—	—	—	—
Naive Cache	203.08	0.03	17.28 $\times$	151.41	2.79	0.00%	
Ours	168.77	-1.23	20.79$\times$	110.80	1.91	26.82%	

Table 3: Cache mechanism ablation. *Growth/step* is the slope from a least-squares linear fit over steps 1–50. *Speedup* is relative to “no cache”. *Mem Saving* is relative to the peak cache memory of “naive cache”.

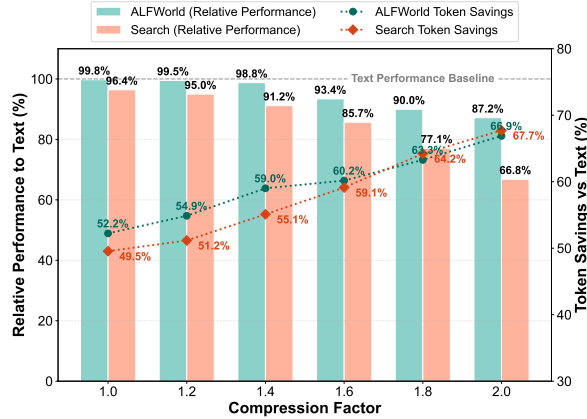


Figure 3: Vision-text compression efficiency. The bars (left axis) denote the success rate relative to the text-based agent baseline, while the lines (right axis) indicate the percentage of tokens saved.

effective per-step cost of $O(1)$ in typical rollouts. However, because it treats each newly arrived line as distinct and permanently appends the rendered result, its cache memory still grows with rollout length (151.41 MB peak and 2.79 MB/step), i.e., space scales with the accumulated rendered pixels.

Segment optical caching further reduces per-step rendering work. By memoizing newline-level segments in a content-keyed dictionary, newly appended context can often be satisfied by cache hits and thus requires no rendering. Therefore, the number of cache-miss segments may shrink as the cache warms up, producing a negative time growth (-1.23 ms/step) and the best average latency (168.77 ms). Meanwhile, segment-level reuse also reduces redundant storage (110.80 MB), corresponding to a 26.82% peak-memory saving relative to *naive cache* and aligning with space scaling dominated by the number of unique segments rather than the number of steps.

5.5 Analysis of Self-Compression

At last, we conduct ablation studies to analyze the effectiveness of the agentic self-compression mechanism using the Qwen2.5-VL-3B-Instruct on the ALFWorld. The results are summarized in Tab. 4.

Configuration	SR (%)	Avg. c_t	Avg. Vis. Tok.
<i>Without RL</i>			
w/o Self-Compression	12.1	1.00	441.2
Self-Compression	11.8	1.05	436.9
<i>RL Training</i>			
w/o Self-Compression	78.4	1.00	458.1
Self-Compression ($K=1$)	45.3	4.91	193.2
Self-Compression ($K=5$)	78.2	1.28	381.7

Table 4: Ablation study on self-compression. *SR*, *Avg. c_t* , and *Avg. Vis. Tok.* denote the success rate, average compression factor, and average vision tokens.

The results demonstrate that RL is essential for effectively leveraging the self-compression mechanism. In the absence of RL, AgentOCR lacks the prior knowledge required to modulate the compression factor, resulting in a negligible change in token usage and a slight performance decline compared to the fixed baseline. When RL is applied with a dense reward schedule ($K=1$), the agent prioritizes the immediate compression reward by aggressively increasing the compression factor to 4.91, which degrades visual fidelity and causes the success rate to plummet to 45.3%. However, by adopting the intermittent reinforcement schedule ($K=5$), AgentOCR successfully balances the trade-off between information density and token cost. This configuration learns a favorable compression rate, reducing average visual token consumption from 458.1 to 381.7 while maintaining a success rate of 78.2%, comparable to the 78.5% achieved by the uncompressed visual baseline.

6 Conclusions

In this work, we present AgentOCR as an exploration into the potential of visual tokens as a compact history medium for multi-turn LLM agents. By integrating segment optical caching to mitigate rendering overhead and agentic self-compression to adaptively balance cost and fidelity, our method demonstrates that the visual modality can effectively complement textual history. Empirical results on ALFWorld and search-based QA suggest

that this optical approach allows agents to retain the majority of their decision-making capabilities while significantly reducing token consumption, offering a resource-efficient alternative to text-only processing. We envision future research expanding on this foundation to explore hybrid storage architectures and unified multimodal interfaces, moving closer to the versatile and efficient information processing found in biological systems.

Limitations

While AgentOCR demonstrates promising results in agentic tasks, several limitations warrant discussion and suggest directions for future work:

First, AgentOCR relies on off-the-shelf VLMs (Qwen2.5-VL series) that were not specifically designed for OCR intensive tasks. Although the proposed AgentOCR framework is model-agnostic in principle, we do not evaluate its behavior across a broader range of VLM architectures, like DeepSeek-OCR (Wei et al., 2025), with different visual tokenization strategies or patch resolutions. Performance and compression efficiency may vary depending on the inductive biases of the underlying vision encoder.

Second, AgentOCR relies on a deterministic text-to-image renderer with fixed hyperparameters such as font size, line spacing, color schemes, and image resolution. While we observe stable performance under our default configuration, we do not systematically explore the sensitivity of the agent to different rendering choices. Suboptimal rendering settings may reduce text legibility or distort layout cues, potentially affecting downstream reasoning.

Last, the current design assumes agent history consists primarily of text (observations, actions, tool outputs) that can be rendered as text-as-image. However, many realistic agent scenarios involve inherently visual elements: GUI screenshots with complex layouts, scientific plots, diagrams, and structured tables. Investigating compression strategies for such multimodal histories could significantly expand AgentOCR’s application scope beyond text-centric domains.

Ethical Considerations

In this work, all experiments were conducted using established public benchmarks and synthesized environments, which do not involve any sensitive personal information or data privacy concerns. Since our work currently does not involve real-world

deployment, its immediate social impact is limited. However, from a long-term perspective, the proposed self-compressing optical context significantly reduces the token overhead for agentic tasks, which may contribute to the development of energy-efficient "Green AI" and facilitate deployment on resource-constrained devices in the future. We used LLMs as writing assistants to refine language. Specifically, their use is limited to grammar correction, style improvement, and wording adjustments for clarity and conciseness.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its Industry Alignment Fund - Prepositioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Andrey Kravchenko, Mikhail Burtsev, and Evgeny Burnaev. 2024. AriGraph: Learning knowledge graph world models with episodic memory for LLM agents. *arXiv preprint arXiv:2407.04363*.
- Vladimir Arlazarov, Elena Andreeva, Konstantin Bulatov, Dmitry Nikolaev, O.O. Petrova, B.I. Savelev, and Oleg Slavin. 2022. Document image analysis and recognition: A survey. *Computer Optics*.
- Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. 2019. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*.
- Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.
- Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and Na Zou. 2025. MAIN-RAG: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Jiale Cheng, Yusen Liu, Xinyu Zhang, Yulin Fei, Wenyi Hong, Ruiliang Lyu, Wei Han Wang, Zhe Su, Xiaotao Gu, Xiao Liu, and 1 others. 2025. Glyph: Scaling context windows via visual-text compression. *arXiv preprint arXiv:2510.17800*.
- Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, and 1 others. 2025a. PaddleOCR 3.0 technical report. *arXiv preprint arXiv:2507.05595*.
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, and 1 others. 2025b. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.
- Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. LongRoPE: Extending LLM context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*.
- David Doermann. 1998. The indexing and retrieval of document images: A survey. *Computer Vision and Image Understanding*.
- Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, and 1 others. 2025. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*.
- Lang Feng, Weihao Tan, Zhiyi Lyu, Longtao Zheng, Haiyang Xu, Ming Yan, Fei Huang, and Bo An. 2025a. Towards efficient online tuning of VLM agents via counterfactual soft reinforcement learning. In *Proceedings of the International Conference on Machine Learning*.
- Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025b. Group-in-group policy optimization for LLM agent training. In *Proceedings of the Advances in Neural Information Processing Systems*.
- Tianyu Fu, Haofeng Huang, Xuefei Ning, Genghan Zhang, Boju Chen, Tianqi Wu, Hongyi Wang, Zixiao Huang, Shiyao Li, Shengen Yan, and 1 others. 2024. Moa: Mixture of sparse attention for automatic large language model compression. *arXiv preprint arXiv:2406.14909*.
- Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2023. In-context autoencoder for context compression in a large language model. *arXiv preprint arXiv:2307.06945*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Shuo He, Lang Feng, Qi Wei, Xin Cheng, Lei Feng, and Bo An. 2026. Hierarchy-of-groups policy optimization for long-horizon agentic tasks. *arXiv preprint arXiv:2602.22817*.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.
- Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. 2025a. HiAgent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32779–32798.

- Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, and 1 others. 2025b. Memory in the age of AI agents. *arXiv preprint arXiv:2512.13564*.
- Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2016. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*.
- Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H Abdi, Dongsheng Li, Chin-Yew Lin, and 1 others. 2024. MInference 1.0: Accelerating pre-filling for long-context LLMs via dynamic sparse attention. In *Proceedings of the Advances in Neural Information Processing Systems*.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.
- Minki Kang, Wei-Ning Chen, Dongge Han, Huseyin A Inan, Lukas Wutschitz, Yanzhi Chen, Robert Sim, and Saravan Rajmohan. 2025. ACON: Optimizing context compression for long-horizon LLM agents. *arXiv preprint arXiv:2510.00615*.
- Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2D documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-free document understanding transformer. In *Proceedings of the European Conference on Computer Vision*.
- Zhenglun Kong, Yize Li, Fanhu Zeng, Lei Xin, Shvat Messica, Xue Lin, Pu Zhao, Manolis Kellis, Hao Tang, and Marinka Zitnik. 2025. Token reduction should go beyond efficiency in generative models—from vision, language to multimodality. *arXiv preprint arXiv:2505.18227*.
- Tom Kwiakowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*.
- Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *Proceedings of the International Conference on Machine Learning*.
- Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Haowei Liu, Xi Zhang, Haiyang Xu, Yuyang Wanyan, Junyang Wang, Ming Yan, Ji Zhang, Chunfeng Yuan, Changsheng Xu, Weiming Hu, and 1 others. 2025a. PC-Agent: A hierarchical multi-agent collaboration framework for complex task automation on pc. *arXiv preprint arXiv:2502.14282*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding R1-Zero-Like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.
- Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. 2018. TextSnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European Conference on Computer Vision*.
- Miao Lu, Weiwei Sun, Weihua Du, Zhan Ling, Xuesong Yao, Kang Liu, and Jiecao Chen. 2025. Scaling LLM multi-turn RL with end-to-end summarization-based context management. *arXiv preprint arXiv:2510.06727*.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.
- Karthik Narasimhan, Tejas Kulkarni, and Regina Barzilay. 2015. Language understanding for text-based games using deep reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the Advances in Neural Information Processing Systems*.
- Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. MemGPT: Towards LLMs as operating systems.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring

- and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP*.
- Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. 2024. Agent Q: Advanced reasoning and learning for autonomous AI agents. *arXiv preprint arXiv:2408.07199*.
- Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiushi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. 2025. ToolRL: Reward is all tool learning needs. *arXiv preprint arXiv:2504.13958*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Proceedings of the Advances in Neural Information Processing Systems*.
- Johannes Rausch, Octavio Martinez, Fabian Bissig, Ce Zhang, and Stefan Feuerriegel. 2021. DocParser: Hierarchical document structure parsing from renderings. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. 2023. Android in the wild: A large-scale dataset for android device control. In *Proceedings of the Advances in Neural Information Processing Systems*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. 2024. FlashAttention-3: Fast and accurate attention with asynchrony and low-precision. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 37, pages 68658–68685.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. DeepSeek-Math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*.
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWorld: Aligning text and embodied environments for interactive learning. In *Proceedings of the International Conference on Learning Representations*.
- R. Smith. 2007. An overview of the tesseract ocr engine. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Proceedings of the Advances in Neural Information Processing Systems*.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*.
- Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. 2025. ZeroSearch: Incentivize the search capability of LLMs without searching. *arXiv preprint arXiv:2505.04588*.
- Richard S Sutton and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction*. MIT press.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.
- Weixun Wang, Shaopan Xiong, Gengru Chen, Wei Gao, Sheng Guo, Yancheng He, Ju Huang, Jiaheng Liu, Zhendong Li, Xiaoyang Li, and 1 others. 2025a. Reinforcement learning optimization for large-scale learning: An efficient and user-friendly scaling library. *arXiv preprint arXiv:2506.06122*.
- Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, and 1 others. 2025b. RAGEN: Understanding self-evolution in LLM agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*.
- Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, and 1 others. 2024. General OCR theory: Towards OCR-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.
- Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. DeepSeek-OCR: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits its reasoning in large language models. *Proceedings*

of the Advances in Neural Information Processing Systems.

Guangxuan Xiao, Jiaming Tang, Jingwei Zuo, Junxian Guo, Shang Yang, Haotian Tang, Yao Fu, and Song Han. 2024. DuoAttention: Efficient long-context LLM inference with retrieval and streaming heads. *arXiv preprint arXiv:2410.10819*.

Bin Xie, Rui Shao, Gongwei Chen, Kaiwen Zhou, Yinchuan Li, Jie Liu, Min Zhang, and Liqiang Nie. 2025. GUI-explorer: Autonomous exploration and mining of transition-aware knowledge for GUI agent. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Ling Xing, Alex Jinpeng Wang, Rui Yan, Xiangbo Shu, and Jinhui Tang. 2025. Vision-centric token compression in large language model. *arXiv preprint arXiv:2502.00791*.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, and 1 others. 2024. Effective long-context scaling of foundation models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025a. A-MEM: Agentic memory for LLM agents. *arXiv preprint arXiv:2502.12110*.

Yifan Xu, Xiao Liu, Xueqiao Sun, Siyi Cheng, Hao Yu, Hanyu Lai, Shudan Zhang, Dan Zhang, Jie Tang, and Yuxiao Dong. 2025b. AndroidLab: Training and systematic benchmarking of android autonomous agents. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, Zejun Ma, and Bo An. 2025. SimpleTIR: End-to-end reinforcement learning for multi-turn tool-integrated reasoning. *arXiv preprint arXiv:2509.02479*.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Kristian Kersting, Jeff Z Pan, Hinrich Schütze, and 1 others. 2025. Memory-R1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *Proceedings of the International Conference on Learning Representations*.

Jiabo Ye, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, Zhaoqing Zhu, Ziwei Zheng, Feiyu Gao, Junjie Cao, Zhengxi Lu, and 1 others. 2025. Mobile-Agent-v3: Fundamental agents for GUI automation. *arXiv preprint arXiv:2508.15144*.

Chanwoong Yoon, Taewhoo Lee, Hyeon Hwang, Minbyul Jeong, and Jaewoo Kang. 2024. COMPACT: Compressing retrieved documents actively for question answering. *arXiv preprint arXiv:2407.09014*.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and 1 others. 2025a. MemAgent: Reshaping long-context LLM with multi-conv RL-based memory agent. *arXiv preprint arXiv:2507.02259*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Lingjun Liu, and 1 others. 2025b. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. 2025a. G-Memory: Tracing hierarchical memory for multi-agent systems. *arXiv preprint arXiv:2506.07398*.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, and 1 others. 2025b. The landscape of agentic reinforcement learning for LLMs: A survey. *arXiv preprint arXiv:2509.02547*.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2024. AdaComp: Extractive context compression with adaptive predictor for retrieval-augmented large language models. *arXiv preprint arXiv:2409.01579*.

Taolin Zhang, Dongyang Li, Qizhou Chen, Chengyu Wang, and Xiaofeng He. 2025c. BELLE: A bi-level multi-agent reasoning framework for multi-hop question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025d. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.

Yifei Zhou and Andrea Zanette. 2024. ArCHer: Training language model agents via hierarchical multi-turn RL. In *Proceedings of the International Conference on Machine Learning*.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. 2025. MEM1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint arXiv:2506.15841*.

Algorithm 1 Environment Wrapper of AgentOCR

```

1: Input: Reward param  $\lambda$ , Compression interval  $K$ .
2: Internal State: Memory  $\mathcal{M}$ , Cache  $\mathcal{C}^{(e)}$ , Renderer  $\mathcal{R}$ .
3: Function Step( $\mathbf{a}_t, c_t, t$ )
4:   // 1. Physical environment step
5:   Execute  $\mathbf{a}_t$ , receive  $\mathbf{o}_{t+1}, r_t$ , and success_flag
6:    $\mathcal{M}_{t+1} \leftarrow \mathcal{M}_t \cup \{(\mathbf{o}_{t+1}, \mathbf{a}_t)\}$ 
7:   // 2. Optical memory rendering (tool execution  $\mathcal{R}$ )
8:    $\mathbf{h}_{t+1} \leftarrow \text{Fetch}(\mathcal{M}_{t+1})$ 
9:   Segments  $(\ell_{t+1,1}, \dots, \ell_{t+1,K}) \leftarrow \text{Split}(\mathbf{h}_{t+1})$ 
10:   $\mathbf{L}_{\text{imgs}} \leftarrow []$ 
11:  for  $i = 1$  to  $K$  do
12:     $k_i \leftarrow \text{Hash}(\ell_{t+1,i})$ 
13:    if  $k_i \notin \mathcal{C}^{(e)}$  then
14:       $\mathcal{C}^{(e)}[k_i] \leftarrow \mathcal{R}(\ell_{t+1,i}; \psi)$  {Cache miss: render segment}
15:    end if
16:    Append  $\mathcal{C}^{(e)}[k_i]$  to  $\mathbf{L}_{\text{imgs}}$ 
17:  end for
18:   $\mathbf{I}_{\text{raw}} \leftarrow \text{Stack}(\mathbf{L}_{\text{imgs}})$ 
19:  // Apply compression  $c_t$  (Eq. (7))
20:   $\mathbf{I}_{t+1} \leftarrow \text{Resize}(\mathbf{I}_{\text{raw}}, \text{scale} = 1/\sqrt{c_t})$ 
21:  // 3. Compression-aware reward (Eq. (8))
22:  if success_flag is True then
23:     $r_t^{\text{comp}} \leftarrow \ln(c_t)$  if success
24:  else
25:     $r_t^{\text{comp}} \leftarrow 0$ 
26:  end if
27:  // Apply sparse reward injection
28:   $\tilde{r}_t \leftarrow r_t + \lambda \cdot r_t^{\text{comp}} \cdot \mathbb{I}(t \bmod K = 0)$ 
29:  return  $(\mathbf{o}_{t+1}, \mathbf{I}_{t+1}), \tilde{r}_t$ 
30: End Function

```

A Pseudo Code

We provide the detailed pseudo code for AgentOCR to facilitate reproduction. Algorithm 1 handles the backend mechanics: it maintains the segment cache, executes the rendering tool with the requested compression c_t , and computes the efficiency-aware reward. Algorithm 2 details the interaction loop, where the policy explicitly selects the tool parameter c_t to control the resolution of the subsequent visual observation.

B Experiments

B.1 Details of Benchmarks

ALFWorld. ALFWorld (Shridhar et al., 2021) is an embodied environment comprising 3,827 tasks, which is publicly available for non-commercial research purposes. The objective for LLM agents is to accomplish household tasks spanning six categories: Pick & Place, Examine in Light (Look), Clean & Place (Clean), Heat & Place (Heat), Cool & Place (Cool), and Pick Two & Place. At each interaction step, the LLM agent selects an action based on the current observation and interaction history, then receives feedback from the environment to verify task completion.

Algorithm 2 AgentOCR Policy Rollout

```
1: Input: Task Instruction  $\mathcal{I}$ , Total training steps  $T_{\max}$ .
2: Initialize: Policy  $\pi_\theta$ , Env Wrapper  $\mathcal{E}$ .
3: Initialize: Global training step  $t \leftarrow 0$ .
4: while  $t < T_{\max}$  do
5:    $\mathcal{E}.\text{Reset}()$ 
6:   Get initial observation  $(\mathbf{o}_t, \mathbf{I}_t)$  from  $\mathcal{E}$ 
7:   while episode not done do
8:     // 1. Decision making (policy generates action +
      // tool call)
9:     Sample action  $\mathbf{a}_t \sim \pi_\theta(\cdot \mid \mathcal{I}, \mathbf{I}_t)$ 
10:    // Parse specific tag as described in Sec. 4.3
11:    Parse output  $\rightarrow (\mathbf{a}_t, c_t)$  via <compression> tag
12:    // 2. Execute action and tool call
13:     $(\mathbf{o}_{t+1}, \mathbf{I}_{t+1}), \tilde{r}_t \leftarrow \mathcal{E}.\text{Step}(\mathbf{a}_t, c_t, t)$ 
14:    Store transition for optimization using  $\tilde{r}_t$ 
15:     $(\mathbf{o}_t, \mathbf{I}_t) \leftarrow (\mathbf{o}_{t+1}, \mathbf{I}_{t+1})$ 
16:     $t \leftarrow t + 1$ 
17:   end while
18:   Update  $\pi_\theta$  via RL optimizer
19: end while
```

Search-based QA. We utilize the QA dataset used in Search-R1 (Jin et al., 2025), which is publicly available for non-commercial research purposes. It contains two categories of benchmark datasets. The first category is single-hop question answering, which includes NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2022). The second category is multi-hop question answering, which includes HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (2Wiki) (Ho et al., 2020), MuSiQque (Trivedi et al., 2023), and Bamboogle (Press et al., 2023). In this scenario, the agent autonomously generates search queries during step-by-step reasoning. It uses the E5 retriever (Wang et al., 2022) to retrieve relevant documents from a knowledge base, returning the top-3 most relevant passages for each query. The agent then analyzes and reasons over the retrieved information, iteratively refining its queries and incorporating new evidence until it arrives at the final answer.

B.2 Details of Training

Hyperparameters for ALFWorld. For ALFWorld, we follow the default settings used in GiGPO (Feng et al., 2025b). Text-only variants are assigned a maximum prompt length of 5120 tokens. Conversely, optical-history variants are constrained to a maximum prompt length of 2048 tokens. Across both modalities, the maximum response length is standardized to 512 tokens. The agent is permitted to interact with the environment for a maximum of 50 steps per episode, with full history. The learning rate is fixed at 1e-6. Dur-

ing the training phase, rollouts are executed on 16 samples per iteration, generating 8 trajectories per sample. We grant a reward of 10 for successful actions and 0 otherwise. The temperature parameter is set to 1.0 during rollouts and reduced to 0.4 for validation. The mini-batch size is maintained at 256. Consistent with standard practices, no KL-divergence loss is applied during training.

Hyperparameters for search-based QA. For the search-based QA tasks, text-only variants operate with an expanded maximum prompt length of 14000 tokens. For optical-history variants, the limit is set to 4096 tokens. Similar to ALFWorld, the maximum response length is capped at 512 tokens. The interaction horizon is shorter, allowing up to 4 environmental steps per episode and retaining a full history. The learning rate remains at 1e-6. Training rollouts involve 128 samples per iteration, with each sample producing 8 trajectories. The reward structure assigns 1 for correct answers and 0 for incorrect ones. Rollout temperature is set to 1.0, while validation uses a greedy decoding strategy (temperature 0.0). The mini-batch size is 256, and the KL-divergence loss is excluded.

B.3 Computing Details

All experiments utilizing the Qwen2.5-VL-3B-Instruct model for both ALFWorld and search-based QA were executed on 2×H100 GPUs. Conversely, the larger Qwen2.5-VL-7B-Instruct model required a configuration of 4×H100 GPUs. The agents were trained for a total duration of 150 iterations.

B.4 Optical Rendering Details

The optical memory is generated via a deterministic renderer $\mathbf{I}_t = \mathcal{R}(\mathbf{h}_t; \psi)$. We employ specific typographic settings and semantic color codes to facilitate efficient parsing by the VLM. The detailed rendering hyperparameters ψ for both benchmarks are provided in Tab. 5.

B.5 Prompts

The specific prompts employed for agents on the ALFWorld and search-based QA tasks are illustrated in Fig. 4, 5, 6, and 7. Specifically, Fig. 4 and 6 detail the templates for text-only baselines, whereas Fig. 5 and 7 present the templates for optical-history variants.

These templates are constructed using Python-style string formatting, where placeholders in curly

Parameter	ALFWorld	Search
<i>Typography & Layout</i>		
Font Family	Monospace	Monospace
Font Size	10pt	12pt
Line Spacing	1.2	1.2
Max Width	392px	560px
<i>Semantic Color Mapping</i>		
Task & Context	Black	Black
[Observation]	Blue (0,0,255)	–
[Action]	Red (255,0,0)	–
<search>	–	Blue (0,0,255)
<information>	–	Red (255,0,0)

Table 5: Rendering hyperparameters ψ for optical memory generation.

braces ({}) mark semantic slots. For instance, {task_description} denotes the task definition, and {current_observation} indicates the immediate environmental feedback. These slots are populated with dynamic content during training.

To structure the model’s reasoning and outputs, we utilize specific control tags. The <think> </think> tags enclose the mandatory step-by-step reasoning chain. Final decisions are wrapped within <action> </action> tags. In the context of search agents, queries are issued between <search> </search> tags, with retrieved evidence presented inside <information> </information> tags; final answers are enclosed in <answer> </answer>. Uniquely for our AgentOCR method, the model is instructed to specify the compression ratio using <compression> </compression> tags.

C Case Study

We present the complete trajectory of multi-turn interaction between the agent and the search tool, and demonstrate the reasoning behavior of AgentOCR trained by GRPO. As shown in Fig. 8 and Fig. 9, the agent progressively accumulates search results in its optical memory and adaptively adjusts compression factors at each step. The agent successfully retrieves relevant information across multiple search iterations and arrives at the correct answer while maintaining efficient token usage through visual history compression.

Prompt Template of Text Agent on ALFWorld

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.

Figure 4: The prompt template of text agent on ALFWorld.

Prompt Template of AgentOCR on ALFWorld

`<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 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. Additionally, select an image compression factor larger than 1.0 for the next image. Higher compression lowers cost, but too much compression harms image quality. You must provide the next compression factor within `<compression> </compression>` tags (e.g., `<compression>1.1</compression>`).

Figure 5: The prompt template of AgentOCR on ALFWorld.

Prompt Template of Text Agent on Search-based QA

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.

Figure 6: The prompt template of text agent on search-based QA.

Prompt Template of AgentOCR on Search-based QA

<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 you have sufficient information to answer the question with high confidence, provide your final answer within <answer> </answer> tags.

Additionally, select an image compression factor larger than 1.0 for the next image. Higher compression lowers cost, but too much compression harms image quality. You must provide the next compression factor within <compression> </compression> tags (e.g., <compression>1.1</compression>).

Output format:

1. Reasoning: state what you found in the image.
2. <search>...</search> or <answer>...</answer>
3. <compression>...</compression>

Figure 7: The prompt template of AgentOCR on search-based QA.

[illegible]

Figure 8: Case study of AgentOCR on HotpotQA (part I)

