ALDEN: Reinforcement Learning for Active Navigation and Evidence Gathering in Long Documents

Anonymous ACL submission

Abstract

While Vision–language models (VLMs) interpret text-rich images effectively, they struggle with reasoning across long, multi-page documents. We present **A**ctive **L**ong-**D**ocum**E**nt **N**avigation (ALDEN), a multi-turn reinforcement learning framework that fine-tunes VLMs as interactive agents capable of actively navigating long, visually rich documents rather than passive readers. ALDEN features a novel `fetch` action that allows direct page indexing, complementing the classic `search` action and better exploiting document structure. To ensure training efficiency and stability, we introduce a rule-based cross-level reward for dense supervision and a visual-semantic anchoring mechanism utilizing dual-path KL-divergence constraints. We train ALDEN on a curated corpus built from open-source datasets where trivial samples are filtered, and queries are rewritten to incentivize multi-turn navigation and fetch usage. Empirically, ALDEN achieves state-of-the-art results on five long-document benchmarks, offering a more accurate and efficient path for long-document understanding.

1 Introduction

Visually rich documents (VRDs) are fundamental to real-world knowledge storage, as they combine text, tables, and figures within complex, semantically structured layouts. Unlike plain text, understanding VRDs requires joint reasoning over both textual and visual content and structural organization. This has given rise to the task of visually rich document understanding (VRDU) (Wang et al., 2023; Ding et al., 2022) which aims to automate analysis and question answering across these multimodal formats, underpinning critical tasks such as scientific multi-modal question answering (Liang et al., 2024) and key information extraction from business documents (Rombach and Fettke, 2024).

While vision-language models (VLMs) excel on single-page or short documents analysis (Xie et al., 2024; Lv et al., 2023; Hu et al., 2024), they struggle with long documents where full-context processing is computationally prohibitive and noisy (Cho et al., 2024). Current solutions typically adopt Retrieval-Augmented Generation (RAG) pipelines (Cho et al., 2024; Chen et al., 2025a), utilizing retrievers to select query-relevant pages (Faysse et al., 2025) and prompting VLMs to perform fixed subtasks like query reformulation, retrieved content summarization, or final answer synthesis (Han et al., 2025; Wang et al., 2025b). While effective, these systems rely on static reasoning patterns and rigid workflows, limiting their ability to generalize or adapt strategies to diverse user queries. This motivates a shift toward the **Agentic VRDU** (A-VRDU) task, which requires the model to act as an agent that can actively navigate and reason over long documents to deliver accurate and adaptive question answering beyond fixed RAG pipelines.

Recent studies (Chen et al., 2025b; Jin et al., 2025; Song et al., 2025) show that modeling search as an action and optimizing the workflow with outcome-based reinforcement learning (RL) yields more generalizable agents capable of active information gathering. While promising for A-VRDU, adapting this framework to VLMs presents unique challenges. Standard semantic retrieval lacks the precision for queries referencing specific pages or requiring reasoning across consecutive pages. Moreover, document-level information gathering typically demands multi-turn interaction, where sparse and delayed outcome-based rewards fail to reinforce helpful intermediate steps or discourage redundant actions (Li et al.). A further challenge arises from the high-dimensional visual inputs. We empirically observe that fully masking the visual tokens when computing the policy gradient, as done in existing approaches (Jin et al., 2025; Song et al., 2025), leads to unstable training dynamics and can even cause collapse.

These limitations motivate our framework,

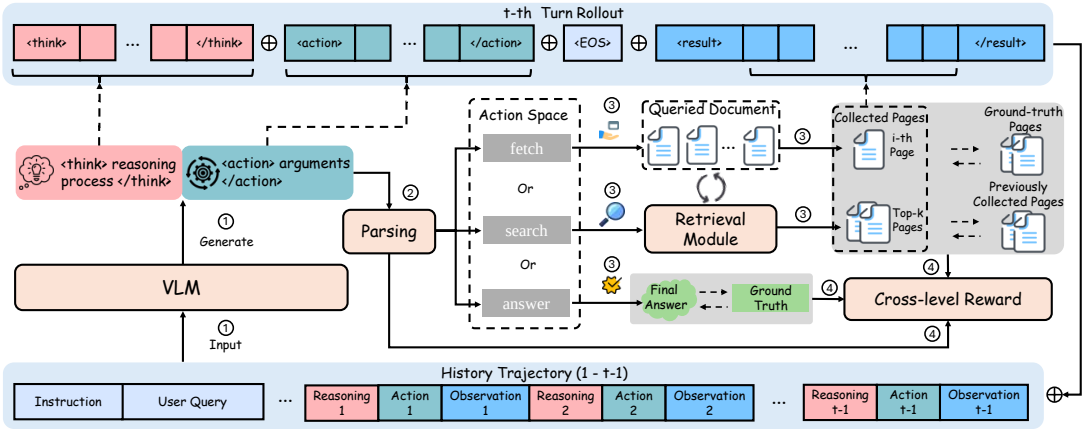

Figure 1: Overview of the rollout process. At each turn: (1) the VLM generates a response conditioned on the dialogue history; (2) the response is parsed into an action (`search`, `fetch`, or `answer`); (3) the action is executed, where `search` or `fetch` collect document pages and `answer` terminates the process; and (4) the cross-level reward function assigns rewards based on execution outcomes and parsing results. $\oplus$ denotes the concatenation operation.

Active **L**ong-**D**ocum**E**nt **N**avigation (ALDEN), a multi-turn RL framework that trains VLMs as interactive agents for navigation in long, visually-rich documents. The overall reasoning-action rollout of ALDEN is illustrated in Fig. 1. We expand the action space with the `fetch` action, enabling direct page access to complement search-based retrieval and efficiently handle diverse queries. To overcome sparse rewards, we incorporate a *cross-level reward function*, which integrates rule-based turn-level supervision with a token-level repetition penalty to provide fine-grained process supervision, encouraging informative evidence collection while discouraging redundant action invoking. Finally, we incorporate a *visual semantic anchoring* mechanism, which constrains the hidden states of generated and visual tokens separately during training to preserve the grounding of visual-token representations and improve overall training robustness.

To train ALDEN, we curate a corpus of 30k samples from DUDE (Van Landeghem et al., 2023), MPDocVQA (Tito et al., 2023a), and SlideVQA (Tanaka et al., 2023a), where we filter trivial documents and rewrite the user queries through LLMs to incentivize multi-turn navigation and the use of fetch actions. Experimental results across five benchmarks demonstrate that ALDEN achieves state-of-the-art performance, significantly outperforming strong baselines. Overall, A-VRDU marks a departure from passive processing toward autonomous navigation, and ALDEN 's performance validates this scalable framework for robust document understanding.

Overall, our main contribution can be summarized as follows:

- We propose the agentic visually-rich document understanding (A-VRDU) task that aims to develop agents that can actively navigate and reason over long visually rich documents.
- To perform the A-VRDU task, we introduce **ALDEN**, a multi-turn RL framework with three key components: an expanded action space featuring a novel `fetch` action, a cross-level reward function, and a visual semantic anchoring mechanism, which together enable efficient and robust training.
- We construct a training corpus for training the A-VRDU agent and conduct extensive experiments on five commonly used VRDU benchmarks, showing that ALDEN significantly outperforms the strongest baseline, improving the answer accuracy by 9.14% on average.

2 Related Work

2.1 Visually-rich documents understanding

Existing VLMs (Hu et al., 2024; Xie et al., 2024; Feng et al., 2024; Liu et al., 2024b) achieve high performance on single-page documents (Mathew et al., 2021; Masry et al., 2022) but face scalability issues with long, multi-page contexts (Deng et al., 2024; Ma et al., 2024b). While semantic retrieval mitigates the computational cost of full-context processing (Tito et al., 2023a; Hu et al., 2024), current methods are limited to passive, prompting-based workflows (Han et al., 2025; Wang et al., 2025b). We advance this paradigm by treating

VRDU as an agentic task (A-VRDU), employing RL to train agents that actively navigate and reason over document structures rather than relying on static retrieval pipelines.

2.2 RL Training for LLMs/VLMs

The success of RLHF (Ziegler et al., 2019; Ouyang et al., 2022) based on the Proximal Policy Optimization (PPO) algorithm (Schulman et al., 2017) and RL with Verifiable Rewards (RLVR) (Shao et al., 2024) based on the Group Relative Policy Optimization (GRPO) algorithm has inspired a new class of "active" RAG agents that optimize retrieval workflows via reinforcement learning (Jin et al., 2025; Song et al., 2025). While concurrent work VRAG-RL (Wang et al., 2025c) applies RL to enhance visual RAG, it focuses on intra-image perception via cropping and zooming. Consequently, fine-tuning VLMs for A-VRDU remains largely unexplored. Unlike open-domain retrieval, A-VRDU requires exploiting explicit document structure (e.g., page indices), denser supervision to guide multi-turn navigation, and stability against the large number of unconstrained visual tokens introduced by high-resolution document pages, which motivates the development of new RL frameworks tailored for this task.

3 Preliminaries

3.1 Problem Formulation

We define A-VRDU as an interactive task where an agent answers a query q_u by navigating a document $\mathcal{D} = (p_1, p_2, \cdots, p_{|\mathcal{D}|})$, composed of a sequence of pages p_i. Since direct access to the full document is restricted, the agent must plan a trajectory of navigational or answering steps to generate a final answer y'. We model this as a Hierarchical Markov Decision Process (MDP) to handle the dual granularity of the task. A **High-Level** MDP manages the discrete action choices at each turn, while a low-level MDP handles the token-by-token realization of those actions. Formally, the high-level MDP operates at the turn level: given a state s_t encapsulating the interaction history, the agent selects a textual action a_t, receives pages as an observation o_t, and a reward r_t. The transition is a deterministic concatenation: $s_{t+1} = [s_t, a_t, o_t]$. The generation of a_t is handled by the **Low-Level** MDP. Here, the state $s_t^i = (s_t, a_t^{1:i-1})$ includes the high-level context and tokens generated so far. The agent selects a token a_t^i from the vocabulary $\mathcal{V}$, receives a reward r_t^i and transitions to $s_t^{i+1} = [s_t^i, a_t^i]$.

3.2 Joint Policy Optimization

The previously defined hierarchical MDP reduces to a standard RLHF formulation when the horizon is collapsed to a single turn. In this setting, the distinction between high-level planning and low-level generation dissolves, and a unified policy π_θ generates the complete token sequence directly from the initial state. This policy is typically optimized using PPO algorithm, which maximizes the expected reward via a clipped surrogate objective:

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}_i \left[\min \left(\rho_i A_1^i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon) A_1^i \right) \right] \quad (1)$$

where ρ_i is the probability ratio between the current and old policies. To balance bias and variance in credit assignment, the advantage term A_1^i is computed using Generalized Advantage Estimation (GAE) (Schulman et al., 2015). This involves estimating the temporal-difference (TD) error $\delta_i = r_1^i + \gamma_{\text{token}} V_\phi(s_1^{i+1}) - V_\phi(s_1^i)$ using a learned value function V_ϕ, and aggregating these errors over the trajectory: $A_1^i = \sum_{k=0}^{L-i-1} (\gamma_{\text{token}} \lambda_{\text{token}})^k \delta_{i+k}$, where γ_{token} is the discount factor, λ_{token} controls the trade-off between bias and variance and L denotes the length of the response. To extend this optimization paradigm from single-turn generation to the multi-turn interaction required for A-VRDU, we propose the ALDEN framework, which is detailed in the following section

4 Methodology

We present **A**ctive **L**ong-**D**ocum**E**nt **N**avigation (ALDEN), a framework for training interactive VLM agents to navigate VRDs via a multi-turn reasoning–action loop. To this end, ALDEN introduces three key components. **(i) Expanded action space:** the agent is equipped with both a semantic `search` action for retrieving pages and a novel `fetch` action for direct page access, enabling flexible exploitation of document structure (§4.1). **(ii) Cross-level reward function:** supervision is provided jointly at the turn level and the token level, guiding the agent toward effective evidence collection and accurate answer generation (§4.2). **(iii) Visual semantic anchoring:** to stabilize RL training, ALDEN constrains the hidden-state evolution of generated and visual tokens separately, mitigating drift and preserving semantic grounding during optimization (§4.3). The overall RL training pipeline of ALDEN is illustrated in Fig. 2 and Alg. 1.

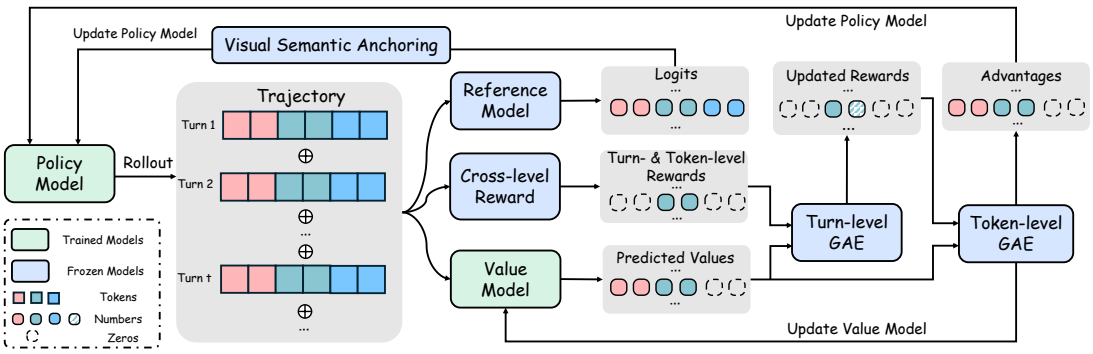

Figure 2: Overview of RL training in ALDEN. The policy model generates multi-turn trajectories, which are scored by a **cross-level reward function** and a **value model**. **Turn-level GAE** integrates future rewards to update the cross-level reward, and **token-level GAE** produces advantages for policy updates. A **reference model** supplies logits for both generated and visual tokens, which the **visual semantic anchoring** mechanism uses to constrain hidden-state evolution during optimization.

4.1 Expanded Action Space

In A-VRDU, agents must navigate using both semantic and structural cues. Standard semantic retrieval struggles with structural dependencies, such as specific page indices (e.g., "see page 12") or sequential reading. ALDEN bridges this gap by introducing a `fetch` action for direct page access, complementing the classic `search` operation. The action space thus consists of three options, each expressed in a structured format that combines free-form reasoning with executable commands as shown in Fig. 1:

- **Search.** The agent generates a reasoning trace within `<think>` tags, followed by a semantic query enclosed in `<search>` tags. An external retrieval module then returns a ranked list of pages based on semantic similarity.
- **Fetch.** Distinct from search, the agent specifies a target page index within `<fetch>` tags to access a page directly, bypassing semantic matching.
- **Answer.** The agent outputs a reasoning trace followed by the final response within `<answer>` tags. This action terminates the rollout and provides the final output for the user query.

Once the action is parsed, the document returns the corresponding page images enclosed within the `<result>` tag. For the `search` action, the associated page numbers are also returned to provide cues of document structure.

4.2 Cross-level Reward Modeling

Since sparse outcome rewards are insufficient for guiding complex navigation, ALDEN introduces a cross-level reward function to provide dense process supervision. This mechanism operates on two levels: turn-level rewards for assessing the strategic utility of actions, and token-level rewards for shaping local generation dynamics.

Turn-level Reward. The immediate turn-level reward r_t is defined as $r_t = f_t + u_t$, comprising a format constraint f_t and a utility score u_t. The format reward f_t is given by:

$$f_t = \begin{cases} 0, & \text{if the format is correct} \\ -1, & \text{otherwise} \end{cases} \tag{2}$$

Thus, only well-formed responses avoid penalty, enforcing structural validity. The utility reward u_t evaluates the action outcome. Let $\mathcal{H}_{t-1}$ denote the history of visited pages, $\mathcal{G}$ denote the ground-truth evidence pages. For a `fetch` action targeting page p_i, the current page set is $\mathcal{C}_t = \{p_i\}$. For a `search` action, the current page set is defined as the collection of the top-K retrieved pages, i.e., $\mathcal{C}_t = \{p_1, \ldots, p_K\}$. The reward is defined as:

$$u_t = \begin{cases} \alpha \cdot \mathrm{F1}(y, y') & \text{if } a_t = \texttt{answer} \\ f_{\text{prox}}(p_i, \mathcal{G}) - \cdot f_{\text{rep}}(\mathcal{C}_t, \mathcal{H}_{t-1}) & \text{if } a_t = \texttt{fetch} \\ |(\mathcal{C}_t \cap \mathcal{G}) \setminus \mathcal{H}_{t-1}| & \text{if } a_t = \texttt{search} \end{cases} \tag{3}$$

where $f\text{prox}(p, \mathcal{G}) = \exp\left(-\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} |p-g|\right)$ rewards the geometric proximity of page p_i to the evidence set $\mathcal{G}$, the character-level F1 score evaluates the character-level overlap between the generated answer and the ground-truth answer. The `fetch` action is subject to a repetition penalty $f_{\text{rep}} = \frac{|\mathcal{C}_t \cap \mathcal{R}|}{|\mathcal{C}_t|}$, penalizing the re-acquisition of known information. We scale the reward of `answer` action by $\alpha > 1$ to ensure answer quality dominates the cumulative intermediate rewards.

While standard RLHF assigns rewards to the final token of a single response, this approach is

insufficient for multi-turn tasks where actions have delayed consequences. We therefore replace r_t with a value estimate $\hat{V}_t$ derived via GAE (Wang et al., 2025a). We first define the turn-level TD error as $\delta_k = r_k + \gamma_{\text{turn}} V_\phi(s^L_{k+1}) - V_\phi(s^L_k)$. The effective reward signal is then computed as: $\hat{V}_t = V_\phi(s^L_t) + \sum_{k=0}^{T-1-t} (\gamma_{\text{turn}}\lambda_{\text{turn}})^k \delta_{t+k}$, where T denotes the total number of turns. By replacing r_t with $\hat{V}_t$, we inject long-horizon strategic information into the token-level optimization.

Token-level Reward. Unlike the atomic `fetch` argument (a single page index), the `search` action generates a multi-token query. Coarse turn-level penalties often fail to identify specific redundant phrases within these queries, leading to inefficient loops. To address this, we introduce a token-level repetition penalty applied specifically to the search query span. For any `search` action after the first, we quantify redundancy by computing the maximum Jaccard similarity between the n-grams of the current query q_t and those of all historical queries $\{q_j\}_{j<t}$:

$$\text{overlap}_t = \max_{j<t} \frac{|Q_n(q_t) \cap Q_n(q_j)|}{|Q_n(q_t) \cup Q_n(q_j)|} \tag{4}$$

where $Q_n(q)$ denotes the set of n-grams in query q. We distribute this penalty to individual tokens to precisely penalize repeated segments. For each token u in the query span a_t^{query}, we assign a weight $w_u = \frac{c_u}{\sum_{v \in a_t^{\text{query}}} c_v}$, where c_u counts the number of overlapping n-grams that contain token u.

Finally, the reward r_t^i assigned to each generated token a_t^i within turn t is defined by combining turn-level and token-level signals:

$$r_t^i = \begin{cases} \hat{V}_t & \text{if } i = L \\ -w_i \cdot \text{overlap}_t & \text{if } t > 1 \wedge a_t = \texttt{search} \\ & \wedge\ a_t^i \in a_t^{\text{query}} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

This yields a unified cross-level signal that penalizes local redundancy without disrupting global credit assignment. Advantages are subsequently derived by applying token-level GAE to this reward stream.

4.3 Visual Semantic Anchoring

While existing studies generally mask observation tokens to isolate the generative action space (Jin et al., 2025; Song et al., 2025), this approach proves insufficient for A-VRDU. Document pages introduce a massive volume of visual tokens, creating a fundamental vulnerability, i.e., an under-constrained visual manifold. Since the mapping from the visual latent space to the textual action space is non-injective, relying solely on action-space KL divergence leaves a high-dimensional null space in the optimization landscape (verified formally in Appx. B). Consequently, aggressive reward-driven gradients can distort visual representations without triggering the trust region penalty, leading to severe training instability and rapid entropy collapse (see Fig. 3).

To remedy this, we propose Visual Semantic Anchoring, a stabilization mechanism that introduces dual-path KL regularization. Beyond the standard textual constraint, we impose a secondary KL penalty on the visual hidden states. This enforces a smoothness constraint on the visual encoder, preserving Lipschitz continuity and maintaining the robustness of pre-trained features against drastic policy updates. Formally, we define:

$$\begin{aligned} \mathcal{L}_{\text{policy}} = \mathbb{E}_t[\mathbb{E}_i[\min\Big[\rho_t^i A_t^i, \text{clip}\left(\rho_t^i, 1-\epsilon, 1+\epsilon\right) A_t^i\Big] \\ +\beta_{\text{gen}} KL(\pi_\theta(a_t^i|s_t^i)||\pi_{\text{ref}}(a_t^i|s_t^i))] \\ +\mathbb{E}_j[\beta_{\text{obs}} KL(\pi_\theta(o_t^j|o_t^{<j}, a_t, s_t)||\pi_{\text{ref}}(o_t^j|o_t^{<j}, a_t, s_t))]] \end{aligned} \tag{6}$$

where β_{gen} and β_{obs} are independent coefficients. In practice, we set $\beta_{\text{obs}} > \beta_{\text{gen}}$ to tightly regularize the much larger observation-token set while allowing more flexibility for generated tokens to adapt to the task.

5 Experiments

We conduct experiments on long VRDU benchmarks to (i) compare ALDEN with strong baselines and (ii) assess the contribution of its key components, including expanded action space, cross-level reward, and visual semantic anchoring, to navigation accuracy, answer quality, and training stability. We first outline datasets, baselines, implementation details, and evaluation metrics (§5.1), then present main results (§5.2), followed by ablations (§5.3) and detailed component analyses (§5.4).

5.1 Experimental Setup

Datasets. We build a challenging training corpus by filtering DUDE (Van Landeghem et al., 2023), MPDocVQA (Tito et al., 2023b), and SlideVQA (Tanaka et al., 2023b) for documents exceeding 3 pages. To enrich query diversity, we use GPT-4o (Hurst et al., 2024) to rewrite part of

MPDocVQA, increasing the proportion of page-index–referenced queries in the final training corpus. The evaluation is conducted mainly on the following VRDU benchmarks: **MMLongBench** (Ma et al., 2024b), **LongDocURL** (Deng et al., 2024), **PaperTab** (Hui et al., 2024), **PaperText** (Hui et al., 2024), and **FetaTab** (Hui et al., 2024). To validate the `fetch` mechanism, we introduce DUDE-sub, a balanced validation set comprising 960 queries that contain both general and structure-dependent questions (e.g., sequential cues). More details about the dataset can be seen in Appx. A.

Baselines. To validate ALDEN's effectiveness, we compare it with three categories of baselines. (1) **Full-Document Input**: SoTA VLMs prompted with the entire document as context to answer user queries. (2) **Visual RAG**: methods that retrieve the most relevant document pages using the user query, including M3DocRAG (Cho et al., 2024), and an ALDEN variant trained with GRPO adapted from a fully textual method ReSearch (Chen et al., 2025b). (3) **Hybrid RAG**: approaches that augment page images with OCR-extracted text for retrieval and reasoning, including MDocAgent (Han et al., 2025), VidoRAG (Wang et al., 2025b). More details are presented in Appx. C.

Implementation Details. Both the policy and value models are initialized from Qwen2.5-VL-7B-Instruct (Bai et al., 2025), and all Visual RAG and Hybrid RAG baselines use the same backbone for fairness. During training, we adopt the single-vector retriever vdr-2b-v1 (Ma et al., 2024a) for images and e5-large-v2 (Wang et al., 2022) for text. For evaluation, we also report results with the multi-vector retrievers ColQwen2-v1.0 (ColQwen) (Faysse et al., 2025) for images and ColBERT-v2.0 (ColBERT) (Santhanam et al., 2021) for text. Unless otherwise noted, each `search` action retrieves the top-1 candidate page, with a maximum of $T = 6$ reasoning–action turns. On average, ALDEN collects 1.87 unique pages per query; hence, single-turn RAG baselines are set to retrieve the top-2 pages for a fair comparison. Further implementation details are provided in Appx. D.

Evaluation Metrics. The primary evaluation metric is GPT-4o–judged answer accuracy (**Acc**) on each benchmark. For finer-grained analysis of ALDEN's components, we further assess navigation quality using trajectory-level retrieval recall (**Rec**), precision (**Pre**), F1-score (**F1**), and the number of unique collected pages (**#UP**). Detailed definitions of these metrics are provided in Appx. E.

5.2 Main Results

Table 1 reports answer accuracy across all baselines. Directly prompting large VLMs with the entire document performs poorly (Acc < 0.30), confirming the difficulty of long-document reasoning where irrelevant content overwhelms true evidence. Retrieval-based methods achieve substantially better results. Among Visual RAG approaches, ALDEN with ColQwen attains the highest average accuracy (0.410), surpassing M3DocRAG by 3.2 points. In Hybrid RAG, baselines such as VidoRAG and MDocAgent benefit from textual signals but are limited by fixed reasoning pipelines. ALDEN with hybrid retrievers achieves the best overall performance, exceeding the strongest hybrid baseline by +7.47% relative improvement. These results highlight ALDEN 's ability to generalize across benchmarks by actively collecting and reasoning over evidence, though modest performance on scientific-paper datasets (PaperText, PaperTab) suggests domain knowledge remains a limiting factor. The notably larger gain over GRPO underscores the limitations the limitation of outcome-based rewards for training multimodal agents in multi-turn, long-horizon settings from base VLMs, which is one of the key motivations for this work. Moreover, ALDEN achieves higher accuracy with a multi-vector retriever at inference despite being trained with a single-vector retriever, indicating that strategies learned with a weaker retriever generalize to stronger ones and suggesting a path to more efficient training. A case study demonstrating learned action patterns is provided in Appx. F.

5.3 Ablation Study

To understand the contribution of each component in ALDEN, we further conduct ablation studies on the five benchmarks. Table 2 reports the Acc metric results for the full model and three variants: (i) *w/o Fetch*, which removes the index-based `fetch` action and relies solely on semantic retrieval; (ii) *w/o Cross-level Reward*, which uses only outcome-level supervision without our designed turn- and token-level reward shaping; and (iii) *w/o Visual Semantic Anchoring*, which omits the constraint on visual hidden states during optimization. Removing any component consistently lowers accuracy, with the largest drop from omitting `fetch`, underscoring the value of direct page-index access. Excluding the cross-level reward also substantially

Table 1: Answer accuracy comparison on five VRDU benchmarks. † indicates the strongest non-ALDEN baseline used to compute the relative improvement (%). **Bold** indicates the best result per dataset.

Method	MMLongBench	LongDocUrl	PaperTab	PaperText	FetaTab	Avg
Full Document Input						
SmolVLM-Instruct (Marafioti et al.)	0.072	0.165	0.065	0.142	0.148	0.118
Phi-3.5-Vision-Instruct (Abdin et al.)	0.141	0.285	0.068	0.174	0.232	0.180
mPLUG-DocOwl2 (Hu et al.)	0.159	0.273	0.072	0.162	0.288	0.191
Qwen2-VL-7B-Instruct (Wang et al.)	0.177	0.280	0.077	0.146	0.339	0.203
LEOPARD (Jia et al.)	0.196	0.313	0.112	0.189	0.341	0.230
Qwen2.5-VL-7B-Instruct (Bai et al.)	0.221	0.375	0.131	0.265	0.336	0.265
InternVL3.5-8B-Instruct (Wang et al.)	0.219	0.381	0.130	0.271	0.348	0.270
Visual RAG methods						
GRPO (ColQwen)	0.274	0.384	0.150	0.295	0.406	0.302
M3DocRAG (ColQwen)†	0.330	0.464	0.201	0.350	0.547	0.378
ALDEN (vdr-2b-v1)	0.335	0.513	0.201	0.342	0.542	0.386
ALDEN (ColQwen)	0.367	0.526	0.211	0.345	0.603	0.410
Relative Improvement (%)	11.21	13.36	4.98	-1.43	10.23	10.81
Hybrid RAG methods						
ViDoRAG (ColQwen + ColBERT)	0.215	0.323	0.158	0.264	0.358	0.264
MDocAgent (ColQwen + ColBERT)†	0.347	0.494	0.221	0.408	0.607	0.415
ALDEN (vdr-2b-v1 + e5-large-v2)	0.385	0.542	0.228	0.416	0.611	0.436
ALDEN (ColQwen + ColBERT)	**0.392**	**0.551**	**0.245**	**0.421**	**0.623**	**0.446**
Relative Improvement (%)	12.97	11.54	10.86	3.18	2.63	7.47

Table 2: Answer accuracy for different ablations of ALDEN on five VRDU benchmarks. **Bold** indicates the best result per dataset.

Method	MMLongBench	LongDocUrl	PaperTab	PaperText	FetaTab	Avg
Full ALDEN	**0.335**	**0.513**	**0.201**	**0.342**	**0.542**	**0.386**
w/o Fetch	0.301	0.469	0.140	0.258	0.443	0.322
w/o Cross-level Reward	0.329	0.483	0.148	0.301	0.518	0.356
w/o Visual Semantic Anchoring	0.326	0.502	0.181	0.328	0.529	0.373

hurts performance, confirming the importance of fine-grained reward shaping. In contrast, removing visual-semantic anchoring causes milder yet consistent degradation. Building on these results, we next provide a detailed component analysis to understand the specific roles of each key design choice in ALDEN.

5.4 Component Analysis

Fetch vs. Search To assess the effect of the proposed `fetch` action, we compare the full ALDEN agent with a *search-only* variant that disables direct page-index access and relies solely on semantic retrieval. Evaluation on the DUDE-sub dataset, which contains explicit page references and structured navigation queries, shows clear benefits of `fetch` (Table 3). Acc improves from 0.545 to 0.653 and Rec from 0.471 to 0.598, while Pre and F1 also increase, indicating more accurate evidence retrieval. The number of unique pages rises from 1.03 to 1.19, reflecting broader coverage. These results confirm that combining index-based `fetch` with semantic `search` enables more flexible and efficient navigation, especially for queries that reference specific pages or require traversal across consecutive pages.

Table 3: Comparison between `search`-only and full ALDEN on the DUDE-sub dataset.

Method	Acc	Rec	Pre	F1	#UP
Search-only	0.545	0.471	0.841	0.531	1.03
Full ALDEN	**0.653**	**0.598**	**0.874**	**0.628**	**1.19**

Table 4: Effect of reward design of outcome-based, turn-level and outcome-based, and full ALDEN on LongDocURL.

Method	Acc	Rec	Pre	F1	#UP
Outcome-based Only	0.483	0.483	0.612	0.520	1.27
Turn-level + Outcome	0.509	0.497	0.608	0.522	1.22
Full ALDEN	**0.513**	**0.506**	**0.612**	**0.526**	**1.39**

Effect of Reward Design. We evaluate how different reward schemes affect ALDEN 's retrieval and reasoning (Table 4). (i) Outcome-based Only assigns a single scalar reward for final answer correctness. (ii) Turn-level + Outcome adds rule-based turn-level supervision, improving Acc from 0.483 to 0.509 and Rec from 0.483 to 0.497, showing that denser feedback aids evidence localization. (iii)

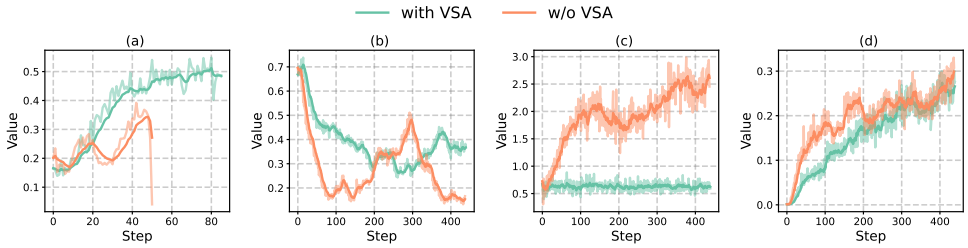

Figure 3: Training dynamics of ALDEN with and without Visual Semantic Anchoring (VSA). Panel (a) shows the turn-level reward of the `answer` action, panel (b) shows token-level entropy, panels (c) and (d) plot the KL divergence of visual tokens and generated tokens, respectively.

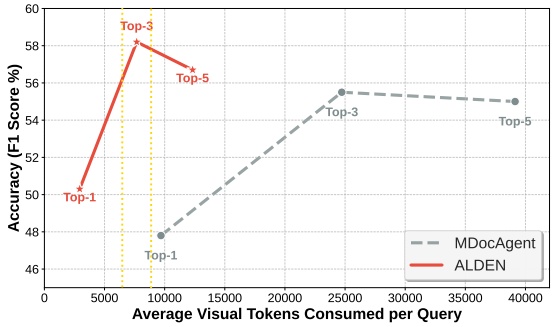

Figure 4: Efficiency and accuracy comparison between ALDEN and the MDocAgent baseline.

Full ALDEN further introduces token-level shaping, yielding a smaller but consistent gain (Acc 0.513, Rec 0.506) and increasing unique pages from 1.22 to 1.39, indicating reduced query repetition and broader exploration. Overall, the cross-level reward design fosters richer query reformulation and more thorough evidence gathering, enhancing both navigation and answer quality.

Effect of Visual Semantic Anchoring. We evaluate the effect of Visual Semantic Anchoring (VSA) on training stability and representation drift, as shown in Figure 3. With a larger batch size (512) than in the main experiments (128), the VSA-enabled model achieves steadily increasing `answer` rewards, while the non-VSA variant fluctuates and collapses (a). VSA also maintains higher policy entropy, supporting healthier exploration (b). Besides, KL divergence of visual tokens grows unchecked without VSA, indicating hidden-state drift, whereas VSA constrains these values while allowing moderate growth for action tokens (c,d). Overall, VSA achieves stabilizing RL training and preventing drift in visual representations.

5.5 Efficiency Analysis

We investigate the token efficiency of ALDEN on the LongDocURL dataset. Specifically, ALDEN is trained using the top-3 retrieved results as observations. During evaluation, we broaden the scope to compare ALDEN against the strongest baseline, MDocAgent, across top-1, 3, and 5 retrieval settings. For a fair comparison, both methods utilize only query-image retrieval. As illustrated in Fig. 4, the baseline relies on expanding the retrieval scope to Top-5 for performance gains, incurring a linear cost penalty. In contrast, ALDEN reaches saturation at Top-3 by effectively identifying signals within noisy contexts. This results in a Pareto improvement, surpassing the baseline's best configuration (+3.2% accuracy) with an approximate $3\times$ reduction in token usage. This confirms that equipping the model with agentic reasoning is a fundamentally more efficient strategy than the passive reading, even when the latter is enhanced by multi-agent mechanisms.

6 Conclusions

We introduced the **Agentic VRDU** task and proposed **ALDEN**, a reinforcement-learning framework that trains VLMs as autonomous agents capable of multi-turn navigation and evidence gathering. The framework integrates a `fetch` action for direct page access, a cross-level reward for fine-grained reward modeling, and a visual semantic anchoring mechanism for stable training. Extensive experiments on multiple long-document benchmarks demonstrate that ALDEN achieves state-of-the-art accuracy while maintaining significantly lower token consumption. Furthermore, ablation studies validate the critical role of the `fetch` action in evidence localization and the effectiveness of visual semantic anchoring in stabilizing the training process, offering broader insights for multi-turn RL in multimodal agents. Ultimately, the A-VRDU paradigm marks a fundamental shift from passive document reading to autonomous navigation across vast information landscapes. The robust performance of ALDEN highlights the potential of such agents to deliver scalable, adaptive, and accurate understanding of complex, visually rich documents.

Limitations

Despite the promising results, ALDEN has several limitations. First, the agent still faces challenges in optimally balancing exploration and exploitation within large document spaces; it may occasionally terminate the search prematurely or navigate redundantly when the target information is buried deep in the document. Second, identifying true evidence pages remains non-trivial, as the agent can sometimes be misled by pages that are visually or semantically similar to the target but lack the precise answer.

Future work could address these issues by constructing larger-scale datasets with high-quality trajectory annotations to improve sample efficiency. Additionally, we plan to leverage trajectories from stronger, closed-source models to guide training, integrating validation and reflection mechanisms to reduce hallucinations. Finally, adopting curriculum learning—starting from shorter documents and progressively moving to complex ones—could help the agent better generalize across tasks of varying difficulty.

LLM Usage Statement

Large Language Models (LLMs) were used as general-purpose writing and editing aids. Specifically, OpenAI's ChatGPT (GPT-5) assisted in polishing grammar, improving clarity, and suggesting alternative phrasings. All research ideas, experimental design, data processing, model development, and analysis were conceived and executed solely by the authors. The LLM provided no novel research insights or substantive scientific contributions.

Reproducibility Statement

We are committed to ensuring the reproducibility of our results. To this end, we will release:

- All source code for training, evaluation, and data preprocessing, including scripts for dataset construction, reward computation, and reinforcement-learning training with ALDEN.
- The processed training corpus derived from DUDE, MPDocVQA, and SlideVQA, along with instructions to regenerate it from the original public datasets.
- Detailed configuration files specifying model hyperparameters, random seeds, and hardware settings.
- Checkpoints for both the policy and value models, and prompts used for GPT-4o evaluation.

Our experiments were run on NVIDIA A100 GPUs (80GB) with PyTorch 2.4 and HuggingFace Transformers 4.49; exact package versions will be provided in the released code. These resources will allow other researchers to fully reproduce our training, evaluation, and analysis results.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jian Chen, Ruiyi Zhang, Yufan Zhou, Tong Yu, Franck Dernoncourt, Jiuxiang Gu, Ryan A. Rossi, Changyou Chen, and Tong Sun. 2025a. SV-RAG: LoRA-contextualizing adaptation of MLLMs for long document understanding. In *The Thirteenth International Conference on Learning Representations*.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, Fan Yang, and 1 others. 2025b. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multimodal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, and 1 others. 2024. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. *arXiv preprint arXiv:2412.18424*.

Yihao Ding, Zhe Huang, Runlin Wang, YanHang Zhang, Xianru Chen, Yuzhong Ma, Hyunsuk Chung, and Soyeon Caren Han. 2022. V-doc: Visual questions answers with documents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21492–21498.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. 2025. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Hao Feng, Qi Liu, Hao Liu, Jingqun Tang, Wengang Zhou, Houqiang Li, and Can Huang. 2024. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *Science China Information Sciences*, 67(12):1–14.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*.

Yulong Hui, Yao Lu, and Huanchen Zhang. 2024. Uda: A benchmark suite for retrieval augmented generation in real-world document analysis. *Advances in Neural Information Processing Systems*, 37:67200–67217.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Mengzhao Jia, Wenhao Yu, Kaixin Ma, Tianqing Fang, Zhihan Zhang, Siru Ouyang, Hongming Zhang, Dong Yu, and Meng Jiang. 2024. Leopard: A vision language model for text-rich multi-image tasks. *arXiv preprint arXiv:2410.01744*.

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*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Xiangyang Li, Kuicai Dong, Yichao Wang, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Derong Xu, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and 1 others. Process vs. outcome reward: Which is better for agentic rag reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. 2024. SceMQA: A scientific college entrance level multimodal question answering benchmark. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 109–119, Bangkok, Thailand. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024b. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*.

Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, and 1 others. 2023. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024a. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, Miami, Florida, USA. Association for Computational Linguistics.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024b. MMLONGBENCH-DOC: Benchmarking long-context document understanding with visualizations. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve noyan, Elie Bakouch, Pedro Manuel Cuenca Jiménez, Cyril Zakka, Loubna Ben allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. 2025. SmolVLM: Redefining small and efficient multimodal models. In *Second Conference on Language Modeling*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

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. *Advances in neural information processing systems*, 35:27730–27744.

Alexander Michael Rombach and Peter Fettke. 2024. Deep learning based key information extraction from business documents: Systematic literature review. *ACM Computing Surveys*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023a. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023b. Slidevqa: a dataset for document visual question answering on multiple images. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023a. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023b. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recogn.*, 144(C).

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, and 1 others. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540.

Kangrui Wang, Pingyue Zhang, Zihan Wang, Yaning Gao*, Linjie Li, Qineng Wang, Hanyang Chen, Chi Wan, Yiping Lu, Zhengyuan Yang, Lijuan Wang, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, Yejin Choi, and Manling Li. 2025a. Reinforcing visual state reasoning for multi-turn vlm agents.

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*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025b. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. 2025c. VRAG-RL: Empower vision-perception-based RAG for visually rich information understanding via iterative reasoning with reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025d. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2023. Vrdu: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5184–5193.

Xudong Xie, Hao Yan, Liang Yin, Yang Liu, Jing Ding, Minghui Liao, Yuliang Liu, Wei Chen, and Xiang Bai. 2024. Wukong: A large multimodal model for efficient long pdf reading with end-to-end sparse sampling. *arXiv preprint arXiv:2410.05970*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Datasets

A.1 Training dataset

Training. We construct our training dataset by combining samples from three publicly

Table 5: Statistics of the training dataset. #GQ and #PQ denote the numbers of general user queries and page-index–referenced queries, respectively.

Sub-dataset	DUDE	SlideVQA	MPDocVQA
#GQ	6,943	10,615	7,992
#PQ	1,011	2	4,165
Sum	7,954	10,617	12,157

available multi-page document understanding datasets: DUDE (Van Landeghem et al., 2023), MPDocVQA (Tito et al., 2023b), and SlideVQA (Tanaka et al., 2023b). These datasets provide diverse document layouts and question-answering formats, making them well-suited for training models on complex multi-turn document question answering tasks.

DUDE is a large-scale benchmark designed for multi-page, visually rich document understanding. It covers diverse domains such as scientific articles, financial and legal reports, technical manuals, and presentations. Each example consists of a full PDF document rendered into page images, paired with a natural-language query and a free-form textual answer, along with page-level ground-truth evidence annotations. SlideVQA contains questions grounded in slide decks, where understanding layout and inter-slide referencing is crucial. It contains slide decks from diverse topics such as education, business, and research talks, requiring models to reason across sequential pages that mix text, charts, and images. Each example provides a slide deck rendered as ordered page images, a natural-language question, and a free-form textual answer, with annotations of relevant slides for evidence grounding. MPDocVQA extends the traditional single-page VQA setting (originally based on DocVQA) by concatenating additional pages to the original single-page input, while retaining the same set of user questions. However, since many of these questions were authored under the assumption that only one page is visible (e.g., “What is the date?" or “Who is the author?"), they often lack sufficient context to guide document retrieval or navigation. To address this, we first use GPT-4o (Hurst et al., 2024) to automatically identify this kind of samples. Then we integrate the index of referred pages into the questions to get page-index-referenced questions, e.g., “In page 5, what is the date?". The prompt we used is shown below:

Prompt for Filtering Queries

You are given a question from a multi-page document VQA dataset. Some questions are not suitable for training an agent to autonomously locate the target page, because they assume the agent already knows which page is relevant. These questions are often vague, layout-based, or refer to elements only visible on a known page (e.g., "What is the PVR no given in the approval sheet?", or "What is written at the top right?"). Your task is to assign a label to each question:
- 1 if the question belongs to this kind of problem, i.e., it assumes the correct page is known and cannot be answered without it.
- 0 if the question does not belong to this kind of problem, i.e., it can be answered after locating the page based on content in the question.
Respond with a JSON object containing only the field "label". Examples:
Question: What is the PVR no given in the approval sheet? Answer: { "label": 1 }
Question: What is the project name mentioned in the title block? Answer: { "label": 0 }
Question: What is the symposium organized by Division of Agricultural and Food Chemistry? Answer: { "label": 0 }
Question: What is written on the top right corner? Answer: { "label": 1 }
Question: What is the page number? Answer: { "label": 1 }
Question: What is the Date? Answer: { "label": 1 }
Now, label the following question:
Question: {question}

To ensure that our model is consistently exposed to multi-page reasoning scenarios, we additionally discard any documents with fewer than 10 pages from all three datasets. This helps avoid biasing the model toward short-context behavior and ensures a consistent level of document complexity.

After merging and filtering, we obtain a training set consisting of 30,728 samples, each comprising a user query and its corresponding multi-page document context, answer and the index of evidence pages. Finally, we proportionally sample 1,024 samples from the validation set of these three datasets as our validation set.

A.2 Benchmarks

We evaluate our method on a diverse set of benchmarks: MMLongBench (Ma et al., 2024b), LongDocURL (Deng et al., 2024), PaperTab (Hui et al., 2024), PaperText (Hui et al., 2024), and FetaTab (Hui et al., 2024). These datasets span a wide range of scenarios, including both open-domain and closed-domain tasks, and include textual as well as visual content. The documents also vary in length and structure, ranging from short forms to complex, multi-page documents. This diversity ensures a comprehensive and fair evaluation of our model's performance across real-world document understanding tasks.

- **MMLongBench-Doc** is a large-scale benchmark designed to evaluate how multimodal large language models handle long, visually rich documents. It contains over a thousand expert-annotated questions drawn from lengthy PDFs (averaging 50 pages and 20k tokens) that mix text, tables, charts, and images. Tasks require single-page, cross-page, and sometimes unanswerable reasoning, testing a model's ability to retrieve and integrate evidence across multiple modalities and extended contexts.

- **LongDocURL** is a benchmark for evaluating large vision-language models on long, multimodal documents by combining three core task types: understanding, numerical reasoning, and element locating. It includes 2,325 high-quality question-answer pairs over 396 documents totaling over 33,000 pages, with an average of 85.6 pages per document. Tasks vary in their evidence requirements: some require single-page evidence, others multi-page, and many involve locating evidence across different layout elements (text, tables, figures, and layout).

- **PaperText** is a subset in the UDA benchmark made up of academic papers (in PDF form) used for retrieval-augmented generation / document question answering tasks. Each document comes with multiple question-answer pairs drawn from "Qasper" (an academic paper reading comprehension dataset), where questions may be extractive, yes/no, or free-form. The dataset preserves full documents to allow answering from context, rather than just small passages.

- **PaperTab** is another subset in UDA also based on academic papers, but the focus is on Q&A pairs where evidence comes from or interacts with tables inside papers. Like PaperText, it retains full PDF documents so that models must locate and reason over tabular content, as well as textual content. The questions are similarly diverse (extractive, yes/no, free-form), and the average size is modest (10–11 pages per document).

- **FetaTab** is a subset of the UDA (Unstructured Document Analysis) benchmark that focuses on free-form question answering over Wikipedia tables in both HTML and PDF formats. It comprises 878 documents and 1,023 QA pairs, averaging about 14.9 pages per document. The questions are "free-form" (i.e. natural language answers, not limited to extractive spans or simple yes/no), which requires models to understand table content, context, and sometimes cross-format layout.

B Derivation about the Visual Semantic Anchoring

Let the input to the VLMs head be a concatenation of visual tokens $Z_v \in \mathbb{R}^{N_v \times d}$ and textual tokens $Z_t \in \mathbb{R}^{N_t \times d}$. The policy output (logits) is a function $f : \mathbb{R}^{(N_v+N_t)\times d} \to \mathbb{R}^V$. We examine the update of the visual encoder parameters ϕ, which solely influence Z_v.

Our verification starts with the Block-Jacobian Decomposition. The linearization of the change in logits Δy with respect to the input features is given by the total derivative. Since the input is concatenated $X = [Z_v, Z_t]$, the Jacobian J_f can be decomposed into two block matrices:

$$J_f = \left[\frac{\partial f}{\partial Z_v} \;\middle|\; \frac{\partial f}{\partial Z_t} \right] = [J_v \mid J_t]$$

where $J_v \in \mathbb{R}^{V \times (N_v \cdot d)}$ is the Partial Jacobian with respect to visual features.

During the PPO update, we are concerned with the gradients flowing into the visual encoder parameters ϕ. These parameters only affect Z_v. The textual tokens Z_t are either fixed (from history) or updated by separate parameters (LLM weights). Thus, relative to the visual encoder update, $\Delta Z_t = 0$. The constraint imposed by the text-only KL divergence ($||\Delta y|| < \epsilon$) simplifies to:

$$\begin{aligned} \Delta y &\approx J_v \cdot \mathrm{vec}(\Delta Z_v) + J_t \cdot \mathbf{0} \\ &\approx J_v \cdot \mathrm{vec}(\Delta Z_v) \end{aligned} \tag{7}$$

The "freedom" available to the visual encoder to drift without triggering the KL penalty is determined solely by the null space of the Partial Jacobian J_v.The dimensionality of this subspace is:

$$\dim(\text{null}(J_v)) = (N_v \cdot d) - \text{rank}(J_v)$$

Even with the presence of Z_t, the rank of J_v is still upper-bounded by the bottleneck dimension of the model (or vocabulary size V). It does not gain rank from the text tokens. Thus the available "drift space" scales linearly with the number of visual tokens:

$$\dim(\text{null}(J_v)) \propto N_v$$

The Null Space represents the "freedom" the visual encoder has to change its weights (ΔZ) without violating the text-KL constraint. As the number of visual tokens N increases, the dimensionality of this unconstrained subspace grows. Consequently, the optimizer has significantly more degrees of freedom to introduce aggressive, potentially destructive updates to the visual representations that satisfy the short-term reward while remaining "invisible" to the text-only KL penalty. This confirms that VLMs with higher visual token counts are mathematically more susceptible to representation drift when lacking visual-semantic anchoring.

C Baselines

To evaluate the effectiveness of ALDEN, we compare it against three categories of methods:

- **Base VLMs supporting multi-image input.** These models directly take the entire multi-page document as context without retrieval, leveraging their built-in multi-page visual processing capabilities. For fairness, we select open-source VLMs of similar scale to Qwen2.5-VL-7B, including LLaVA-v1.6-Mistral-7B (Liu et al., 2024a), Phi-3.5-Vision-Instruct (Abdin et al., 2024), LLaVA-One-Vision-7B (Li et al., 2024), SmolVLM-Instruct (Marafioti et al., 2025), mPLUG-DocOwl2 (Hu et al., 2024), LEOPARD (Jia et al., 2024), InternVL3.5-8B-Instruct (Wang et al., 2025d).

- **Visual RAG methods.** These methods use the user query to retrieve the most relevant document pages and feed them into the model as context. We include M3DocRAG (Cho et al., 2024) as a strong baseline, as well as our proposed ALDEN. To isolate the impact of our reward function design, we additionally evaluate a variant that trains the same backbone with GRPO using only outcome-based rewards (no turn-level shaping), mirroring common text-only RLHF setups as in ReSearch (Chen et al., 2025b). Specifically,

  - M3DocRAG is a multi-modal document understanding framework designed for multi-page and multi-document question answering. It first encodes each page into joint visual-text embeddings using a multi-modal encoder, then retrieves the top-K relevant pages via a MaxSim-based retrieval mechanism, optionally accelerated with FAISS for large-scale documents. Finally, a multi-modal language model processes the retrieved pages to generate precise answers, effectively handling complex queries that require reasoning over both textual and visual content.
  - ReSearch introduces a framework that trains large language models to integrate reasoning and search in a unified process. The model learns, via reinforcement learning, when and how to perform search actions during multi-step reasoning, using search results to guide subsequent reasoning steps. By treating search as part of the reasoning chain, ReSearch enables LLMs to solve complex multi-hop tasks, demonstrate self-correction and reflection, and generalize effectively across benchmarks, achieving significant performance gains over baseline models.

- **Hybrid RAG methods.** These approaches combine visual and textual retrieval by first applying an OCR tool to extract all text from the document. The query is then used to retrieve both the most relevant page image and the most relevant OCR-extracted text, which are jointly fed into the model. We evaluate MDocAgent (Han et al., 2025) and VidoRAG (Wang et al., 2025b) as a representative method in this category.

  - MDocAgent is a multi-modal, multi-agent framework for document understanding that combines Retrieval-Augmented Generation (RAG) with specialized agents to handle complex documents. The system employs a General Agent for multi-modal context retrieval, a Critical Agent for identifying

key information, a Text Agent for analyzing textual content, an Image Agent for interpreting visual elements, and a Summarizing Agent to synthesize results. By coordinating these agents, MDocAgent effectively integrates textual and visual reasoning, achieving significant improvements in accuracy and error reduction compared to existing large vision-language models and RAG-based methods. For all five agents in this framework, we consistently use the original LLaMA3.1-8B as the LLM for the text agent, while employing a consistent VLMs, i.e., Qwen2.5-VL-7B, for remaining agents.

– ViDoRAG is a multi-agent framework designed to enhance the understanding of visually rich documents. It employs a Gaussian Mixture Model (GMM)-based hybrid retrieval strategy to effectively handle multi-modal retrieval, integrating both textual and visual information. The framework incorporates a dynamic iterative reasoning process, utilizing agents such as Seeker, Inspector, and Answer to iteratively refine the understanding and generation of responses. This approach addresses challenges in traditional Retrieval-Augmented Generation (RAG) methods by improving retrieval accuracy and enabling complex reasoning over visual documents. We use Qwen2.5-VL-7B as backbone for all agents in this methods.

D Implementation Details

Our implementation is based on the EasyR1[1] framework. We adopt the default optimization hyperparameters from the EasyR1 framework, specifically the learning rates and KL coefficients, to ensure training stability. Both the policy model and the value function are initialized from Qwen2.5-VL-7B-Instruct (Bai et al., 2025). We use a batch size of 128, with fixed learning rates of 1×10^{-6} for the policy model and 1×10^{-5} for the value function. The maximum number of interaction turns is set to $T = 6$. For visual inputs, we constrain the number of image pixels to lie between 261,070 and 2,508,800. Based on these settings, we set the maximum number of tokens in the trajectory as 19000. The KL coefficients for generated tokens and observation tokens are set to $\beta_{\text{gen}} = 0.001$ and $\beta_{\text{obs}} = 0.01$, respectively. For the `search` actions, we used only the top-1 retrieved pages. Besides, we set the scale coefficient $\alpha = 5$. The weight of repetition penalty is set as $\eta = 0.5$. For the calculation of GAE, we set $\gamma_{\text{token}} = 1.0$, $\gamma_{\text{turn}} = 0.9$ and $\lambda_{\text{token}} = \lambda_{\text{turn}} = 1.0$. During training, we adopt the single-vector retriever vdr-2b-v1 (Ma et al., 2024a) for images and e5-large-v2 (Wang et al., 2022) for text for training efficiency. For evaluation, we also report results with the multi-vector retrievers ColQwen2-v1.0 (ColQwen) (Faysse et al., 2025) for images and ColBERT-v2.0 (ColBERT) (Santhanam et al., 2021) for text. All experiments are conducted on 16 NVIDIA A100-80Gb GPUs.

The system prompt that we used during training of Visual RAG variant of ALDEN is shown in Fig. 5.

The system prompt that we used during training of Hybrid RAG variant of ALDEN is shown in Fig. 5.

[1] https://github.com/hiyouga/EasyR1

E Evaluation Metrics

We evaluate models using both answer quality and intermediate navigation metrics.

Model-based Accuracy (Acc). Answer quality is assessed with an LLM-as-judge protocol. Given a predicted answer and the ground-truth reference, GPT-4o is prompted to classify the prediction as *Correct*, *Incorrect*, or *Tie/Unclear*. We compute accuracy for each benchmark as the percentage of responses judged *Correct* over all responses:

$$\text{Acc} = \frac{\#\text{Correct}}{N}, \tag{8}$$

where N is the number of test instances.

Trajectory-level Recall (Rec). Let $\mathcal{G}$ denote the set of ground-truth evidence pages for a given query, and let $\mathcal{T}$ denote the set of pages collected by the agent along a trajectory. The trajectory-level recall is defined as:

$$\text{Rec} = \frac{|\mathcal{T} \cap \mathcal{G}|}{|\mathcal{G}|}. \tag{9}$$

This metric measures the fraction of ground-truth pages successfully retrieved by the agent over the course of a trajectory, providing an indicator of how effectively the agent gathers relevant information.

Trajectory-level Prevision (Pre). Let $\mathcal{G}$ denote the set of ground-truth evidence pages for a given query, and let $\mathcal{T}$ denote the set of pages collected by the agent along a trajectory. The trajectory-level precision is defined as:

System prompt of ALDEN with Visual RAG

You are a helpful assistant designed to answer user questions based on a user-provided multi-page document. The document can not be input directly with the question, you must reason step by step to determine how to obtain evidence document pages by optimally utilizing tools and analyze the relevant content in the obtained document pages to precisely answer user's question. Your reasoning process MUST BE enclosed within <think> </think> tags. Your answer MUST BE enclosed within <answer> </answer> tags. In the last part of the answer, the final exact answer is enclosed within \boxed{{}} with latex format. The available tool is a **search tool**. After reasoning, you can invoke the search tool by generating <search> your search query here </search> to retrieve document pages most relevant to your search query. For example, your response could be in the format of '<think> your reasoning process </think> <search> search query </search>', or '<think> your reasoning process </think> <answer> your answer here. The final answer is \[\boxed{{answer here}} \] </answer>'. After invoking a tool, the user will return obtained document pages inside <result> </result> tags to you. Besides, the user will additionally provide the page number of the obtained page.
Important constraints:
- Only if you get all the potential evidence pages and find that the there is no evidenced answer or the document content is irrelevant to the user query, you can respond with '<think> your reasoning process </think> <answer> The final answer is \[\boxedThe problem is not answerable \] </answer>'.
- If multiple valid answers are found, return them separated by semicolons.
- You may not get the true evidence page in one-shot, carefully check whether the obtained pages are the true evidence page. If not, try different rewritings of your query or try different tool usage strategy several times.

Figure 5: System prompt of ALDEN with Visual RAG

$$\text{Pre} = \frac{|\mathcal{T} \cap \mathcal{G}|}{|\mathcal{T}|}. \tag{10}$$

This metric measures the fraction of pages collected by the agent that are actually relevant, providing an indicator of how accurately the agent identifies evidence pages during a trajectory.

F1 Score (F1). Based on the trajectory-level precision and recall, the trajectory-level F1 score is defined as the harmonic mean of the two:

$$\text{F1} = 2 \cdot \frac{\text{Pre} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}. \tag{11}$$

This metric provides a balanced measure of the agent's performance, accounting for both its ability to collect relevant pages (recall) and to avoid collecting irrelevant ones (precision) over a trajectory.

Number of uniquely collected pages (#UP).

F Case Study

In this section, we present typical examples from the LongDocURL benchmark to analyze the behavior of the ALDEN model. As illustrated in the successful cases in Tab. 6 and Tab. 7, the agent effectively learns to combine the `search` and `fetch` actions to gather evidence and correctly answer the question. In particular, the example in Tab. 7 shows the agent using `fetch` to complement `search` when the latter fails to retrieve the true evidence page in a single step, highlighting the benefit of ALDEN 's expanded action space. Alongside these successes, we also present failure cases in Tab. 8 and Tab. 9. These examples reveal that the model still struggles to verify evidence pages and explore sufficiently when the `search` action returns results far from the ground-truth pages—a challenge we leave for future work.

User Question	What's name of the figure at the page which contains a table whose name is "Table 18: Quality control related uses by type of use"?
Ground-truth Pages	40
Generated Trajectory	**assistant** <think> The question asks for the name of the figure at the page that contains a table with the name "Table 18: Quality control related uses by type of use". To answer this, I should use the search tool to look for the table with the specified name in the document. The fetch tool can be used to obtain the specific page where this table is located. </think> <search> name of the figure at the page which contains a table whose name is "Table 18: Quality control related uses by type of use" </search> **user** <result> Page 39: Image tokens </result> **assistant** <think> The table mentioned in the search query is not found on the provided page. The table in the document is "Table 17: Regulatory uses by main types of uses". The search query asks for the figure at the page which contains the table "Table 18: Quality control related uses by type of use". I should use the fetch tool to obtain the page with the table "Table 18: Quality control related uses by type of use". </think> <fetch> 40 </fetch> **user** <result> Image tokens </result> **assistant** <think> The table mentioned in the search query is found on the provided page. The table in the document is "Table 18: Quality control related uses by type of use". The search query asks for the figure at the page which contains the table "Table 18: Quality control related uses by type of use". The figure mentioned in the document is "Figure 20: Evolution of total numbers of uses of animals for pyrogenicity testing between 2015 and 2018". The figure name is "Figure 20". </think> <answer> The final answer is \[\boxe{{Figure 20}} \]</answer>

Table 6: Example of a generated case showing the user question and ALDEN's full reasoning–action trajectory.

System prompt of ALDEN with Hybrid RAG

You are a helpful assistant designed to answer user questions based on a user-provided multi-page document. Each page exists in two modalities: the original image and an OCR text extraction. You cannot access the full document directly; instead, you must reason step by step to determine how to obtain evidence document pages by optimally utilizing tools and analyze the relevant content in the obtained document pages to precisely answer user's question. Your reasoning process MUST BE enclosed within <think> </think> tags. Your answer MUST BE enclosed within <answer> </answer> tags. In the last part of the answer, the final exact answer should be enclosed within \boxed{{}} with latex format. The available tools include a **search tool** and a **fetch tool**. After reasoning, you can invoke either the search tool by generating <search> your search query here </search> to retrieve relevant document pages in both modalities or the fetch tool by generating <fetch> modal, page number </fetch> to obtain a specific document page in the specified modal, where the modal should be 'image' or 'text' and the page number should be a integrity number chosen from the user specified page number range. For example, your response could be in the format of '<think> your reasoning process </think> <search> search query </search>', or '<think> your reasoning process </think> <fetch> image, page number </fetch>', or '<think> your reasoning process </think> <fetch> text, page number </fetch>', or '<think> your reasoning process </think> <answer> your answer here. The final answer is \[\boxed{{answer here}} \] </answer>'. After invoking a tool, the user will return obtained document pages inside <result> </result> tags to you. For the search tool, the user will return both the relevant image pages and the relevant OCR text pages and attach them with corresponding page numbers. For the fetch tool, the user will only return either the image page or the OCR text page according to your input arguments.

Important constraints:

- Only if you get all the potential evidence pages and find that the there is no evidenced answer or the document content is irrelevant to the user query, you can respond with '<think> your reasoning process </think> <answer> The final answer is \[\boxedThe problem is not answerable \] </answer>'.
- If multiple valid answers are found, return them separated by semicolons.
- Only one page can be fetched at a time using the fetch tool.
- You may not get the true evidence page in one-shot, carefully check whether the obtained pages are the true evidence page. If not, try different rewritings of your query or try different tool usage strategy several times.
- Page numbers shown in the document pages may not be consistent with user specified page number range. In case of any discrepancy, the user defined parge number range shall prevail.
- You need to invoke the tools at least once and can invoke up to 5 times. When you output the answer, the interaction stops.

Figure 6: System prompt of ALDEN with Hybrid RAG.

User Question	What is the percentage decrease in hospitalizations between May 14 and May 20, 2020? (A). -6.5(B). +6.5(C). -65(D). +65Choose the letter name in front of the right option from A, B, C, D.
Ground-truth Pages	21
Generated Trajectory	**assistant** <think> The question asks for the percentage decrease in hospitalizations between May 14 and May 20, 2020. To answer this, I need to find information in the document that discusses hospitalizations for these specific dates and the percentage change between them. I will use the search tool to look for the relevant information in the document. </think> <search> percentage decrease in hospitalizations between May 14 and May 20, 2020 </search> **user** <result>Page 21: Image tokens </result> **assistant** <think> The image contains the information needed to answer the question. According to the document, there has been a -6.5% decrease in hospitalized patients between May 14 and May 20, 2020. The correct answer is (A). </think> <answer> The final answer is \[\boxed{{A}} \]</answer>

Table 7: Example of a generated case showing the user question and ALDEN's full reasoning–action trajectory.

User Question	What is the interval between specific lunar/solar events in days?
Ground-truth Pages	46, 47, 48, 55, 57, 58
Generated Trajectory	**assistant** <think> The question asks for the interval between specific lunar/solar events in days. To answer this, I should look for information in the document that discusses lunar and solar events and their intervals. The search tool seems the best option to find this information. I will use the search tool first. </think> <search> interval between lunar/solar events in days </search> **user** <result>Page 25: Image tokens </result> **assistant** <think> The image contains the information needed to answer the question. The document states that the interval between the lunar eclipse on April 15, 2014, and the lunar eclipse on September 28, 2015, is exactly 266 days. This matches the information in the question. The interval between specific lunar/solar events in days is 266 days. The final answer is \[\boxed{266} \]</think> <answer> The final answer is \[\boxed{266} \]</answer>

Table 8: Example of a generated case showing the user question and ALDEN's full reasoning–action trajectory.

User Question	What penalties are specified for possession of category l or category 2 items?
Ground-truth Pages	21
Generated Trajectory	**assistant** <think> The question asks for the penalties specified for possession of category 1 or category 2 items. To answer this, I should look for information in the document that discusses the penalties for these categories. The search tool seems the best option to start with as it can help locate the relevant information in the document. I will use the search tool first. </think> <search> penalties for possession of category 1 or category 2 items </search> **user** <result>Page 21: Image tokens </result> **assistant** <think> The image contains the information needed to answer the question. According to the document, the penalties for possession of a category 1 item or a category 2 item are specified in the Misuse of Drugs Act 1981. For a simple offence, the penalty is $12,000 or imprisonment for 3 years or both. The relevant information is found in the document. The final answer is \[\boxed{{$12,000 or imprisonment for 3 years or both}} \]</think> <answer> The final answer is \[\boxed{{$12,000 or imprisonment for 3 years or both}} \]</answer>

Table 9: Example of a generated case showing the user question and ALDEN's full reasoning–action trajectory.

Algorithm 1 PPO with Dual KL Regularization for Multi-Turn VRDU Agents

Require: Actor π_θ, Critic V_ϕ, Reference model π_{ref}, KL weights $\beta_{\text{gen}}, \beta_{\text{obs}}$, discount factors $\gamma_{\text{token}}, \gamma_{\text{turn}}$, GAE parameters $\lambda_{\text{token}}, \lambda_{\text{turn}}$, replay buffer $\mathcal{B}$

Initialize replay buffer $\mathcal{B}$
for iteration $= 1, 2, \ldots$ **do**
  Sample $|\mathcal{B}|$ queries from the dataset
  for each query **do**
    Reset: query q, empty retrieval history, $t \leftarrow 1$
    while $t < T$ **and** $a_{t-1} \neq$ `answer` **do**
      π_θ generates a token sequence $a_t \sim \pi_\theta(\cdot|s_t)$
      Parse the discrete action (`search`, `fetch`, or `answer`) from a_t
      Execute action $\rightarrow$ obtain new state s_{t+1} and turn reward r_t
      Store $\{a_t, s_{t+1}, r_t\}$ in $\mathcal{B}$
      $t \leftarrow t + 1$
  Turn-level value estimation:
  for each episode in $\mathcal{B}$ **do**
    Estimate $V_\phi(s_t)$ at final token of each turn
    Compute target turn value $\hat{V}_t$ via turn-level GAE
    Assign token-level reward $\tilde{r}_t \leftarrow \hat{V}_t$
  Dual KL penalty computation:
  for each token in $\mathcal{B}$ **do**
    if token is generated **then**
      Compute A_t^i via token-level GAE using $\tilde{r}_t$
      Compute $\mathrm{KL}(\pi_\theta(\cdot|s) \,\|\, \pi_{\text{ref}}(\cdot|s))$ with weight β_{gen}
    else if token is observation **then**
      Compute $\mathrm{KL}(\pi_\theta(\cdot|s) \,\|\, \pi_{\text{ref}}(\cdot|s))$ with weight β_{obs}
  PPO update:
  Update θ by maximizing policy loss $\mathcal{L}_{\text{policy}}$
  Update ϕ by minimizing value loss $\mathcal{L}_{\text{value}}$