# REMEMBER THE LAB: ADAPTIVE VISUAL ABSTRACTION FOR LONG-HORIZON LABORATORY UNDERSTANDING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Modern laboratory research involves long-horizon experimental processes, where states, procedures, and outcomes continuously evolve across repeated experiments. With the emergence of multimodal foundation models, AI assistants have shown great potential for laboratory research. However, effective assistance in dynamic laboratory environments requires the ability to maintain and utilize knowledge accumulated throughout experimental processes. Existing multimodal systems mainly focus on current observations, limiting their ability to reason over evolving laboratory processes across long temporal horizons. We introduce **RA-Lab**, a multimodal memory-augmented agent framework for long-horizon laboratory video understanding. RA-Lab constructs persistent experimental memories through two key components: ***Adaptive Visual Abstraction***, which transforms raw observations into experience-level visual memories, and ***Targeted Multimodal Retrieval***, which retrieves relevant historical experiences based on the current observation and query. The retrieved memories are combined with current observations to support memory-augmented laboratory reasoning. To evaluate long-horizon laboratory understanding, we introduce **LabVQA**, a benchmark for laboratory video reasoning over experimental procedures, states, and temporal relationships. Extensive experiments demonstrate that RA-Lab effectively constructs and utilizes multimodal experimental memories, improving long-horizon laboratory reasoning compared with existing multimodal models and memory-augmented agents.

## 1 INTRODUCTION

Recent advances in artificial intelligence have shown great potential for scientific discovery, enabling AI systems to assist researchers in tasks such as literature search, coding, hypothesis generation, and experiment planning Phan et al. (2025); Novikov et al. (2025); Aygün et al. (2026). Beyond digital assistance, self-driving laboratory Tom et al. (2024); Lu et al. (2024); Cong et al. (2025); Gottweis et al. (2026) aims to bring AI into physical experimental environments, where agents can interact with instruments, samples, and scientific procedures. As AI assistants move toward active participation in experiments, they need to continuously understand ongoing laboratory processes.

Recent advances in multimodal large language models (MLLMs) Zheng et al. (2026); Yang et al. (2025); Tian et al. (2025) have enabled AI agents to perceive and reason about activities from egocentric and fixed-view videos. However, current multimodal agents are mainly designed for short-term visual understanding and remain limited in supporting long-horizon laboratory scenarios. To study this capability, we introduce **LabVQA**, a long-horizon laboratory video question answering benchmark constructed from real-world experimental videos. LabVQA evaluates whether multimodal agents can understand laboratory procedures, experimental states, and evolving processes over time. As compared in the bottom of figure 1, our evaluation shows that existing models struggle when solving tasks that require information beyond the current observation.

In this work, we introduce **RA-Lab** (Remember those abstraction image on laboratory work), a multimodal memory-augmented agent framework for long-horizon laboratory video understanding. RA-Lab builds experimental memories through two key components: ***Adaptive Visual Abstraction***,



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

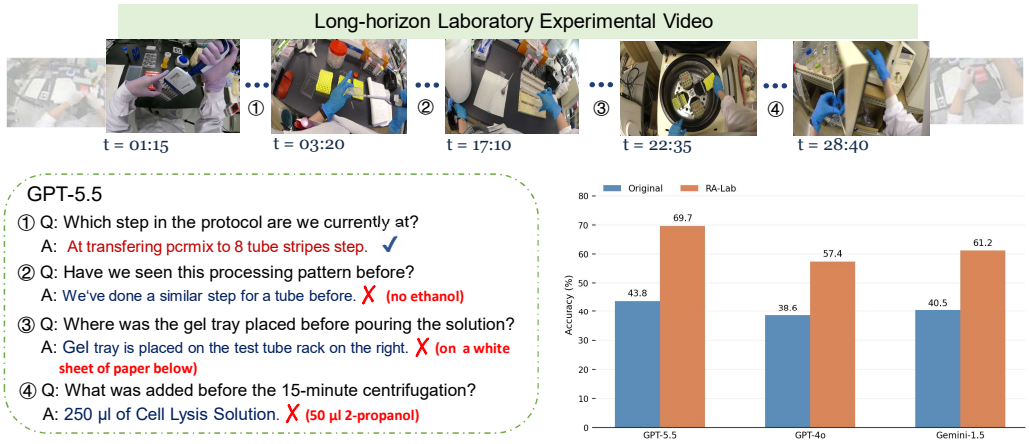


Figure 1: **Multimodal Large Models struggle with long laboratory understanding task.** Top: Examples of a long-horizon laboratory video. Bottom: The curved box on the left shows examples of GPT-5.5 answers in our designed video question-answering setting, where each number corresponds to a number in the video stream above, indicating the time point at which a question was asked to the model. The bar chart on the right shows a comparison of question-answering accuracy between the original model and the model after integrating the multimodal agent framework RA-Lab.

which converts raw observations into experience-level visual memories by preserving task-relevant information, and ***Targeted Multimodal Retrieval***, which retrieves relevant past experiences based on the current observation and query. The retrieved memories are combined with current observations to support memory-augmented laboratory reasoning.

The main contributions of this paper are summarized as follows:

*i.* We formulate long-horizon laboratory understanding as a multimodal memory problem, highlighting the need to recall historical experimental experience beyond the current visual observation.

*ii.* We propose a multimodal memory framework with adaptive visual abstract and Targeted Multimodal Retrieval, designed to preserve and recall decision-relevant experimental experience.

*iii.* We establish a comprehensive evaluation across laboratory reasoning tasks and long-horizon laboratory video QA, analyzing the trade-offs among memory utility, visual detail preservation, retrieval quality, and computational efficiency.

## 2 Related Work

### 2.1 Memory in Streaming and Long Video Understanding

Online understanding of long-horizon videos has become increasingly important Wu et al. (2022); Tang et al. (2025); Zhang et al. (2024). Existing approaches Ren et al. (2024); Yao et al. (2025); Di et al. (2025); Wu & Krahenbuhl (2021) often equip video models with streaming memory to extend temporal coverage. Longvila Chen et al. (2025) and LongVA Zhang et al. (2024) expand context windows and compress visual tokens, while StreamReady Azad et al. (2026) introduces a Ready-Gate for time-sensitive long-video QA. StreamArena Zhang et al. (2026) further separates foreground interaction from backend memory for efficient hourly video QA. However, these methods primarily operate at the frame or token level, limiting scalability to increasingly long histories. Recent external agent memory approaches provide a more flexible alternative for long-term video understanding.




Figure 2: **Overview of the proposed multimodal memory framework for long-horizon laboratory video understanding.** Historical laboratory activities are organized into multimodal experimental experiences, whose visual evidence is adaptively abstracted into compact visual memories. Given the current observation and task query, targeted multimodal retrieval selects relevant historical experiences, which are then used by the MLLM to support protocol alignment, error detection, next-step prediction, and long-video question answering.

### 2.2 Long-Term Memory for Multimodal Agents

Early long-term memory methods Feng et al. (2024); Liu et al. (2024); Zheng et al. (2026) mainly target ultra-long contextual dialogue. With the emergence of minute- and hour-level video benchmarks Wang et al. (2025); Song et al. (2024); Fu et al. (2025), multimodal agents with long-term memory have recently been explored for video understanding. M3-Agent Long et al. (2026) maintains an entity-centric multimodal graph, while WorldMM Yeo et al. (2026) constructs multi-scale episodic, semantic, and visual memories. However, existing methods either compress visual content into text, sacrificing fine-grained visual information, or retain large numbers of raw frames, introducing substantial redundancy and token-level reasoning costs Ma et al. (2026). These limitations motivate a more adaptive way to determine what visual information should be preserved in long-term memory.

### 2.3 Scientific Laboratories with AI

Scientific laboratories are increasingly becoming important settings for AI-driven experimentation Li et al. (2024); Yoshikawa et al. (2022); Huang et al. (2025). Laboratory experiments involve extended procedures, observations, and evolving experimental states Lan et al. (2025). Self-driving laboratories pursue closed-loop experiment planning and execution Szymanski et al. (2023); Granda et al. (2018), while recent systems such as LabOS Cong et al. (2025) and LabVLA Ren et al. (2026) explore AI-assisted scientific workflows. However, existing MLLM-based systems remain limited in observing and guiding complex laboratory activities Cong et al. (2025); Phan et al. (2025), partly because information required for understanding the present experiment may reside in previously observed experiences rather than the current video alone. **Laboratories therefore provide a natural and challenging setting for memory-augmented multimodal agents, where historical visual experience can complement the immediate context.**

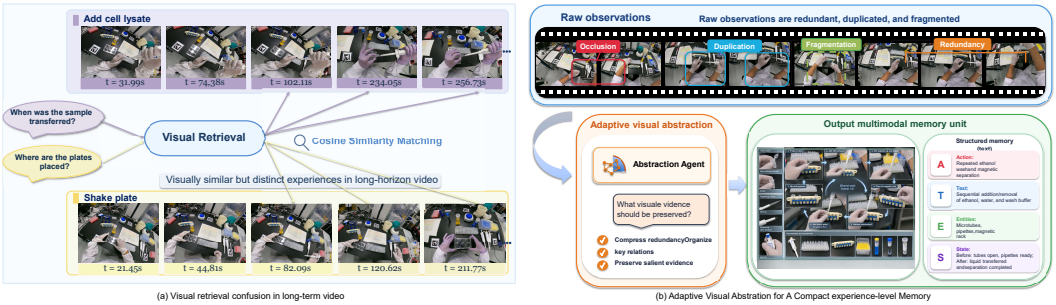


Figure 3: **Motivation and overview of adaptive visual abstraction for long-term laboratory memory.** (a) In long-horizon videos, different experiences can exhibit highly similar visual patterns, making similarity-based retrieval ambiguous. (b) Adaptive Visual Abstraction identifies and reorganizes salient visual evidence from redundant and fragmented observations into a compact experience-level visual memory, while preserving the underlying multimodal context and source evidence.

## 3 METHODOLOGY

We introduce RA-Lab, a multimodal memory-augmented agent for long-horizon laboratory video understanding. The core idea is to transform continuous laboratory observations into reusable, step-level multimodal memories, combining structured information with a compact visual representation. As illustrated in Figure 2, RA-Lab first divides long laboratory videos into local temporal segments and aligns each segment with an experimental protocol step. Consecutive segments corresponding to the same step are then consolidated into a step-level experience. For each experience, an abstraction agent determines which visual information should be preserved and how it should be organized, producing a compact visual abstraction. The resulting structured experience and visual abstraction are stored as a multimodal memory and can later be retrieved to support reasoning over the current observation.

Formally, let $\mathcal{H}_{<t} = \{V_1, \ldots, V_{t-1}\}$ denote the historical laboratory videos and $V_t$ the current observation. Given a task or question $q_t$, RA-Lab incrementally constructs a memory bank $\mathcal{M}_t = \{M_1, \ldots, M_n\}$ from historical observations. Each memory $M_i$ contains a structured experience and its corresponding visual abstraction. The relevant memories are retrieved from $\mathcal{M}_t$ and provided to the reasoning model together with the current observation and query:

$$y_t = \mathcal{G}\left(V_t, q_t, \mathcal{R}(\mathcal{M}_t, V_t, q_t)\right), \tag{1}$$

where $\mathcal{G}$ denotes the reasoning model and $\mathcal{R}$ denotes the retrieval process. The following sections describe memory construction, visual abstraction, retrieval, and memory-augmented reasoning in detail.

### 3.1 MULTIMODAL EXPERIMENTAL EXPERIENCE MEMORY

We first divide each laboratory video into fixed-length local temporal segments. Each segment serves as a local observation unit. For each segment, we sample a set of video frames and provide them together with the temporal context and the experimental protocol to a multimodal large language model. The model identifies the protocol step corresponding to the current segment and determines whether the step is still in progress or has been completed. This step alignment connects local visual observations with the semantic structure of the experimental procedure.

Consecutive segments assigned to the same protocol step are then consolidated into a single step-level experience. The consolidation combines the information accumulated across the corresponding segments, including the experimental actions, entities and materials, states and state transitions, textual description, and temporal context. Let $\mathcal{V}_i$ denote the sampled visual observations associated with the $i$-th step, and let the resulting structured experience be

$$e_i = \{T_i, E_i, A_i, S_i, \tau_i\}, \tag{2}$$

where $T_i$ denotes the textual and procedural description, $E_i$ denotes the relevant entities and materials, $A_i$ denotes the actions performed, $S_i$ denotes the experimental states and their transitions, and $\tau_i$ denotes the temporal context.

The visual observations $\mathcal{V}_i$ are retained as source evidence for subsequent visual abstraction. Together, the structured experience $e_i$ and its corresponding visual evidence $\mathcal{V}_i$ provide the basis for constructing the multimodal memory.

### 3.2 Adaptive Visual Abstraction

Long laboratory videos contain substantial visual redundancy, with many different experimental steps repeatedly exhibiting similar low-level patterns such as hands, tubes, pipettes, and laboratory work surfaces. Directly retaining sampled frames therefore provides limited information and may make visually similar experiences difficult to distinguish. The challenge is not only to reduce the number of visual observations, but also to determine which visual information is important for representing an experimental experience.

We therefore introduce *Adaptive Visual Abstraction* to transform the visual evidence of each step-level experience into a compact visual representation. Given a structured experience $e_i$ and its associated visual observations $\mathcal{V}_i$, an abstraction agent first determines *what to preserve*, *what to compress*, and *how to organize* the retained visual information. The abstraction focuses on information that characterizes the experimental step, including salient actions, relevant entities and states, critical relationships, visual changes, and meaningful temporal structure, while repeated or low-information observations are compressed.

Specifically, the abstraction agent produces an abstraction plan

$$p_i = \mathcal{A}_{\text{plan}}\left(e_i, \mathcal{V}_i\right), \tag{3}$$

where $p_i$ specifies the visual evidence to preserve, the information to compress, and the organization used to represent the experience. The organization is determined according to the content of each experience rather than following a fixed layout, allowing different experimental steps to be represented in different visual forms.

The abstraction plan is then converted into image-generation instructions:

$$\pi_i = \Phi(p_i), \tag{4}$$

where $\Phi$ translates the structured plan into instructions for an image generation or editing model. The model synthesizes a single visual abstraction from the plan and the source visual observations:

$$\hat{V}_i = \mathcal{G}_{\text{img}}\left(\pi_i, \mathcal{V}_i\right). \tag{5}$$

The resulting $\hat{V}_i$ is neither a selected frame nor a direct concatenation of the source observations. Instead, it provides a compact visual representation of the experimental experience by preserving salient visual evidence and relationships while suppressing redundant observations. Because both the retained content and its visual organization are determined from the individual experience, different experimental steps can yield different visual abstractions.

Finally, each multimodal memory is represented as

$$m_i = \{e_i, \hat{V}_i\}, \tag{6}$$

where $e_i$ is the structured step-level experience and $\hat{V}_i$ is its corresponding visual abstraction. These multimodal memories are stored in the memory bank for subsequent retrieval and reasoning.

### 3.3 Targeted Multimodal Retrieval

As the memory bank grows, RA-Lab performs targeted retrieval to efficiently identify relevant historical experiences.

Given the current observation and query, we encode them into a retrieval representation $z_t = \mathcal{E}_q(V_t, q_t)$. Each historical experience is represented by its structured memory and visual abstraction,

Table 1: QA accuracy (%) under different visual memory varints. For each question, the same ground-truth relevant experience is directly provided to GPT-5.5, with retrieval excluded.

| **Memory Variants** | **Visual Evidence** | **Accuracy (%)** |
|---|---|---|
| No Memory | – | 21.8 |
| Text-only Memory | None | 49.3 |
| Text + 2 Raw Frames | 2 frames | 45.6 |
| Text + 4 Raw Frames | 4 frames | 52.1 |
| Text + 8 Raw Frames | 8 frames | 47.4 |
| Text + Visual Abstraction (Ours) | 1 abstraction | **57.7** |

forming a multimodal memory representation $h_i = \mathcal{E}_m(m_i, \hat{V}_i)$. We then compute the relevance between the current query and each step-level experience as $r_i = \mathrm{sim}(z_t, h_i)$, and retrieve the top-$K$ experiences with the highest relevance scores.

The retrieved experiences retain both their structured experimental context and compact visual abstractions, which are subsequently provided to the reasoning module for multimodal memory-augmented inference. Thus, targeted retrieval operates at the experience level rather than over redundant raw observations, enabling RA-Lab to directly access compact, reusable historical experiences.

### 3.4 Memory-Augmented Reasoning

The retrieved memories are provided to the reasoning model together with the current observation and query:

$$y_t = \mathcal{G}\left(V_t, q_t, \mathcal{M}_t^*\right), \tag{7}$$

where $\mathcal{M}_t^*$ denotes the top-$K$ retrieved step-level experiences. Each retrieved experience contains a structured memory representation and its corresponding visual abstraction. The reasoning model therefore operates on a compact multimodal representation of historical experiences, allowing previously observed experimental knowledge to be reused without repeatedly processing the full historical video.

## 4 Experiment

**Models** For Laboratory Understanding tasks, we compared RA-Lab against a comprehensive set of baseline models, including streaming video models, such as MovieChat Song et al. (2024) and Streamingbridge Wang et al. (2026); vision-language models, including GPT-5.5 Singh et al. (2025), Qwen2.5-VL Bai et al. (2023), and Gemini-1.5 Team et al. (2024); and memory-augmented multimodal agents, including M3-Agent Long et al. (2026), WorldMM Yeo et al. (2026). For WorldMM, we use GPT-5.5 (as set in the original paper) as the base model. For M3-Agent, its base model is the Qwen2.5-Omini Bai et al. (2023) series, and we directly use the fine-tuned weights provided by these models. As described in section 3, our method involves step-level memory construction, abstraction plan generation, abstraction image generation based on the abstraction plan and memory-augmented reasoning. Specifically, we employ vision-language models such as GPT-5.5 OpenAI (2025) for image-to-text understanding and reasoning tasks (step-level memory construction, abstraction plan generation, etc.) and image generation models such as GPT-Image-2 OpenAI (2026) to generate abstraction images.

**Datasets and Evaluation Tasks.** We constructed a dataset called LabVQA using laboratory videos from three open-source datasets: **FineBio** Yagi et al. (2025), **BioVL2** Nishimoto et al. (2025), and **EgoExoLearn** Huang et al. (2024). Together, these datasets provide approximately 250 videos, totaling nearly 800 hours, depicting real-world laboratory activities with temporal, procedural, and visual annotations that support the construction of long-horizon evaluation tasks.

From these recordings, we construct a 80-video benchmark covering historical experience recall, state evolution reasoning, temporal-spatial reasoning, experience comparison, and experience-grounded reasoning. We additionally evaluate protocol alignment and issue identification tasks as

Table 2: Performance comparison on LabVQA benchmark across five capability dimensions. The benchmark covers historical experience recall, state evolution reasoning, temporal-spatial reasoning, experience comparison, and experience-grounded reasoning.

| Method | Hist. Recall ↑ | State Evol. ↑ | Temp.-Spatial ↑ | Exp. Comp. ↑ | Exp. Grounded ↑ |
|---|---|---|---|---|---|
| *Streaming Video Model* | | | | | |
| MovieChat Song et al. (2024) | 18.9 | 15.6 | 19.2 | 13.8 | 14.5 |
| Gemini 1.5 pro Team et al. (2024) | 39.5 | 40.2 | 37.3 | 42.7 | 38.8 |
| *Vision-Language Model* | | | | | |
| GPT-5.5 OpenAI (2025) | 48.1 | 47.2 | 48.6 | 39.1 | 49.5 |
| Qwen2.5-VL-7B Bai et al. (2023) | 27.7 | 30.4 | 35.0 | 29.3 | 37.7 |
| *Memory-augmented Multimodal Agent* | | | | | |
| M3-Agent Long et al. (2026) | 45.9 | 51.0 | 41.4 | 38.9 | 47.1 |
| WorldMM Yeo et al. (2026) | 51.3 | 41.8 | 44.1 | **48.0** | 52.9 |
| **RA-Lab(Ours)** | **62.7** | **60.4** | **59.5** | 43.9 | **58.6** |

described in LabOS Cong et al. (2025) to examine whether multimodal memory can provide useful historical context for downstream laboratory understanding tasks. All evaluation instances are grounded in the original video evidence, and the generated question-answer pairs are manually verified to ensure their quality. For each question, the expected answer is determined based on the corresponding experimental steps, visual observations, and temporal context provided by the source videos. We evaluate model responses by comparing them with the reference answers using both exact matching for objective questions and semantic evaluation for open-ended reasoning questions. For open-ended responses, answers are considered correct if they preserve the key experimental information, including relevant actions, entities, states, temporal relationships, or procedural explanations required by the question. Detailed benchmark construction, task definitions, question statistics, and examples are provided in the Appendix A.

**Implementation Details** The input video is segmented into video segments of fixed 15s duration and sampled for MLLM understanding. We sampled 8 frames each segements for models to understanding or generation. Considering memory overhead, we delete sampled frames from the local memory after each input to the model to construct the memory. All embeddings are embedded using VLM2Vec-V2 Meng et al. (2025). The relevance between the query and historical experiences is computed based on their representations, and the top-$K$ experiences are retrieved and provided to the reasoning model. LabVQA is constructed from real-world laboratory videos and protocols, with question-answer pairs generated by a multimodal large language model based on temporal experimental trajectories and visual observations. When generation questions in LabVQA, for each query, 6 frames are sampled from the historical context and 6 frames from the current segment, resulting in 12 visual observations provided to the question generation process. More detailed prompts, sampling settings, and retrieval parameters are provided in the Appendix B.1.

### 4.1 Main Result

**Analysis of Memory Modality and Scalability under LabVQA.** To examine the role of visual representations in multimodal memory, we compare different ways of representing the same relevant historical experience, as shown in Table 1. For each question, the relevant historical experience is identified from the annotated source video and directly provided to the reasoning model, thereby excluding the effect of memory retrieval. We consider a text-only memory, text memory plus memory augmented with different numbers of raw video frames, and the proposed visual abstraction representation. The text-only setting provides only the structured step-level experience, while the raw-frame settings additionally provide 2, 4, or 8 sampled frames from the same experience. The visual abstraction setting instead provides the structured experience together with a single generated visual abstraction. All settings use the same historical memory, query, and reasoning model (GPT-5.5), with only the visual representation varying. As the table shows, increasing the number of visual memories does not necessarily lead to higher QA accuracy; in fact, it can even decrease accuracy when the number of images is 12 (compared to text-only memory). However, almost all multimodal memory methods outperform text-only memory, demonstrating the critical importance of visual memory in this laboratory setting. Our method achieves the best results with only one

Table 3: Evaluation on protocol alignment and issue identification with GPT-5.5 and Qwen2.5-VL-7B. We compared three settings: no memory (Baseline), text-only memory, and RA-Lab (Ours).

| | GPT-5.5 | | Qwen2.5-VL | |
|---|---|---|---|---|
| **Memory Setting** | Protocol Alignment ↑ | issue identification ↑ | Protocol Alignment ↑ | issue identification ↑ |
| Baseline | 41.7 | 42.5 | 26.9 | 25.1 |
| + Text-only Memory | 35.0 | 32.9 | 29.8 | 28.2 |
| + RA-Lab (ours) | **59.5** | **49.7** | **45.2** | **42.9** |

image as visual memory, proving our motivation that redundancy in visual memory leads to comprehension difficulties.

**Performance on LabVQA of Existing Models.** Table 2 presents a comprehensive comparison on our LabQA benchmark, covering five capability dimensions and three categories of baseline models. In this setup, GPT-5.5 serves as the basemodel of abstraction agent and reasoning agent for RA-Lab, while GPT-Image-2 is utilized for image generation. It can be seen from the results of stream video models that existing multimodal large language models exhibit weak scene understanding and memory capabilities on long laboratory videos. For visual language models, GPT-5.5 performed best, yet still achieved less than half the accuracy. Nevertheless, RA-Lab boosts historical experience recall and state evolution reasoning scores to 62.7 and 60.4 by equipping GPT-5.5 with our designed multimodal memory-enhanced agent framework. Regarding the memory-augmented multimodal agents, WorldMM and Streambridge deliver the relatively best performance across various metrics. Even so, their accuracy across all problems remains only around 50%. The above results demonstrate the effectiveness and scalability of RA-Lab. All model configurations in this experiment follow their original papers or common practice, with detailed specifications provided in Appendix B.

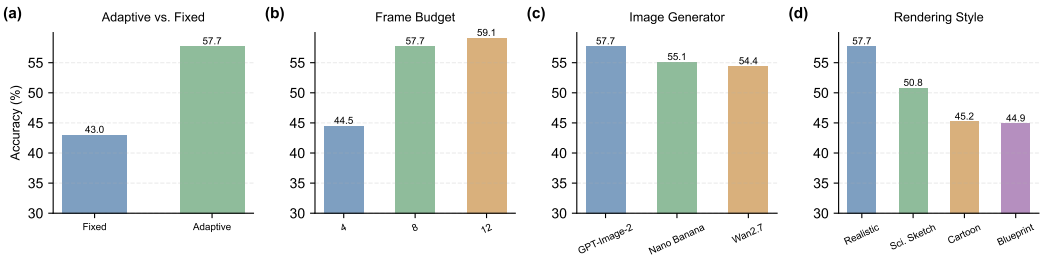


Figure 4: Analysis of visual abstraction under different content organization, frame budgets, image generators, and rendering styles. Here, Nano Banana. represents Nano Banana Pro, and Sci. Sketch represents the rendering style of Scientific Sketch.

**Discussion and Analysis of Visual Abstraction.** To further analyze the adaptive visual abstraction, we conducted quantitative experiments from multiple perspectives. First, we tested the impact of experience-adaptive and fixed-template abstraction plans on QA accuracy. Additionally, we examined the impact of the number of conditional images input to the generation model (e.g., GPT-Image-2) on QA results. The results, as shown in the figure 4, demonstrate that adaptive organization provides a clear improvement over the fixed design, while performance increases significantly once the number of source frames are much more. We also tested the image generation model and the impact of generation style (controlled by the generation prompt). The choice of image generator has a limited impact, whereas rendering style leads to greater variation, with realistic rendering achieving the best performance. These results indicate that visual content is the most important and sensitive element in the visual abstraction strategy. It must not only be semantically prominent but also visually similar to the source image frames. Furthermore, more source image frames lead to better abstraction results. Examples of our visual abstraction are shown in figure 7 and more details can be found in appendix C.1.

**Process-Oriented Evaluation.** Beyond question answering of laboratory assistant, RA-Lab is also evaluated on protocol alignment and issue identification Cong et al. (2025); Yagi et al. (2025). Specifically, Protocol alignment mean that the model needs to generate a stepwise protocol describing procedural actions and parameters and issue identification indicate that the model needs to

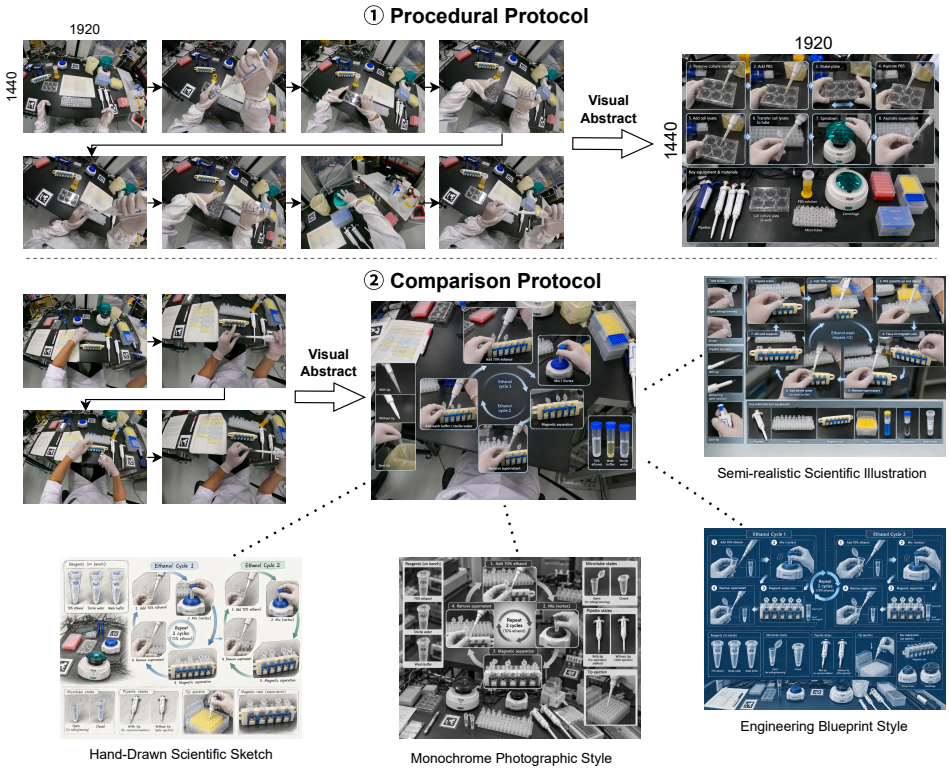


Figure 5: **Examples of visual abstraction from laboratory video experiences.** Each experience is compressed into a single visual memory by adaptively selecting, compressing, and reorganizing salient visual information, including entities, actions, states, relationships, and temporal structure. The resulting abstraction is experience-dependent rather than template-driven, and can be rendered with different visual styles while preserving the underlying information.

troubleshoot based on the video and identify any deviation from gold standard. The accuracy results are shown in table 3, without memory enhancement, the baseline model exhibits relatively poor accuracy. In RA-Lab, each local video segment is first aligned to a protocol step and consolidated into a step-level experience, which is then represented through both structured experimental information and an adaptive visual abstraction image. In short, the best-performing baseline model on the protocol alignment task, such as GPT-5.5, maximizes the advantages of RA-Lab, achieving an average accuracy improvement of 20%. Appendix B explains how RA-Lab is applied to these two types of tasks.

## 5 Conclusion

We present RA-Lab, a multimodal memory-augmented agent framework for long-horizon laboratory video understanding. By converting continuous laboratory observations into structured experimental experiences, RA-Lab enables agents to leverage historical knowledge during reasoning. Specifically, Adaptive Visual Abstraction builds experience-level visual memories, while Targeted Multimodal Retrieval recalls relevant experiences to support current observations. Experiments on laboratory procedural understanding and LabVQA demonstrate that RA-Lab improves long-horizon laboratory reasoning by exploiting past experimental experiences. RA-Lab has several limitations: visual memory quality depends on the underlying multimodal understanding and generation models, and excessive abstraction may discard fine-grained evidence. Step-level experience construction may also introduce missing or merged memories. Future work will explore more robust memory construction and retrieval strategies for scalable scientific agents.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with several aspects of the research and writing process. Specifically, AI tools were used to assist in the design and refinement of the experimental methodology, including the development of a long laboratory video question answering benchmark based on existing datasets. We also used generative AI models to assist in generating visual artifacts related to our proposed adaptive visual abstraction framework, where generative models serve as a component of the proposed method. Additionally, we used AI tools to assist with research code development, debugging, and manuscript refinement to improve readability and presentation. All AI-assisted outputs, including generated artifacts, code, and textual modifications, were carefully reviewed and verified by the authors. We take full responsibility for the final content of this work, including all scientific claims, experimental results, and materials presented in the paper.

### REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed framework, experimental settings, and evaluation protocols in the main paper and supplementary materials. The dataset processing pipeline, benchmark construction details, implementation details, and additional experimental results are provided in the appendix. We will release the source code and necessary resources upon publication to support future research and verification.

## REFERENCES

Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, Marc Coram, Hao Cui, Jake Garrison, Renee Johnston, Anton Kast, Cory Y McLean, Peter Norgaard, et al. An ai system to help scientists write expert-level empirical software. *Nature*, pp. 1–3, 2026.

Shehreen Azad, Vibhav Vineet, and Yogesh S Rawat. Streamready: Learning what to answer and when in long streaming videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40494–40504, 2026.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, volume 2025, pp. 18227–18246, 2025.

Le Cong, David Smerkous, Xiaotong Wang, Di Yin, Zaixi Zhang, Ruofan Jin, Yinkai Wang, Michal Gerasimiuk, Ravi K Dinesh, Alex Smerkous, et al. Labos: The ai-xr co-scientist that sees and works with humans. *arXiv preprint arXiv:2510.14861*, 2025.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. In *International Conference on Learning Representations*, volume 2025, pp. 42115–42127, 2025.

Peiyuan Feng, Yichen He, Guanhua Huang, Yuan Lin, Hanchong Zhang, Yuchen Zhang, and Hang Li. Agile: A novel reinforcement learning framework of llm agents. *Advances in Neural Information Processing Systems*, 37:5244–5284, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with co-scientist. *Nature*, pp. 1–3, 2026.

Jarosław M Granda, Liva Donina, Vincenza Dragone, De-Liang Long, and Leroy Cronin. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature*, 559 (7714):377–381, 2018.

Kefeng Huang, Jonathon Pipe, Alice E Martin, Tianyuan Wang, Barnabas A Franklin, Andy M Tyrrell, Ian JS Fairlamb, and Jihong Zhu. Tarmac: A taxonomy for robot manipulation in chemistry. *arXiv preprint arXiv:2510.19289*, 2025.

Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22072–22086. IEEE, 2024.

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

Zhiqian Lan, Yuxuan Jiang, Ruiqi Wang, Xuanbing Xie, Rongkui Zhang, Yicheng Zhu, Peihang Li, Tianshuo Yang, Tianxing Chen, Haoyu Gao, et al. Autobio: A simulation and benchmark for robotic automation in digital biology laboratory. *arXiv preprint arXiv:2505.14030*, 2025.

Shoujie Li, Yan Huang, Changqing Guo, Tong Wu, Jiawei Zhang, Linrui Zhang, and Wenbo Ding. Chemistry3d: Robotic interaction benchmark for chemistry experiments. *arXiv preprint arXiv:2406.08160*, 2024.

Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models. *arXiv preprint arXiv:2401.02777*, 2024.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *International Conference on Learning Representations*, volume 2026, pp. 146197–146246, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Xiaokang Ma, Yifan Sun, Zhihong Jin, Jie Gu, Yudong Luo, Shenyi Shao, Chu Tang, Jingmin Chen, and Li Pu. Reflectworld-mm: An entity-oriented multi-media memory system for open-ended video streams. *arXiv preprint arXiv:2607.09759*, 2026.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

Tomohiro Nishimoto, Taichi Nishimura, Koki Yamamoto, Keisuke Shirai, Hirotaka Kameko, Yuto Haneji, Tomoya Yoshida, Keiya Kajimura, Taiyu Cui, Chihiro Nishiwaki, et al. Biovl-qr: Egocentric biochemical vision-and-language dataset using micro qr codes. In *2025 IEEE International Conference on Image Processing (ICIP)*, pp. 695–700. IEEE, 2025.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

OpenAI. Update to GPT-5 system card: GPT-5.2. System card, December 2025. URL https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf. Documents the GPT-5.2 model family; accessed 2026-05-26.

OpenAI. GPT Image 2 model. https://developers.openai.com/api/docs/models/gpt-image-2, 2026.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Baochang Ren, Xinjie Liu, Xi Chen, Yanshuo Liu, Chenxi Li, Daqi Gao, Zeqin Su, Jintao Xing, Zirui Xue, Rui Li, et al. Labvla: Grounding vision-language-action models in scientific laboratories. *arXiv preprint arXiv:2606.13578*, 2026.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232. IEEE, 2024.

Nathan J Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E Kumar, Tanjin He, David Milsted, Matthew J McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624(7990):86, 2023.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128. IEEE, 2025.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*, 2025.

Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(16):9633–9732, 2024.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant. *Advances in Neural Information Processing Systems*, 38:132332–132359, 2026.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025.

Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1884–1894, 2021.

Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13577–13587. IEEE, 2022.

Takuma Yagi, Misaki Ohashi, Yifei Huang, Ryosuke Furuta, Shungo Adachi, Toutai Mitsuyama, and Yoichi Sato. Finebio: A fine-grained video dataset of biological experiments with hierarchical annotation. *International Journal of Computer Vision*, 133(10):7352–7367, 2025.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28885–28900. IEEE, 2025.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10807–10816, 2025.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. Worldmm: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026.

Naruki Yoshikawa, Andrew Zou Li, Kourosh Darvish, Yuchi Zhao, Haoping Xu, Artur Kuramshin, Alán Aspuru-Guzik, Animesh Garg, and Florian Shkurti. Chemistry lab automation via constrained task and motion planning. *arXiv preprint arXiv:2212.09672*, 2022.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

Xichen Zhang, Guankai Li, Yinghao Zhu, Shijian Wang, Sitong Wu, Shaozuo Yu, Meng Chu, Yuan Lu, and Jiaya Jia. Streamarena: Toward continuous, interactive, and long-horizon agentic streaming video understanding. *arXiv preprint arXiv:2608.05703*, 2026.

Junhao Zheng, Chengming Shi, Xidi Cai, Qiuke Li, Duzhen Zhang, Chenxing Li, Dong Yu, and Qianli Ma. Lifelong learning of large language model based agents: A roadmap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

# APPENDIX

**A LabVQA Benchmark Construction** **15**

A.1 Source Datasets . . . . . 15

A.2 LabVQA Instance Construction . . . . . 15

A.3 Memory Query Construction . . . . . 15

A.4 Question Generation and Annotation . . . . . 16

A.5 Annotation Verification . . . . . 17

A.6 Query Sampling Details . . . . . 17

A.7 Question Categories . . . . . 17

A.8 LabVQA Statistics and Examples . . . . . 17

**B Implementation Details** **18**

B.1 RA-Lab . . . . . 18

**C Details of Adaptive Visual Abstraction** **19**

C.1 Implementation of Adaptive Visual Organization . . . . . 19

**D Example of RA-Lab on LabVQA** **20**

D.1 Example of Memory-Augmented Laboratory Reasoning . . . . . 20

D.2 Examples of LabVQA Tasks . . . . . 22

**E Limitations** **23**

Table 4: Overview of laboratory video datasets used for LabVQA construction and evaluation.

| Dataset | Domain | Videos | Duration | Annotations |
|---|---|---|---|---|
| FineBio | Biological experiments | 50 | 2.5 h | Steps, atomic operations, objects |
| BioVL2 | Biochemical procedures | 22 | 1.05 h | Protocols, temporal procedures |
| EgoExoLearn | Laboratory activities | 8 | 0.33h | Ego-centric multimodal observations |

# A LabVQA Benchmark Construction

## A.1 Source Datasets

We construct LabVQA from three laboratory video datasets with complementary characteristics: FineBio Yagi et al. (2025), BioVL2, and EgoExoLearn Huang et al. (2024). These datasets provide diverse egocentric laboratory observations together with temporal or procedural annotations, enabling the construction of long-horizon laboratory memory tasks. FineBio serves as the primary source for LabVQA construction, providing fine-grained biological experiments with hierarchical annotations, including experimental steps, atomic operations, and object-level information. BioVL2 provides biochemical procedures paired with corresponding protocol descriptions, enabling trajectory-based memory construction. We additionally include the egocentric laboratory subset of EgoExoLearn to evaluate long-horizon understanding under diverse first-person experimental scenarios.

The statistics of the source datasets are summarized in Table 4.

## A.2 LabVQA Instance Construction

Existing laboratory video benchmarks mainly focus on action recognition, procedural understanding, or object-level perception. To evaluate whether models can recall and reason over previous experimental experiences, we construct LabVQA as a memory-oriented video question answering benchmark. Each LabVQA instance is defined as:

$$\mathcal{I}_i = (H_i, C_i, Q_i, A_i, E_i), \tag{8}$$

where $H_i$ denotes historical visual evidence, $C_i$ represents the current observation, $Q_i$ and $A_i$ denote the question and reference answer, and $E_i$ contains supporting evidence from the original experimental trajectory. Unlike conventional video QA settings, LabVQA questions cannot be answered from the current observation alone. Instead, models must utilize historical experimental context to infer the current state, track state evolution, or reason about previous operations.

## A.3 Memory Query Construction

For each laboratory video, we first extract a structured experimental trajectory according to the available temporal annotations. The trajectory consists of ordered experimental operations:

$$\mathcal{T} = \{(a_i, t_i^s, t_i^e)\}_{i=1}^{N}, \tag{9}$$

where $a_i$ denotes the experimental action and $t_i^s, t_i^e$ represent its temporal boundaries.

We sample query moments from later stages of each trajectory, where sufficient historical context is available. Specifically, a query step is selected only when at least three previous experimental operations exist.

For each query moment, all preceding operations are regarded as historical experience, while the selected operation represents the current observation. To provide compact visual evidence, we uniformly sample six frames from the complete historical trajectory and six frames from the current observation segment.

Table 5: Prompt used for memory-dependent question generation.

**The Prompt for Memory-dependent Question Generation**

You are given historical experimental observations and the current state of a laboratory experiment. Your task is to generate a memory-dependent question-answer pair for evaluating long-term laboratory memory.
The question should require recalling previous experimental experiences and reasoning about how the current experimental state was developed.

**Input:**

- Historical experimental trajectory and temporal annotations.
- Historical visual evidence sampled from previous experimental stages.
- Current observation frames corresponding to the query moment.

**Output Format:**

```
{
 "question": "...",
 "answer": "...",
 "category": "...",
 "memory_required_reason": "...",
 "evidence_steps": [...]
}
```

**Requirements:**
The generated question must require historical experimental memory.
The answer should depend on previous observations or state transitions.
The question should be grounded in the provided experimental evidence.
Questions only requiring action recognition, protocol description, or isolated visual understanding should be avoided.

Therefore, each LabVQA instance contains twelve visual observations: six historical frames describing previous experimental experience and six current frames representing the present experimental state.

### A.4 Question Generation and Annotation

Given the constructed memory queries, we generate memory-oriented question-answer pairs using a multimodal large language model. For each query instance, the model receives the historical experimental trajectory, sampled historical visual evidence, current observation frames, and corresponding temporal annotations. The generated questions are designed to evaluate whether an agent can utilize previous experimental experiences for understanding the current laboratory state. Therefore, the questions are constrained to require information beyond the current observation, such as previous operations, experimental state transitions, persistent entities, or temporal-spatial relationships. Questions that can be answered through isolated visual recognition, current action understanding, or direct protocol description are excluded. The multimodal large language model produces a structured output containing the question, reference answer, reasoning category, memory dependency explanation, and supporting evidence steps from the original experimental trajectory. The prompt used for question generation is shown in Table 5.

After generation, each question-answer pair is associated with the corresponding historical evidence steps from the original trajectory. These evidence annotations explicitly indicate which previous experimental experiences are required for answering the question, enabling the evaluation of memory utilization rather than only answer correctness.

Table 6: Question categories in LabVQA.

| Category | Description |
|---|---|
| Historical Recall | Recall previous experimental events relevant to the current state |
| State Evolution | Track how experimental states change over time |
| Temporal-Spatial Reasoning | Reason about entities and relationships across time and space |
| Experience Comparison | Compare different experimental stages or procedures |
| Integrated Memory Reasoning | Combine historical experience with current observations |

### A.5 Annotation Verification

After question generation, each question-answer pair is reviewed to ensure that it is grounded in the original experimental trajectory. The verification process checks three aspects: (1) whether the question requires information from the provided experimental evidence, (2) whether the reference answer is supported by the corresponding video observations and temporal annotations, and (3) whether the assigned reasoning category matches the required reasoning ability.

Questions that can be answered solely from isolated visual recognition or without historical experimental context are removed. The final benchmark only contains question-answer pairs that are supported by the original laboratory videos and associated annotations.

### A.6 Query Sampling Details

For each laboratory video, we construct query instances based on the temporally ordered experimental trajectory. A query is sampled from an experimental step only when sufficient historical context is available. Specifically, we require at least three previous operations before selecting a query step, ensuring that the question can involve information from earlier experimental stages. For each selected query step, the corresponding current observation is represented by the sampled frames from the target segment, while historical evidence is collected from previous operations in the same experimental trajectory. To avoid bias toward a particular temporal location, query steps are sampled across different stages of the experiment whenever possible.

### A.7 Question Categories

LabVQA covers several types of memory-dependent reasoning abilities required for long-horizon laboratory understanding. The categories are designed to characterize different ways in which historical experimental experiences contribute to interpreting the current observation. They include recalling previous experimental events, tracking state transitions, reasoning about persistent entities across time and space, comparing experimental experiences, and integrating historical evidence with current observations.

The category definitions and representative examples are summarized in Table 6.

### A.8 LabVQA Statistics and Examples

We construct LabVQA from approximately 50 laboratory videos collected from FineBio and BioVL2. Each query instance contains historical visual evidence, current observations, and a memory-dependent question-answer pair. To ensure that the benchmark evaluates long-horizon memory rather than short-term perception, each query is sampled only when sufficient historical experimental context is available. For each query instance, six frames are uniformly sampled from the historical trajectory and six frames are sampled from the current observation segment. The benchmark therefore requires models to reason over both previous experimental experiences and the current laboratory state. The overall statistics of LabVQA are summarized in Table 7.

Table 7: Statistics of the LabVQA benchmark.

| Statistics | Value | Description |
|---|---|---|
| Source videos | ∼50 | Laboratory videos from FineBio and BioVL2 |
| Source datasets | 2 | FineBio and BioVL2 |
| Visual evidence per query | 12 frames | 6 historical frames + 6 current frames |
| Historical evidence | 6 frames | Previous experimental experience |
| Current observation | 6 frames | Current experimental state |
| Minimum history | 3 steps | Previous operations required for reasoning |
| Total queries | 1050 | Memory-dependent QA instances |
| Average history length | 5.2 steps | Average number of previous steps |

Table 8: Overview of RA-Lab memory construction pipeline.

| Stage | Output |
|---|---|
| Experience Construction | Structured experimental experience |
| Visual Memory Generation | Experience-level visual memory |
| Memory Retrieval | Relevant historical experiences |
| Memory-Augmented Reasoning | Memory-enhanced prediction |

# B Implementation Details

We provide implementation details of RA-Lab and the baseline used in our experiments. We first describe the implementation of RA-Lab, including multimodal experimental experience memory, Adaptive Visual Abstraction, Targeted Multimodal Retrieval, and Memory-Augmented Reasoning. We then describe the corresponding baseline implementations details.

## B.1 RA-Lab

RA-Lab constructs multimodal experience memory from laboratory videos through three stages: experience construction, visual memory generation, and memory retrieval. For each laboratory video, we first divide the video into local 15s-segments and sample frames for multimodal processing. GPT-5.5 is used as based model to generate structured experience representations from the sampled frames and available protocol context, including the corresponding protocol step, experience summary, entities, actions, states, and temporal information. Segments belonging to the same protocol step are consolidated into a single experience entry, and the associated source frames are retained for subsequent visual abstraction.

For each experience, GPT-5.5 further generates an adaptive abstraction plan that specifies the important visual information to preserve. The abstraction plan and source frames are then provided to GPT-Image-2 to synthesize a single visual memory image with a resolution of $2048 \times 1536$. The generated visual abstraction is stored together with the structured experience as a multimodal memory entry. During inference, the current observation and task query are encoded into a retrieval representation. Each memory entry is represented by both its structured experience description and corresponding visual abstraction. All query and memory representations are embedded using VLM2Vec-V2 Meng et al. (2025). Specifically, the current observation and task query are encoded into a multimodal query representation, while each historical memory entry is encoded into a corresponding memory representation. The relevance between the query and historical experiences is computed using cosine similarity between their representations, and the top-$K$ experiences with the highest similarity scores are retrieved and provided to the reasoning model.

During inference, the same memory construction procedure is applied to historical laboratory videos. For each query, retrieved memories are concatenated with the current observation and question as input to the reasoning model. Unless otherwise specified, the number of retrieved experiences is fixed to top-$K$ according to the experimental setting.

#### B.1.1 QWEN2.5-VL

Qwen2.5-VL-7B Bai et al. (2023) is used as an open-source vision-language model baseline. We sample frames from each laboratory video segment and provide them together with the task query as input. The model performs reasoning based on the sampled visual observations without additional memory augmentation.

#### B.1.2 GPT-5.5

GPT-5.5 Hurst et al. (2024) means GPT-5.5 without Agent, which can only take images as visual input. For each task, video inputs are converted into sampled frames, which are provided together with the task query. GPT-5.5 performs reasoning directly from the sampled visual observations without external memory.

#### B.1.3 GEMINI-1.5

Gemini 1.5 pro Team et al. (2024) directly accepts each 15s-segments as input. For each task, the video segment and the task query are provided to the model for inference.

#### B.1.4 MOVIECHAT

MovieChat Song et al. (2024) is evaluated as an online video understanding baseline. It extracts frame-level features from laboratory videos using a sliding-window mechanism and maintains a hybrid memory for long-video reasoning. The LLM performs question answering conditioned on the stored video memory.

#### B.1.5 STREAMBRIDGE

StreamBridge Wang et al. (2026) is evaluated as a streaming video understanding baseline. It processes laboratory videos in a streaming manner by maintaining a memory buffer of historical visual information. During inference, the model performs question answering based on the current observation and the accumulated streaming memory.

#### B.1.6 WORLDMM

WorldMM Yeo et al. (2026) is evaluated as a dynamic multimodal memory agent for long-video reasoning. For laboratory videos, it constructs multiple complementary memories, including episodic memory for event information, semantic memory for high-level knowledge, and visual memory for detailed visual evidence. During inference, an adaptive retrieval agent selects relevant memory sources and temporal granularities according to the query, and the retrieved information is used for answer generation.

#### B.1.7 M3-AGENT

M3-Agent Long et al. (2026) is evaluated as a long-term multimodal memory agent. Given laboratory videos, it constructs long-term memory through a memorization process, including episodic and semantic memories organized in a multimodal graph. During inference, the control process retrieves relevant memories based on the task query and performs multi-round reasoning to answer laboratory video questions.

## C DETAILS OF ADAPTIVE VISUAL ABSTRACTION

### C.1 IMPLEMENTATION OF ADAPTIVE VISUAL ORGANIZATION

To analyze the effect of adaptive visual organization, we compare the proposed adaptive strategy with a fixed organization strategy under the same image generation setting. Both strategies take the same source frames and experience information as input and use GPT-Image-2 as the visual memory renderer. The difference lies in how the visual organization of the memory is determined.

Table 9: Implementation comparison between fixed and adaptive visual organization.

| | Fixed Organization | Adaptive Organization |
|---|---|---|
| Planning | No planning module | GPT-5.5 generated abstraction plan |
| Generation Input | Source frames + fixed instruction | Source frames + experience-specific plan |
| Information Selection | Predefined visual elements | Entities, actions, relationships, and states selected by experience |
| Composition | Same organization across experiences | Experience-dependent organization |
| Renderer | GPT-Image-2 | GPT-Image-2 |

Table 10: Example of a multimodal experience memory constructed by RA-Lab.

| **Component** | **Content** |
|---|---|
| Source Video | P01_01_01.mp4 |
| Experimental Step | remove_culture_medium (1.0s–20.43s) |
| Experience Summary | The operator removes culture medium from a multiwell culture plate using a pipette, preparing the sample for subsequent experimental operations. |
| Entities | Multiwell culture plate, pipette, culture medium, PBS container, and pipette tip box. |
| Action | Aspirating liquid from the culture plate and preparing the pipette for the following experimental procedure. |
| State Transition | Before: culture medium is present in the wells.<br>After: the wells become largely emptied after liquid removal. |
| Abstraction Plan | Preserve the pipette–plate interaction, liquid removal process, important laboratory entities, and the corresponding state transition.<br>Suppress repeated observations, irrelevant background objects, and redundant visual information. |

In the fixed setting, we use a predefined organization instruction for all laboratory experiences. The instruction requires the generated memory to consistently include a general laboratory overview, the main experimental objects, the performed action, and relevant state changes. Therefore, all experiences are represented using the same visual organization strategy regardless of their procedural characteristics.

In the adaptive setting, RA-Lab first generates an experience-specific abstraction plan from the structured experience and source observations. The plan specifies which entities, actions, relationships, spatial arrangements, and state changes should be preserved for the current experience. The abstraction plan is then provided to GPT-Image-2 together with the source frames to generate the final visual memory. Therefore, the visual organization is dynamically determined according to the information structure of each laboratory experience rather than following a predefined template.

The implementation differences between the two strategies are summarized in Table 9.

# D Example of RA-Lab on LabVQA

To illustrate how RA-Lab constructs multimodal experience memories from laboratory videos, we provide a representative example from the FineBio dataset. Each memory entry integrates structured experimental information and an experience-level visual abstraction. As shown in Table 10, the memory stores the experimental context, entities, actions, state transitions, and the adaptive abstraction plan used for visual memory generation. The corresponding generated visual memory is shown in Figure 6.

## D.1 Example of Memory-Augmented Laboratory Reasoning

To demonstrate how RA-Lab utilizes retrieved historical experiences for laboratory reasoning, we provide a representative LabVQA example. The example illustrates that the answer requires recalling previous experimental states and relating them to the current observation.




Figure 6: Generated visual memory for the example experience. RA-Lab compresses multiple experimental observations into an experience-level visual representation that preserves key entities, actions, and state transitions.

Table 11: Example of memory-augmented reasoning in LabVQA.

| **Component** | **Content** |
|---|---|
| Video | P01_03_01.mp4 |
| Historical Memory | The sample underwent magnetic bead addition, pipetting, vortexing, spindown, and magnetic rack placement. After the previous magnetic rack placement following the wash-buffer treatment, the supernatant was aspirated while magnetic beads remained retained. |
| Current Observation | The tube is placed in the magnetic rack after a later ethanol-wash cycle, including 70% ethanol addition, mixing, vortexing, and spindown. |
| Question | Earlier in this experiment, what happened after the bead-containing tube was placed in the magnetic rack following the wash-buffer treatment, and how should that prior experience be used to interpret the current tube's state? |
| Answer | After the previous magnetic rack placement, the supernatant was aspirated while magnetic beads remained retained in the tube. The current tube undergoes an analogous separation stage after ethanol treatment, where beads are expected to be retained by the magnet while the liquid phase can be removed. |
| Reason | The answer requires recalling the previous wash-buffer separation cycle and applying the observed temporal state transition to the current ethanol-wash stage. |

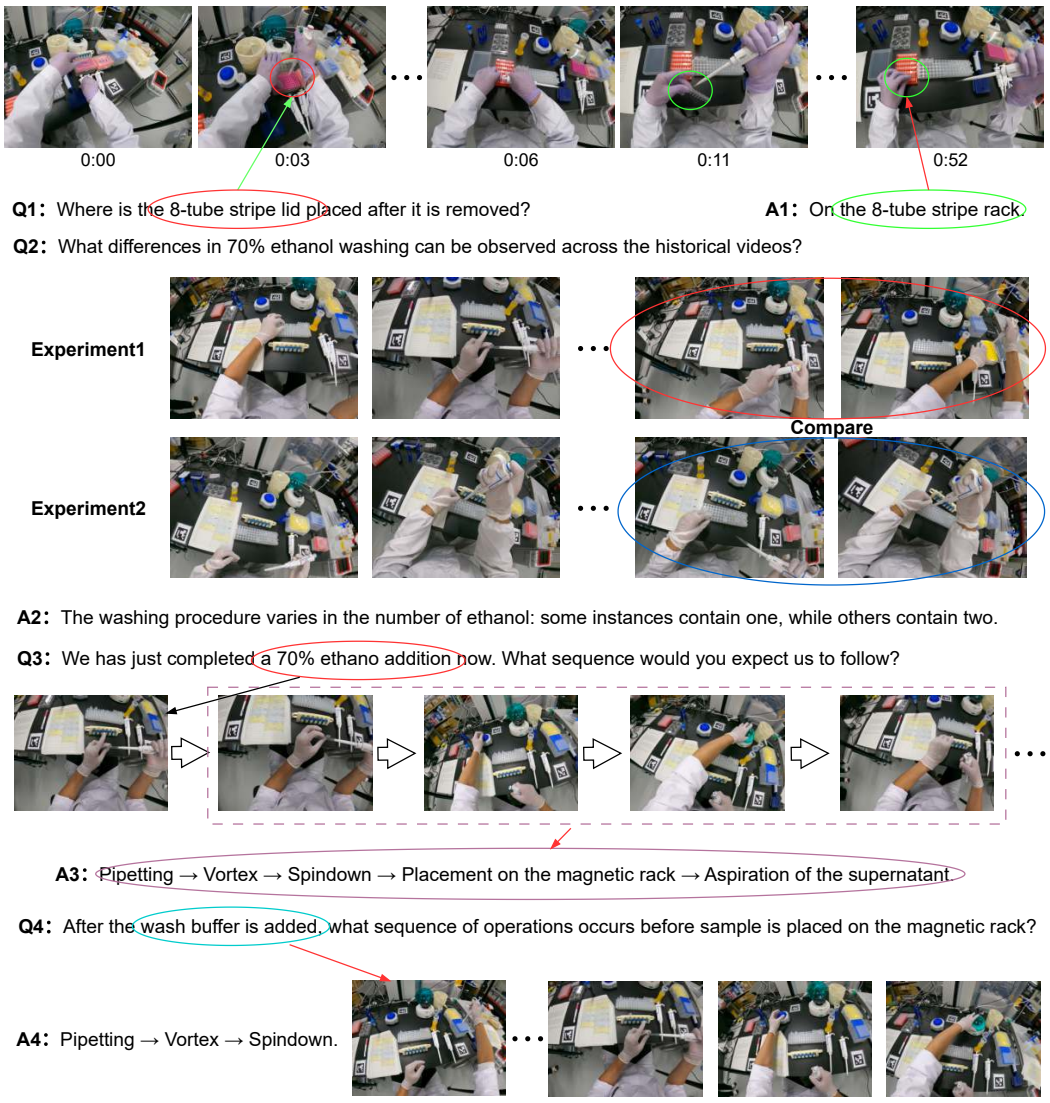


Figure 7: **Examples of LabVQA tasks.** The figure presents representative questions requiring reasoning over historical and current laboratory observations. The examples include historical object-location retrieval (Q1), cross-experience procedural comparison (Q2), current–historical reasoning for next-step prediction (Q3), and temporal reasoning over sequential operations (Q4). The highlighted visual evidence demonstrates that answers rely on fine-grained actions, spatial relationships, and temporal dependencies across laboratory experiences.

### D.2 Examples of LabVQA Tasks

To better illustrate the design of LabVQA, we provide representative examples of memory-oriented laboratory video question answering in Figure **??**. Different from conventional video QA tasks that mainly focus on recognizing visible actions from the current observation, LabVQA requires models to utilize historical experimental observations and temporal context to answer questions. The examples cover several types of memory-dependent reasoning, including retrieving historical object locations, comparing procedural variations across different experiences, predicting subsequent operations based on previous experimental trajectories, and reasoning over temporal dependencies among sequential operations. These examples demonstrate that successful answering requires integrating visual evidence from both historical and current laboratory observations.

# E LIMITATIONS

**Dependence on temporal segmentation and protocol alignment.** The current memory construction pipeline relies on fixed-length video segmentation and protocol-guided step alignment. Although this design provides a practical way to organize continuous laboratory videos into step-level experiences, fixed temporal boundaries may introduce ambiguity and lead to incomplete or inaccurate assignment of video segments to experimental steps. Moreover, the alignment process depends on the capability of the underlying multimodal models to understand experimental procedures and match visual observations with protocol descriptions.

**Dependence on structured protocol information.** RA-Lab currently benefits from available experimental protocols during memory construction, which provide semantic guidance for organizing laboratory experiences. However, in scenarios without structured procedural annotations, the quality of experience construction and memory organization may degrade. Future work may explore protocol-free experience discovery and automatic procedure understanding from raw laboratory observations.

**Information loss in visual abstraction.** Although Adaptive Visual Abstraction aims to preserve important visual evidence while reducing redundancy, the generated visual memories may inevitably omit certain fine-grained details during the abstraction process. Future work may investigate more faithful visual memory representations that better preserve detailed experimental evidence while maintaining compactness.