SciEducator: Scientific Video Understanding and Educating via Deming-Cycle Multi-Agent System

Supplementary Material

A. Tools & Agents

Our system integrates 10 agents and 6 tools, each designed to handle specific tasks with clearly defined input and output specifications to ensure precise execution and seamless integration. The set of tools and agents is categorized into two groups: (i) dynamically invocable components and (ii) fixed-execution components, described as follows.

A.1. Dynamically Invocable Tools/Agents

(i) Planning and Answer Generation (Planner Agent): Our system uses a Planner Agent configured with GPT-4o as its core. Leveraging the powerful text processing and knowledge reasoning capabilities of the LLM, it is responsible for generating and adjusting the solution pool within the system, consolidating all acquired information, evaluating confidence level, and outputting the final answer.

(ii) Video Content Acquisition (Captioner Agent): Our system uses a Captioner Agent configured with Gemini 2.0 Flash to obtain the content of video frames and generate textual descriptions of the video.

(iii) Solution Evaluation (Evaluator Agent): Our system uses an Evaluator Agent configured with GPT-4o to evaluate the various solutions in the solution pool. It selects the best solution from the current pool by considering both subjective and objective metrics and initiates its execution.

(iv) Web Content Search (Web Search Agent): We use the open-source web content search model searchGPT, based on the Google Search engine, to search for web links related to input keywords and organize the web information for output.

(v) Paper Search (Paper Search Agent): We develop a Paper Search Agent configured with GPT-4o for searching academic paper content. It can search for papers related to input keywords across major scientific paper platforms, organize and summarize the found content, and finally output the results.

(vi) Video Super-Resolution Tool (VideoSR Tool): We built a tool for performing super-resolution on video frames based on an open-source model. This enhances low-resolution or blurry frames, ultimately improving the description of the video content.

(vii) Experimental Procedure Search (Procedure Search Agent): We develop a Procedure Search Agent based on the open-source web content search model searchGPT. It searches for corresponding experimental procedures based on input experimental terms or descriptions, organizes the web information, and outputs the results.

(viii) Key Entity Recognition in Content (Entity Recognition Agent): SciEducator uses an Entity Recognition Agent configured with GPT-4o to identify key experimental instruments and materials involved in the current experimental steps, and outputs them in a specified format.

(ix) Experimental Precautions Prompter (Safety Alert Agent): We develop a Safety Alert Agent based on the open-source web content search model searchGPT. It comprehensively searches for corresponding safety precautions based on the current experimental content and equipment, organizes the information, and outputs it to alert readers to avoid potential hazards when conducting the experiment.

A.2. Fixed-Execution Tools/Agents

(i) Knowledge Base Construction and Storage: We develop tools for knowledge base construction and storage. These tools archive knowledge texts, invoke an embedding model to create and store query vectors, thereby facilitating subsequent retrieval. The construction and storage of the knowledge base must be completed before the DCAgent operates.

(ii) Knowledge Base Retrieval (RAG Agent): We construct a small-scale knowledge base containing basic scientific knowledge and fundamental material properties (e.g., electromagnetic induction, properties of air). We configured a Retrieval-Augmented Generation (RAG) Agent based on GPT-4o. Based on the video content, it can extract important keywords, retrieve the most relevant content from the knowledge base based on these keywords, and output an organized summary. If no relevant content exists in the knowledge base, it will output a statement indicating the lack of relevant knowledge.

(iii) IDF Value Calculation for Keywords in Solutions (IDF Calculator Tool): We independently develop a tool capable of calculating the Inverse Document Frequency of each keyword within the knowledge base, indicating the uniqueness of each keyword and its importance to the video content. It is fixedly invoked each time the Evaluator Agent evaluates the solutions in the solution pool. The Evaluator Agent extracts keywords from each solution and inputs them into the IDF Calculator Tool for computation.

(iv) Experimental Equipment Image and Purchase Link Search (Equipment Search Tool): We independently develop a tool for searching images and purchase links of experimental equipment, typically returning one image and one most relevant link per equipment item.

(v) Experimental Procedure Illustration Generation (Illustration Generation Tool): We utilize the Gemini 2.5 Flash Image (Nano Banana) API to generate illustrations of experimental procedures by inputting descriptive text of the steps.

(vi) Text-to-Speech Conversion (Speech Generation Tool): We employ the open-source tool KittenTTS to convert experiment-related text into speech.

(vii) Multi-modal Information Integration for Structurally Organized and Aesthetically Arranged E-booklet Compilation (E-booklet Generation Agent): We independently developed an E-booklet Generation Agent for integrating multi-modal information related to experiments, such as text, images, links, and audio. Implemented in JavaScript/CSS,the agent composes an HTML e-booklet with clear hierarchy and an aesthetically pleasing layout to stimulate readers' scientific interest in a vivid and engaging manner.

B. Details about Empirical Prior

To obtain the resource consumption and the probability of returning expected results for dynamically invocable tools/agents, we build an empirical prior $\mathcal{E}$ by issuing 20 randomized probe calls per tool/agent, from which Evaluator Agent obtains average latency, average token usage, and success probability. Token usage is directly converted into the corresponding platform's API money cost. Results:

Web Search Agent: On average, each call takes 21.89s, costs \$0.0139, and 88% success rate.

Paper Search Agent: On average, each call takes 17.41s, costs \$0.010, and 75% success rate.

Captioner Agent: On average, each call takes about 25s if the frame rate is 1 FPS, costing \$0.0001; both time and financial cost increase in direct proportion to the FPS, 93% success rate.

VideoSR Tool: Approximately 53s if the frame rate is 1 FPS, costing \$0.0001; both time and financial cost increase in direct proportion to the FPS, 100% success rate.

Procedure Search Agent: On average, each call takes 28.17 seconds, costs \$0.0178, and 87% success rate.

Entity Recognition Agent: On average, each call takes 10.45 seconds, costs \$0.0064, and 99% success rate.

Safety Alert Agent: On average, each call takes 26.29 seconds, costs \$0.0153, and 72% success rate.

Based on our practical considerations, we combine time consumption (t) and financial cost (c) into a total cost: t + 1000c. This is provided to the Evaluator Agent as a reference. The Evaluator Agent is instructed to give equal weight to resource consumption and feasibility considerations, and ultimately to select the overall optimal solution.

C. Evaluation Metrics

C.1. Understanding

We employ Qwen3-Max to uniformly evaluate all model-generated responses. We first provide the reference answers and substantial scientific background regarding the query questions, instructing Qwen3-Max to assess the relevance of the model-generated answers to the query, focusing on how well the response aligns with the scientific subdomain involved in the question, regardless of correctness. The objective is to determine whether the model provides misleading or irrelevant information. The detailed scoring strategy is as follows:

- 1.If the answer is entirely relevant to the subfield of the question, it receives a relevance score of 1.
- 2.If an answer contains partially irrelevant content or is only broadly related to the field, it receives a score of 0.5.
- 3.If it is completely irrelevant, it receives a score of 0.

Prompt: Evaluating Relevance

You are a professional scoring teacher and evaluation expert. Your task is to evaluate the relevance of responses based on the reference answer and the scientific background related of question, using three scoring options: 1 point, 0.5 points, or 0 points.

Please follow these scoring criteria in order of priority from highest to lowest:

1.If the answer is entirely relevant to the specific subfield of the question, award 1 point.
2.If the answer contains partially irrelevant content or is only broadly related to the field of the question, assign 0.5 points.
3.If the answer is completely irrelevant, assign 0 points.

Please provide the points directly, only the number, no other output.

We then use Qwen3-Max to analyze the semantic similarity between generated answers and reference answers, scoring the accuracy of model responses. The detailed scoring strategy is as follows:

- 1.If the model's response contains absolute errors—including numerical inaccuracies (e.g., the correct answer is 3, but the model answers 2) or terminological mistakes (e.g., the correct answer is "Magnus Effect," but the model answers "Bernoulli's Principle")—it receives a score of 0 directly.
- 2. Qwen3-Max is prompted to analyze the key points of the reference answer and determine whether the model's response covers all of them. If no key point is covered, the score remains 0. If some key points are covered and the remaining content contains no absolute errors, a score

of 0.5 is assigned.
- 3. If all key points are covered and no absolute errors are present in the remaining content, a full score of 1 is awarded.

Prompt: Evaluating Accuracy

You are a professional scoring teacher and evaluation expert. Your task is to evaluate the quality of responses based on reference answers, using three scoring options: 1 point, 0.5 points, or 0 points.

Please follow these scoring criteria in order of priority from highest to lowest:

1. First check for any absolute errors that contradict the reference answers, such as instances where the reference answer deems something correct but the response considers it incorrect, or numerical mistakes; if such errors occur, award 0 points.
2. Then analyze the key points of the reference answers and compare them to the responses, noting that wording can differ but the meaning must be strictly identical to count as a match; if the response fully covers all key points of the reference answers, give 1 point, and additional correct content without errors can still warrant 1 point.
3. Finally, if the response matches only part of the key points and contains no absolute errors, assign 0.5 points, but exercise caution when awarding this score. Don't be too rigid, you should fully refer to the different expressions of the standard answer. If there are more details than the standard answer, it should be considered correct

Please provide the points directly, only the number, no other output.

C.2. Educating

We employ four metrics and uniformly use Qwen-VL-Plus to evaluate all model responses in a comparative setting. Qwen-VL-Plus receives supplied with substantial background information about each experiment as a reference. Specifically, the metrics are:

- **Relevance**: How well the generated experimental procedures and precautions align with the current experiment and its underlying principles.
- **Instructional Quality (IQ)**: How effectively the generated procedures and precautions guide children in conducting the experiment, with emphasis on detail orientation, completeness, clarity, and safety warnings.
- **Attractiveness**: A comprehensive assessment of how engaging the textual instructions are. For SciEducator, the aesthetic quality of its supporting illustrations is also incorporated into this evaluation to identify the most captivating response.
- **Educational Value (EV)**: How well each model's response stimulates children's scientific interest and guides them to understand the principles through the experiment.

Prompt: Evaluating Performance in Education

You are a fair and professional evaluation expert. Several models have generated experimental procedures and safety precautions for a specific scientific phenomenon. Your task is to compare the responses produced by these different models from four aspects and select the best-performing model for each aspect.

You will be provided with a description of the scientific phenomenon and substantial background information about it as a reference. The four aspects are as follows:

Relevance: How well each model's generated experimental procedures and precautions align with the current experiment and its underlying principles.
Instructional Quality: How effectively each model's generated procedures and precautions guide children in conducting the experiment, with emphasis on detail orientation, completeness, clarity, and safety warnings.
Attractiveness: How engaging the responses are. If any response contains images, the images should also be included in the evaluation.
Educational Value: How well each model's response stimulates children's scientific interest and guides them to understand the principles through the experiment.

Please select the best-performing model for each aspect, referring to the models by their names, and output the result in the following format:
[
{
"Relevance": model name,
"Instructional Quality": model name,
"Attractiveness": model name,
"Educational Value": model name
}
]

D. Main Prompts

tcolorbox

D.1. Plan Stage

Prompt: Plan Stage

"""
User Query: {user_query}
Video Path: {video_path}
Video Content Description: {video_description}
RAG Search Results: {rag_result}

You are a scientific video understanding task planning expert. Your responsibilities are:
1. Generate multiple possible solutions based on video information, user queries, and retrieved knowledge
2. Each solution should include specific descriptions, tool call sequences, and parameters for each tool input

Please output a list of solutions directly in JSON format without any additional text:
[
  {
"Number": Solution number (integer),
"description": "Solution description (detailed explanation of the solution process)",
"steps": [
      {
"tool": "Name of the tool to call (must use one of the following recognized tool names: WebSearch, PaperSearch, Captioner, VideoSR)",
"input": "Tool input parameters (clear and specific input content)"
      }
        ]
  }
]

Tool Description:
- WebSearch: Web search tool, can input a query sentence
- PaperSearch: Academic paper search tool, can input a query sentence
- VideoProcessor: Video understanding tool, inputs include a video path, number of segments to divide the video into, frames per second to process, and information to search for in the video
- VideoSR: Video super-resolution tool, input parameters are the same as VideoProcessor, performs super-resolution on input frames before understanding

Do not include any other fields or text, output only JSON objects.

Note:
1. Your solutions should be as logical and reasonably sequenced as possible, be relevant to user queries, striving to preserve the substances, environment, and specific movements of phenomena occurring in the video. Including the experimental objects and environment in tool input parameters may increase the success probability of the solution.
2. You should output as many possible solutions as possible to facilitate unified evaluation and improve success rate. Your output solutions may have more details, be more complete, and have longer call chains than our examples.
3. Please carefully check whether each step of your solution is conducive to solving the problem, such as querying the miraculous movements, changes, or interactions of objects, the environment or solution in which phenomena occur, etc. Make good use of each tool to obtain information and reduce ineffective calls.
4. Do not rashly conclude what phenomenon the video describes before you have obtained sufficient information. The parameters input to all tools in your solution should be as specific as possible. For example, if the video phenomenon occurs in mate tea, and you want to query the interaction between solid and liquid, or the peculiar movement of solids in liquid, we recommend that you always change the liquid here to mate tea, i.e., "peculiar movement of solids in mate tea".
5. Note that you are working on a scientific phenomenon understanding task. Your solutions should aim to clarify the scientific phenomena in the video content description.

Please strictly adhere to the requirements
"""

D.2. Do Stage

Prompt: Do Stage

"""
You are a professional solution evaluation expert. Please evaluate multiple solutions and select the best one based on the following information:
User Query: {user_query}
RAG Search Results: {rag_result}
Previous Solution Results: {previous_results}
Here are the solutions to evaluate along with their keyword IDF values (higher IDF values indicate more unique keywords): {plans_with_idf}

Please follow these evaluation steps: 1. Analyze the feasibility and relevance of each solution. Evaluate feasibility and relevance according to the following criteria:
2. Consider the IDF values of keywords - higher IDF keywords may be more representative. Then consider the frequency of keywords in the video content (TF values) - higher TF keywords may be more representative.
3. Evaluate the completeness and logical coherence of each solution. Evaluate whether each solution contains scientific phenomena — such as "Why can a person swing higher despite no energy being added?"
4. Comprehensively consider the time consumption and financial cost of tools called in each solution, along with the probability of returning expected results of them. Below is the tool information list:
WebSearch: 21.89s, $0.0139, 88% success rate
PaperSearch: 17.41s, $0.010, 75% success rate
VideoProcessor: 25s if the frame rate is 1 FPS, costing $0.0001; both time and financial cost increase in direct proportion to the FPS, with a 93% success rate.
VideoSR: 53s if frame per second = 1, $0.0001, both time and financial cost increase in direct proportion to the FPS, with a 100% success rate.

We roughly combine time consumption (t) and financial cost (c) into total consumption: t + 1000c. You can reference this value for judgment

5. Considering all steps, giving equal weight to resource consumption and feasibility considerations, and ultimately to select the overall optimal solution.

Please output the best solution directly in JSON format without any additional text:
{
"best_plan": {
"Number": "solution number",
"description": "solution description",
"Why": reasons you choose it as the best plan and reasons other plans are not good
"steps": [
{
"tool": "tool name",
"input": "input parameters"
}
]
}
}
"""

D.3. Study & Act Stage

Prompt: Study & Act Stage

"""
User Query: {user_query}
Video Path: {video_path}
Video Content: {video_descriptions}
RAG Search Results: {rag_result}
Historical Execution Results and Current Execution Result: {history.get('previous_results', [])}
All Available Solutions: {json.dumps(all_plans, ensure_ascii=False, indent=2)}
executing plan number:{Number}

Please analyze the reasons for the failure or poor outcome of this execution, summarize the existing execution results and new knowledge/information, adjust previous solutions, and attempt to generate completely new solutions. Finally, compile them into a new solution collection called new_plans.

Return a JSON object containing:
{
"failure_analysis": "Analysis of failure reasons",
"knowledge_summary": "Summary of knowledge and information contained in existing execution results",
"new_plans": "List of new solutions (same format as the original solution list)"
}

Tool Description:
- WebSearch: Web search tool, can input a query sentence
- PaperSearch: Academic paper search tool, can input a query sentence
- VideoProcessor: Video understanding tool, inputs include a video path, number of segments to divide the video into, frames per second to process, and information to search for in the video
- VideoSR: Video super-resolution tool, input parameters are the same as VideoProcessor, performs super-resolution on input frames before understanding

Here are some examples you can reference for adjusting or generating solutions based on failure reasons:
1.If you find that video information is insufficient, you can adjust or generate solutions by calling the VideoProcessor tool, increasing the input frame rate, and reducing the number of segments to obtain more information. However, you should not let VideoProcessor continue searching for things in existing solution as you might be going off track
2.If you need specific video information to determine the final answer, you can adjust or generate solutions by calling the VideoProcessor tool and specifying the questions you want to query
3.If the VideoProcessor tool returns that the video is blurry and affecting answer generation, you can call VideoSR to clarify the video and then understand.
4.If WebSearch or PaperSearch did not return expected information or timeout, you can try adjusting your search keywords
5.If the search returns information that is too broad, you can try using more precise keywords for searching
6.If the search returns several completely different scenarios and you need more details to determine the final answer, you can call appropriate tools to obtain them
7.If a particular tool suddenly fails, you should reduce the use of that tool

Note: You don't necessarily have to generate new solutions every time; it's also acceptable to simply discard failed solutions.
"""

E. More Visualization Result

In this section, we provide additional visualization results to further demonstrate SciEducator's capabilities in both scientific video understanding and educational content generation.

E.1. Scientific Video Understanding

SciEducator leverages a unique iterative self-evolving mechanism rooted in the Deming Cycle (Plan-Do-Study-Act) to achieve rigorous step-wise reasoning Fig. 7. Unlike standard MLLMs that may hallucinate or provide superficial answers, our system integrates external professional

knowledge and performs failure analysis to refine its understanding. This capability serves as the foundation for the subsequent educational stage, ensuring that the scientific principles identified (e.g., light refraction, chemical reactions) are accurate before being translated into educational materials.

E.2. Educational E-booklet Generation

A core innovation of SciEducator is its ability to generate comprehensive, child-friendly educational E-booklets, which is shown in Fig. 8. The system organizes multimodal content—including text, diagrams, and safety alerts—into a structured format that fosters engagement and safety. The full E-booklet is visualized in five segments:

- The booklet begins with an engaging title and an "Interesting Introduction" designed to capture the learner's curiosity. As shown in the visualization, the system uses evocative language (e.g., "a tiny scientist, exploring the wonders") to transform complex scientific concepts into an accessible narrative, setting the stage for the experiment.
- The E-booklet describes the detailed experimental materials and corresponding pictures and shopping links, which are helpful for the rapid commencement of the experiment.
- The E-booklet provides a step-by-step, detailed description of the specific experimental procedures, along with corresponding illustrations, to assist in conducting the experiments.
- The E-booklet also provides important notes, which highlight some dangerous actions and operations that may lead to experimental failures, thereby ensuring the safety of experiments.
- The E-booklet concludes with a concise summary that reinforces core concepts and takeaways.

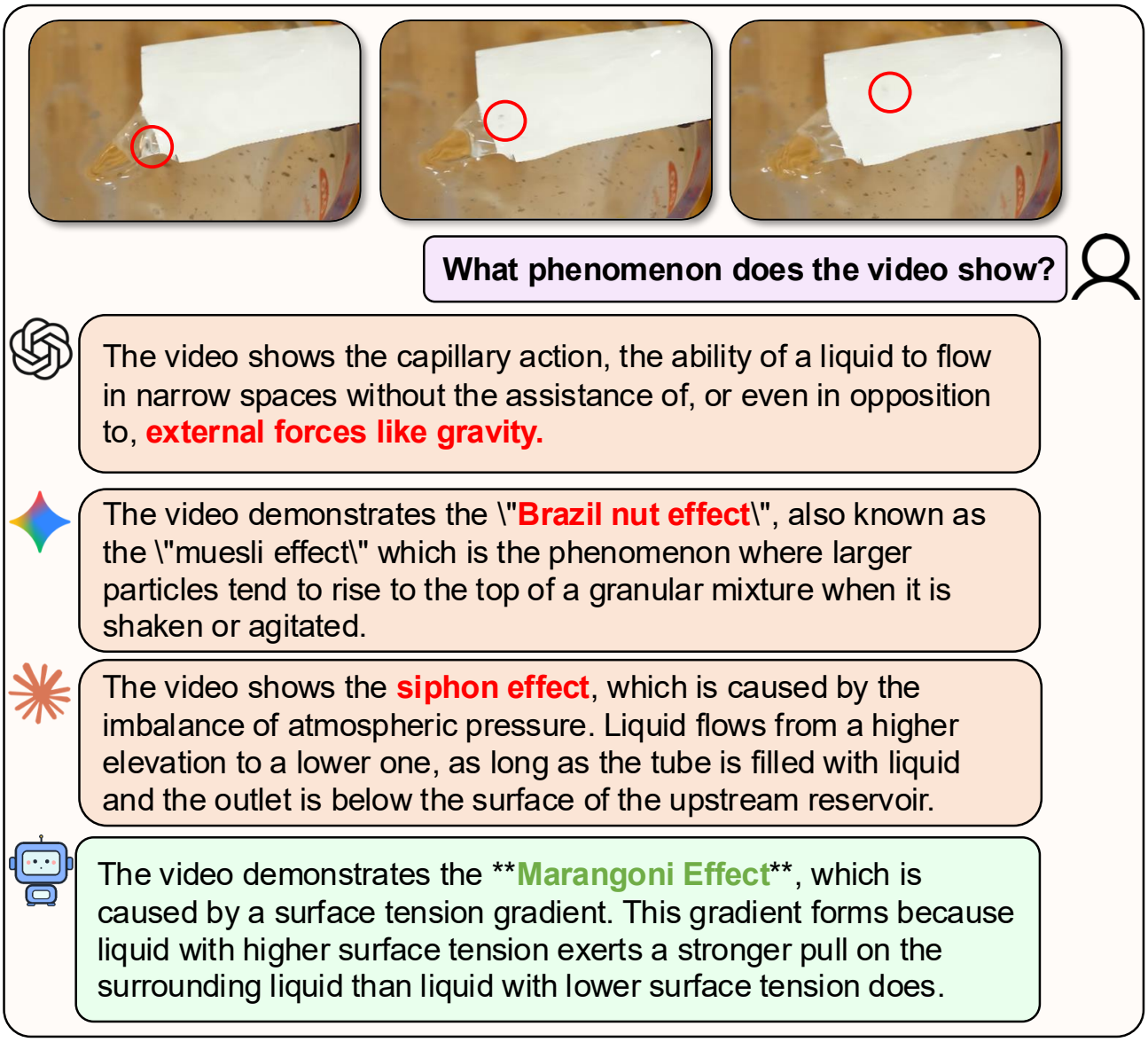

Figure 7. Qualitative comparison between SciEducator and MLLMs. These examples demonstrate SciEducator's ability to generate more comprehensive, better-structured, and more logically coherent answers than the other MLLMs.

Splitting Water: Hydrogen and Oxygen Adventure!

Introduction

Welcome to the Splitting Water: Hydrogen and Oxygen Adventure! Get ready to dive into an exciting world where water transforms into two invisible, magical gases—hydrogen and oxygen. Imagine being a tiny scientist, exploring the wonders of water with just a battery, some pencils, and a sprinkle of science magic! As you set up this simple experiment, you'll witness the power of electricity as it bravely separates water into its two best friends, hydrogen and oxygen, right before your eyes. So, grab your lab coat and goggles, and let's embark on this bubbly journey of discovery!

Figure 8. Our generated e-booklet with comprehensive contents and a well-organized structure. These examples demonstrate SciEducator's ability to generate more comprehensive, better-structured, and more attractive Education E-booklet.
(a) About Title and Interesting Introduction

Materials

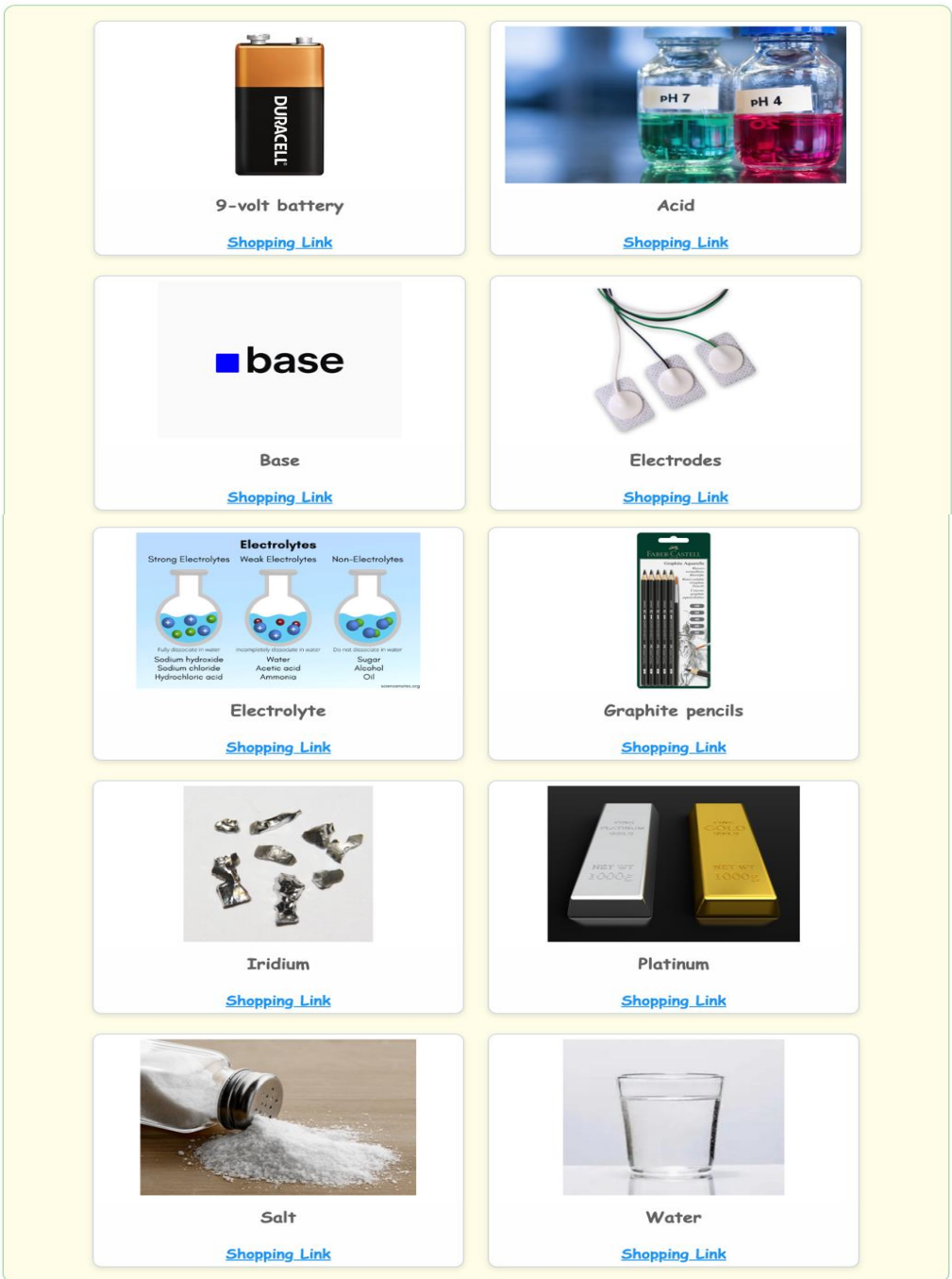

Figure 8. Our generated e-booklet (Continued).
(b) Experiments Materials List

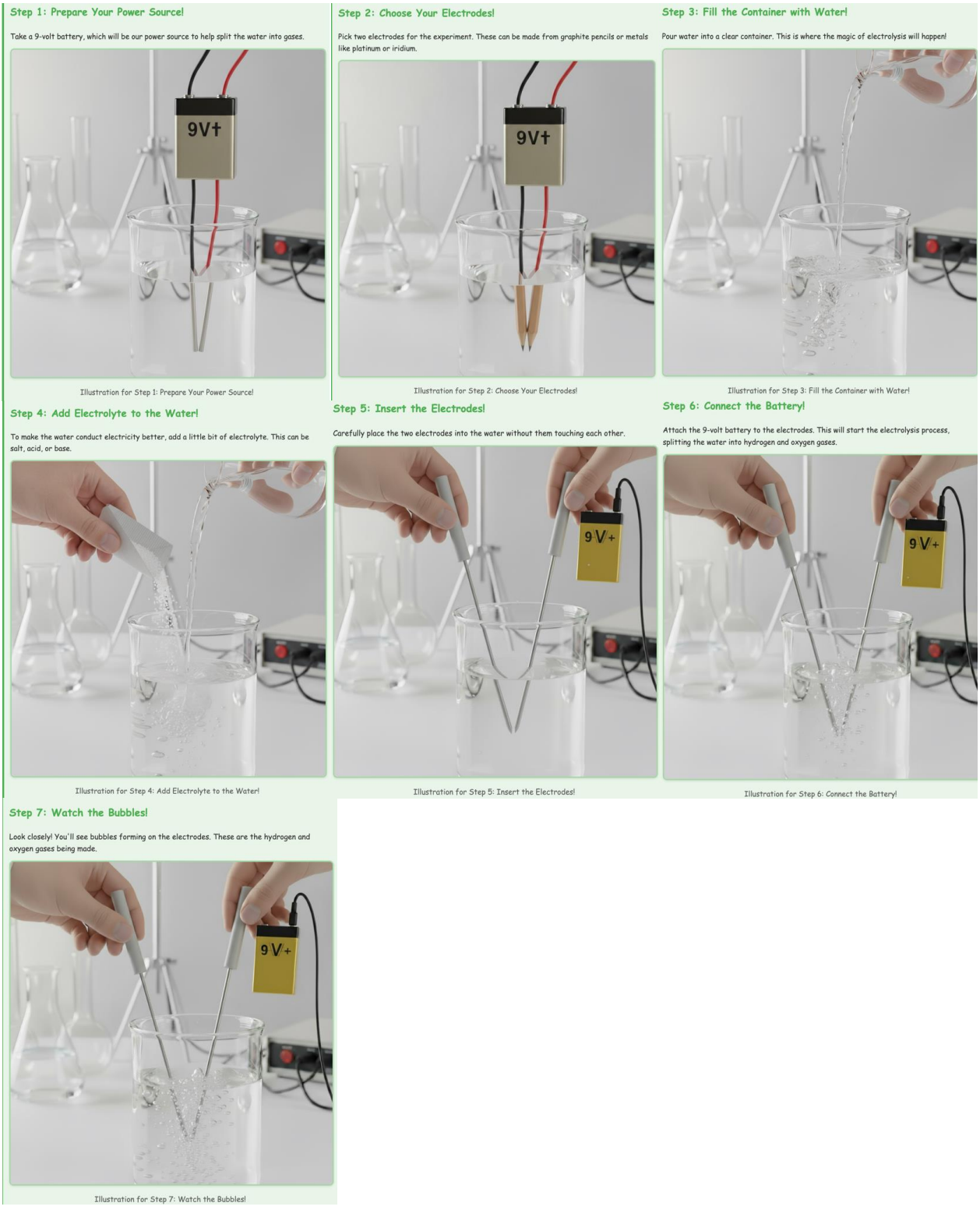

Figure 8. Our generated e-booklet (Continued).

(c) Experiments steps

Important Notes

- Use inert materials like graphite pencils or metals such as platinum or iridium for electrodes to prevent unwanted reactions.

- Ensure the use of a direct current (DC) electrical power source, such as a 9-volt battery, to provide the necessary energy for the reaction.

- Add an appropriate electrolyte, such as a salt, acid, or base, to the water to enhance conductivity and improve the efficiency of the electrolysis process.

- Be aware that the overall reaction is endothermic, requiring an input of energy to proceed.

- Handle all materials, especially the electrolyte, with care to avoid skin contact or ingestion.

- Conduct the experiment in a well-ventilated area to safely disperse the hydrogen and oxygen gases produced.

- Ensure that the setup is stable and secure to prevent accidental spills or short circuits.

- Monitor the experiment closely to ensure that the electrodes are functioning correctly and that the circuit is complete.

- Be cautious of the potential risks associated with the production of hydrogen gas, which is flammable, and take measures to avoid ignition sources.

- Properly dispose of any waste materials and clean the equipment thoroughly after the experiment.

Figure 8. Our generated e-booklet (Continued).
(d) Important Notes

What Did We Learn?

And there you have it, young scientists! You've just completed a fantastic journey, splitting water into hydrogen and oxygen with the power of electricity. Isn't it amazing how something so simple can reveal the secrets of the universe? Remember, every great scientist started with curiosity and experiments just like this one. So, keep exploring, asking questions, and discovering new things. Who knows what incredible adventures await you in the world of science? Keep dreaming big, and let your imagination take you on more exciting journeys!

Figure 8. Our generated e-booklet (Continued).
(e) Summary

F. Extra Information

F.1. Knowledge Base

The knowledge base contains fundamental scientific concepts and detailed explanations in physics and chemistry. For instance, in physics, it covers topics such as Newton's second law, electromagnetic induction, and thermal expansion and contraction. In chemistry, it includes the combustion of metals and the properties of gases in the air. It also incorporates essential physics formulas and chemical equations. The scope of knowledge spans basic science from middle school to high school levels. Beyond this, it strictly excludes any advanced knowledge or uncommon scientific phenomena. The knowledge base is structured into 84 chapters, each with a specific theme. Its primary purpose is to serve as a foundational reference document corpus for IDF (Inverse Document Frequency) value calculation. Additionally, it aims to mitigate hallucinations produced by large language models during the plan stage by providing them with a fundamental knowledge context. We utilize the text-embedding-3-large model to compute and store embeddings for each chapter of the knowledge base. This allows for rapid vector retrieval in subsequent system runs, eliminating the need to recompute the knowledge base document embeddings each time. The concept behind this knowledge base can be applied to other fields as well, and is not limited to the domain of scientific video understanding.

F.2. Average Cost Per Question

We measure the average time consumption and token (monetary) cost per question for SciEducator when the maximum number of PDSA Cycle iterations during video understanding is set to 1, 3, and 5. The results are shown below. Note that in addition to the time and tokens consumed by the PDSA Cycle itself, a fixed call to the Captioner Agent is required beforehand to obtain an initial video description.

- Maximum PDSA rounds = 1: Average time consumption per question is about 105s, with a money cost of $0.0542.
- Maximum PDSA rounds = 3: Average time consumption per question is about 158s, with a money cost of $0.0783.
- Maximum PDSA rounds = 5: Average time consumption per question is about 206s, with a money cost of $0.1051.

All money costs are automatically calculated by the platform of the APIs we call.

F.3. SciVBench Dataset

We provide some statistical analyses of videos and QA pairs in SciVBench. Fig. 9 presents statistics on the average video duration. Fig. 10 and Fig. 11 show statistical analyses of question and answer lengths. All three figures are presented separately for physics, chemistry, and daily life.

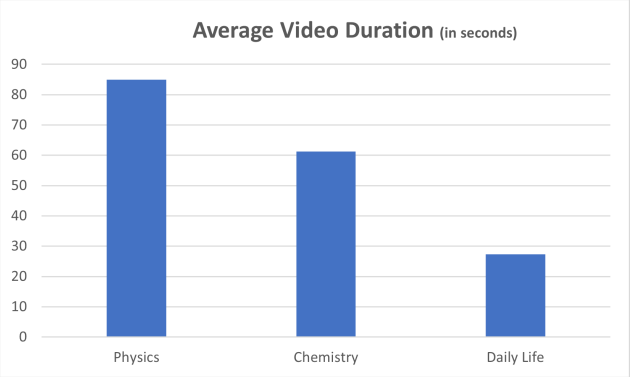

Figure 9. Statistics of average duration for three video categories in SciVBench.

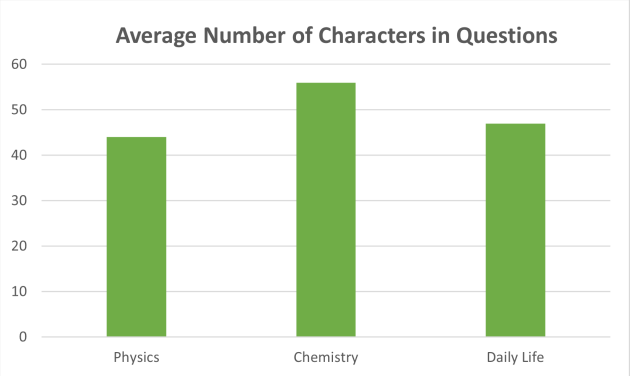

Figure 10. Statistics of average question character length for three video QA categories in SciVBench.

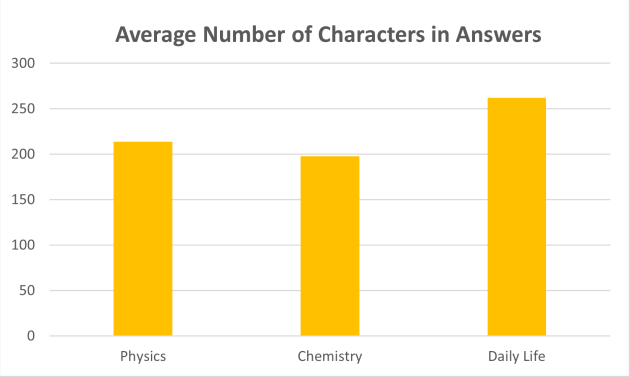

Figure 11. Statistics of average answer character length for three video QA categories in SciVBench.