This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Zhiyu Xu[1], Weilong Yan[2], Yufei Shi[3], Xin Meng[4], Tao He[5],
Huiping Zhuang[6], Ming Li[7*], Hehe Fan[8]

[1]Jinan University [2]National University of Singapore [3]Nanyang Technological University
[4]Peking University [5]University of Electronic Science and Technology of China
[6]South China University of Technology [8]Zhejiang University
[7]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)

Abstract

Recent advancements in multimodal large language models (MLLMs) and video agent systems have significantly improved general video understanding. However, when applied to scientific video understanding and educating*—a domain that demands external professional knowledge integration and rigorous step-wise reasoning—existing approaches often struggle. To bridge this gap, we propose SciEducator, the first iterative self-evolving multi-agent system for scientific video comprehension and education. Rooted in the Deming Cycle from management science, our design reformulates its* Plan–Do–Study–Act *philosophy into a self-evolving reasoning and feedback mechanism, which facilitates the interpretation of intricate scientific activities in videos. Moreover, SciEducator can produce multimodal educational content tailored to specific scientific processes, including textual instructions, visual guides, audio narrations, and interactive references. To support evaluation, we construct SciVBench, a benchmark consisting of 500 expert-verified and literature-grounded science QA pairs across five categories, covering physical, chemical, and everyday phenomena. Extensive experiments demonstrate that SciEducator substantially outperforms leading closed-source MLLMs and state-of-the-art video agents on the benchmark, establishing a new paradigm for the community. Our code and data will be available at SciEducator.*

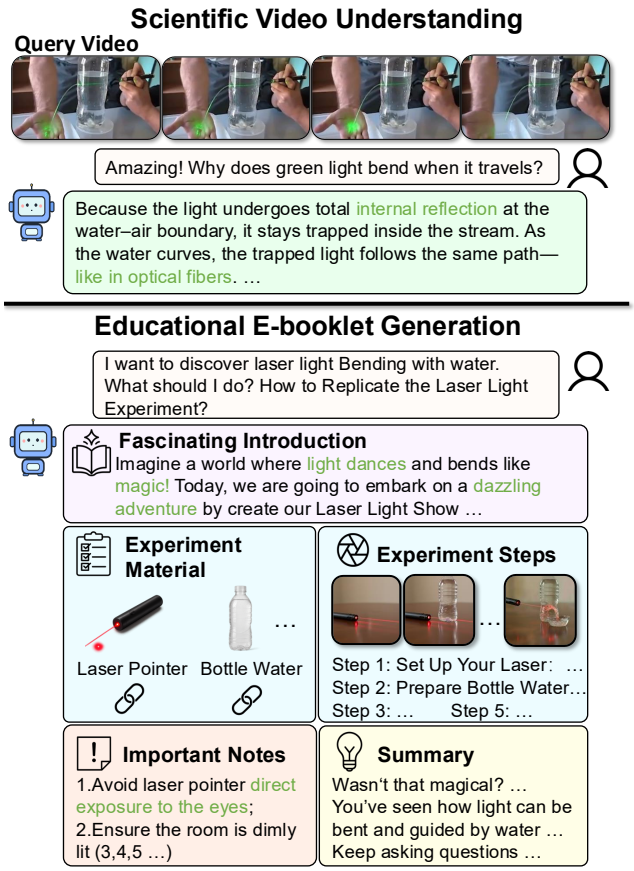


Figure 1. SciEducator, conducting video comprehension and delivering science education, can generate multimodal educational e-booklets that provide detailed, and engaging guidance.

1. Introduction

Existing MLLMs for video have achieved significant progress by focusing on enabling large models to understand videos through the integration of visual encoders, large language models, and temporal modeling, thereby allowing them to perceive, reason about, and discuss dynamic visual content [3, 53, 54]. However, they suffer from a key limitation: inadequate ability to leverage external resources

⋆ Corresponding author.

and integrate tools [8, 52], which substantially constrains their functionality and application scenarios.

Recent agentic systems, capable of integrating and effectively utilizing external tools to ultimately achieve complex and powerful functionalities, thus attracting significant attention [37, 63]. For instance, [43] proposed a PostAgent that can transform a research paper into a well-structured poster, while [44] introduced an agentic system integrating multiple agents capable of fully autonomous software development. These examples demonstrate the substantial application potential of agentic systems. However, they still face several issues at the current stage. Their performance is affected by inherent problems of large language models, such as hallucinations [57] and unstable capabilities [60]. Furthermore, for scientific video tasks that require the integration of external professional knowledge and rigorous step-by-step reasoning, these systems often struggle to generate effective and feasible plans initially and still lack the systematic mechanism to self-evolve and self-optimize their workflows based on previous execution results.

To address the limitations of existing agentic systems, we propose SciEducator, a novel multi-agent system for scientific video understanding and educating, where external professional knowledge integration, complex task planning, and rigorous reasoning are essential. Rooted in the celebrated *Deming Cycle* [41, 50] from management science, SciEducator introduces an iterative optimization mechanism that enables self-evolving reasoning and continuous enhancement. By repeatedly planning, executing, evaluating, and refining its workflows, the system progressively converges toward high-confidence and accurate interpretations as well as reliable educational content generation of complex scientific activities.

Our SciEducator operates in two stages: *understanding* and *educating*. In the *understanding* stage, an LLM-based planner acts as the central controller. It retrieves domain knowledge from both internal and external sources, formulates multiple candidate workflows involving various agents (e.g., Video Content Acquisition Agent, Web or Academic Paper Search Agent), and stores them in a solution pool. An LLM-based evaluator then integrates diverse metrics to assess each candidate and select the optimal workflow for execution. The execution results undergo a confidence evaluation: if the planner deems the current results sufficiently convincing, it synthesizes the available information to produce the final output; otherwise, SciEducator proceeds with further processing. The execution outcomes and feedback, especially summarized failure reasons and newly acquired knowledge, are subsequently used to refine and update the solution pool, thereby embodying the Deming Cycle's iterative improvement loop. In the *educating* stage, SciEducator generates multimodal e-learning materials tailored to the scientific content interpreted in the *understanding* stage. It first retrieves relevant experimental procedures, safety precautions, required equipment images, and shopping links, and then produces textual instructions, visual guides, audio narrations, and interactive e-booklets that present scientific knowledge in an engaging and accessible manner.

To evaluate the superiority of our system, we introduce SciVBench, a new benchmark containing 500 expert-validated and literature-grounded question–answer (QA) pairs across three scientific domains: physics experiments, chemistry experiments, and daily life phenomena. The QA pairs are further categorized into five types, *i.e.,* terminology, principle, prediction, reading, and design. We also propose a comprehensive set of metrics for comparative evaluation of both understanding and educating capabilities. Experimental results show that SciEducator consistently outperforms leading commercial MLLMs and state-of-the-art multi-agent systems in both aspects.

Our main contributions are summarized as follows:

- We propose **SciEducator**, the first multi-agent system for scientific video understanding and educating, which not only comprehends complex and fine-grained scientific videos but also generates multimodal educational e-booklets that integrate diverse information to deliver engaging guidance and stimulate scientific curiosity.
- Inspired by the **Deming Cycle**, we equip SciEducator with an iterative workflow optimization mechanism that progressively refines solutions through failure analysis and newly acquired knowledge, substantially enhancing its capability to handle meticulous scientific activities and offering a novel paradigm for agentic system design.
- We construct **SciVBench**, the first benchmark for scientific-phenomenon video analysis, featuring diverse question–answer pairs across physics, chemistry, and daily life phenomena, along with a comprehensive set of evaluation metrics.

2. Related Work

2.1. Multimodal Large Language Models

Recent years have witnessed substantial progress in vision-language models [4, 32, 33, 39, 48, 49, 56, 61, 62, 64]. VTimeLLM introduces a boundary-aware three-stage training strategy [36], TimeSuite employs temporal adaptive position encoding [58], and Slowfast-LLava designs a low-speed feature sampler to enhance temporal awareness in visual representations [54]. Closed-source models such as GPT-4o [42], Gemini [35], and Claude [1] have also been widely adopted as foundational MLLMs in downstream video tasks. However, these systems are critically limited by their inherent inability to effectively leverage external resources and integrate specialized tools. This fundamental shortfall severely curtails their functional capacity and practical applicability in various scenarios.

2.2. Multi-Agent Systems

With the application of multi-agent systems(MASs) across numerous domains, computer vision researchers have also begun exploring MASs for video understanding [34, 55]. VideoAgent [52] proposed a framework capable of interactive reasoning and planning through confidence evaluation, focusing particularly on long-form video comprehension. In [8], a memory-enhanced MAS for video understanding was introduced, emphasizing structured spatiotemporal information. VideoAgent2 [63] simulated the cognitive process of human video understanding by designing an uncertain chain-of-thought reasoning approach. However, they still face several challenges. Their performance is limited by the inherent weaknesses of LLM, like hallucinations and instability. For scientific video tasks and medical tasks [47] requiring external expertise and rigorous step-by-step reasoning, they often fail to produce viable initial plans and lack systematic mechanisms that enable them to self-evolve and optimize their workflows based on past results.

3. SciEducator

Our pipeline is illustrated in Fig. 2, where we introduce SciEducator, a multi-agent system for scientific video comprehension and education. We begin by giving the problem definition in Sec. 3.1, and introduce the specialized tools & agents integrated into the system in Sec. 3.2, followed by a description of SciEducator's video understanding workflow in Sec. 3.3, and the educational booklet generation process in Sec. 3.4. These designs enable the system to effectively analyze problems, provide feasible solutions, and ultimately produce accurate and robust responses, with an optional education e-booklet generation.

3.1. Problem Definition and Overview

The goal of our multi-agent system $\mathcal{S}$ is to return an accurate, self-consistent answer A given a user query Q and a scientific video V; what's more, it is able to generate an educational booklet after video understanding. The core workflow of our approach can be represented as follows:

$$A = \mathcal{S}(Q, V; P, E, \mathcal{T}), \tag{1}$$

where P and E represent the planner and evaluator, which act as the leading roles inside the Deming (PDSA) cycle. $\mathcal{T}$ denotes the dynamically configured set of tools and agents aiming at different stages inside our system. For example, in the Deming cycle, $\mathcal{T}$ includes $\mathcal{T}_{\text{Plan}}, \mathcal{T}_{\text{Do}}, \mathcal{T}_{\text{Study}}$; and in the educational booklet generation stage, $\mathcal{T}$ refers to $\mathcal{T}_{\text{Booklet}}$.

Speaking of the PDSA iteration, which is the core of our pipeline, it conducts the plan-do-study-act process iteratively to obtain the most refined and reasonable answer, as shown in Fig. 2. Specifically, "plan" refers to the process of retrieving the corresponding knowledge and building up a solution pool from our system's internal knowledge base. The "do" stage targets at assessing to find the most appropriate solution from the pool, and provides confidence to decide either to stop cycles and give a well-structured response to users, or step into the "study" stage. In "study", our system aims to find the reasons for being unable to give answers, and seeks useful new information. After updating the solution pool with the "act" stage, it will have better confidence to provide a good reply to the query. We will illustrate more details in Sec. 3.3.

Another key aspect of our system is the extension of scientific video understanding toward educational content generation, enabling the automatic creation of structured e-booklets that illustrate and guide the replication of scientific phenomena with high relevance, clarity, and appeal to young learners. Further details are provided in Sec. 3.4.

3.2. Tools & Agents Configuration

Our system integrates a total of sixteen specialized components, including ten agents and six tools, each tailored for specific subtasks with well-defined input–output interfaces. These components are organized into two categories: (i) **dynamically invocable tools/agents**, which are adaptively called based on the system's reasoning context; and (ii) **fixed-execution tools/agents**, which operate automatically at predetermined stages of the workflow.

The dynamically invocable components primarily support functions such as task planning, content acquisition, evaluation, and web or literature retrieval. The fixed-execution components focus on knowledge-base maintenance, multimodal synthesis, and e-booklet generation. Detailed descriptions of all tools and agents are provided in the Supplementary Material.

3.3. Deming Cycle within SciEducator

We instantiate SciEducator as a closed-loop Plan–Do–Study–Act (PDSA) controller tailored to scientific video understanding and its downstream educational content creation. The same loop is later reused for e-booklet generation (Sec. 3.4) with a slightly different toolset and, when applicable, without video input. Each of the four stages in the i-th cycle is described as follows.

Plan stage. Given a user query Q and video V sampled at one fps, we prompt a captioner to obtain a temporally grounded description V_{content}. A retrieval-augmented agent then extracts salient entities and keywords and retrieves domain knowledge K from an internal corpus. Noted that, this K is only retrieved in the first cycle for once, and will be updated just by the study stage. The planner composes a diverse pool of candidate plans by combining V_{content} and K, using few-shot exemplars to stabilize call signatures and

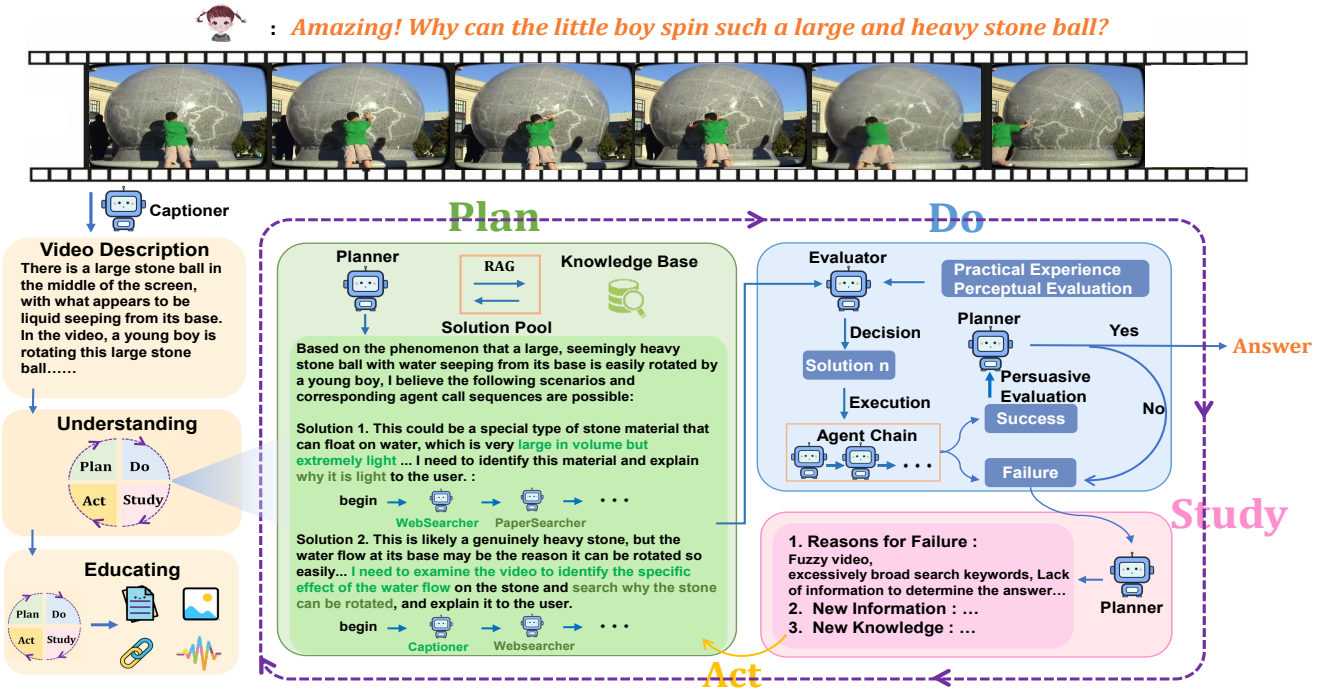


Figure 2. SciEducator architecture overview. We design a multi-agent system capable of implementing the PDSA cycle, which iteratively optimizes output responses through cyclic iterations.

parameter settings. This stage can be summarized as:

$$M_i, K = P\big(Q,\, V,\, \mathcal{T}_{\text{Plan}}\big), \tag{2}$$

where M_i is the candidate plan pool in the i-th cycle and $\mathcal{T}_{\text{Plan}}$ denotes the planning toolset.

Do stage. Given the solution pool M_i, the evaluator scores each plan with a composite of objective A_{obj} and LLM-based perceptual criteria A_{percep} and selects the best trade-off in time and token efficiency, success likelihood, feasibility, and overall performance. To estimate time and token costs, we build an empirical prior $\mathcal{E}$ by issuing 20 randomized probe calls per tool/agent, from which we obtain average latency, average token usage, and success probability. For relevance and feasibility, we weight a plan's keywords using inverse document frequency (IDF) [45], reflecting their ability to retrieve discriminative knowledge for the current video:

$$\text{IDF}(k) = \log\left(\frac{N}{\text{f}(k)+1}\right), \tag{3}$$

where N is the corpus size and $\text{f}(k)$ is the number of documents containing k; we use an internal corpus of 84 physics and chemistry documents. In parallel, a subjective judge compares plans on coverage, logical coherence, scientific soundness, and clarity. The selection of the best solution $s^\star$ is defined as:

$$s^\star = \arg\max_{s \in M_i} \big[\, A_{\text{obj}}(s; \mathcal{E}, \text{IDF}) + \lambda\, A_{\text{percep}}(s) \,\big], \tag{4}$$

which is then executed step-by-step to produce the stage output R_i. We summarize the do stage as

$$R_i = E\big(M_i,\, V,\, \mathcal{T}_{\text{Do}}\big), \tag{5}$$

where $E(\cdot)$ is the evaluator and $\mathcal{T}_{\text{Do}}$ refers to the do-stage toolset. After obtaining R_i, the planner estimates a confidence score using the query Q, the video context V, and the executed plan:

$$C_i = P(R_i,\, Q,\, V). \tag{6}$$

This score indicates whether the available evidence suffices to address the query and yield a convincing answer; if high, SciEducator synthesizes the information and produces a well-structured response.

Study and Act Stage. If confidence is low, the system enters the study phase: the planner diagnoses why the answer is insufficient (e.g., tool failures, overly broad and irrelevant retrieval, or insufficient detail in video captions) and aggregates any useful evidence discovered this round into the knowledge base. Formally,

$$F_i,\, K_{i+1} = P\big(R_i,\, K_i,\, Q,\, V,\, \mathcal{T}_{\text{Study}}\big), \tag{7}$$

where F_i is the failure analysis and $K_{i+1} = K_i \cup K_{\text{new}}$ is the updated knowledge. Guided by F_i, K_i, the act function Γ_{Act} replans the next pool—adjusting structure, granularity, or query specificity (e.g., apply video super-resolution for blurry frames, increase captioning fps for missed actions, and refine queries with more specific entities):

$$M_{i+1} = \Gamma_{\text{Act}}\big(F_i,\ K_{i+1},\ M_i\big). \tag{8}$$

The system executes the next iteration with the updated solution pool M_{i+1} and repeats the confidence check, iterating until $C_i \geq \tau$ or a maximum number of cycles we choose.

3.4. Educational E-booklet Generation

After video understanding, SciEducator identifies the specific scientific phenomenon in the video and its underlying principles. When required, it triggers a multimodal retrieval and generation pipeline to produce a child-friendly e-booklet that includes: **(a)** experimental guidance; **(b)** required equipment with purchase links; **(c)** step-by-step procedures with instructional diagrams; **(d)** safety precautions for both equipment and operations with accompanying audio prompts; and **(e)** a concise summary of the experiment and its principles. The booklet adopts a progressive, logically organized layout to support rapid comprehension and faithful replication by young readers.

To assemble content, SciEducator uses an Entity Recognition agent to extract key entities, then the Procedure Search and Safety Alert agents to retrieve replicable steps and precautions. We target four properties: **(i)** relevance (avoiding content unrelated to the experiment), **(ii)** instructional quality (complete, detailed, and safety-aware), **(iii)** attractiveness (engaging, visually impressive replication), and **(iv)** educational value (steps that expose the governing principles). The process follows the same PDSA loop as Sec. 3.3, but here the input is only the query, confidence is assessed w.r.t. these four properties, while the available tools are different.

After obtaining the procedures, an Equipment Search tool returns images and purchase links for the relevant equipment; an Illustration Generation tool produces stepwise guidance images; and a Speech Generation tool synthesizes audio instructions. Finally, an E-booklet Generation agent integrates text, images, audio, hyperlinks, and layout into a well-structured booklet in the style of children's science literature. Compared with a single-LLM baseline, SciEducator provides richer modalities with higher relevance and instructional quality, while improving engagement through integrated text–image–audio design.

4. Experiments

4.1. SciVBench Dataset

To construct a specialized, accurate, and reasoning-intensive collection of scientific videos with corresponding QA pairs, we systematically gathered rich scientific video resources from major video platforms and science education websites, including 54 physics experiment videos, 54 chemistry experiment videos, and 103 daily life phenomenon videos as shown in Fig. 3. Domain experts in physics and chemistry systematically reviewed the subtitles, explanatory narrations, knowledge manuals or guides, as well as relevant reference materials and academic papers for each video. Two domain experts independently authored and validated each QA pair, and disagreements were adjudicated by a third expert. Through rigorous cross-validation with additional sources, we developed 500 scientifically-grounded question-answer pairs based on the video content: 160, 148, and 192 pairs for physics, chemistry, and daily life phenomenon videos.

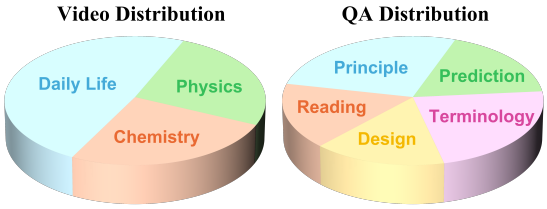


Figure 3. Proportional distribution of video and QA categories. SciVBench comprises three types of scientific videos and five categories of questions, enabling a comprehensive evaluation of a model's ability to acquire diverse domain knowledge and tackle various complex scientific problems.

Only the visual content of the videos was used as input, with all subtitles and audio narrations removed. Each QA pair was designed such that correct answers could only be derived by comprehending the video content, not from the question text alone. All answers maintain logical integrity, forming coherent reasoning chains from fundamental knowledge to final conclusions, while avoiding any semantic deviation from reference materials and papers. These highly specialized, accurate, and reasoning-intensive QA pairs establish a solid foundation for evaluating the performance of existing MLLMs and multi-agent systems (MASs) in scientific video comprehension and validating the effectiveness of our SciEducator.

4.2. Evaluation Metrics

Evaluation Metrics for Understanding. We employ two metrics to evaluate the performance in scientific video understanding, using qwen3-max as the unified evaluator for all model-generated answers. Specifically, these metrics are:

- **Relevance (Rel)**: We first provide the reference answer and extensive scientific background for the current question to Qwen3-Max [2], instructing it to assess how relevant the model-generated answer is to the query—focusing on whether the response aligns with the scientific subdomain involved, regardless of correctness.

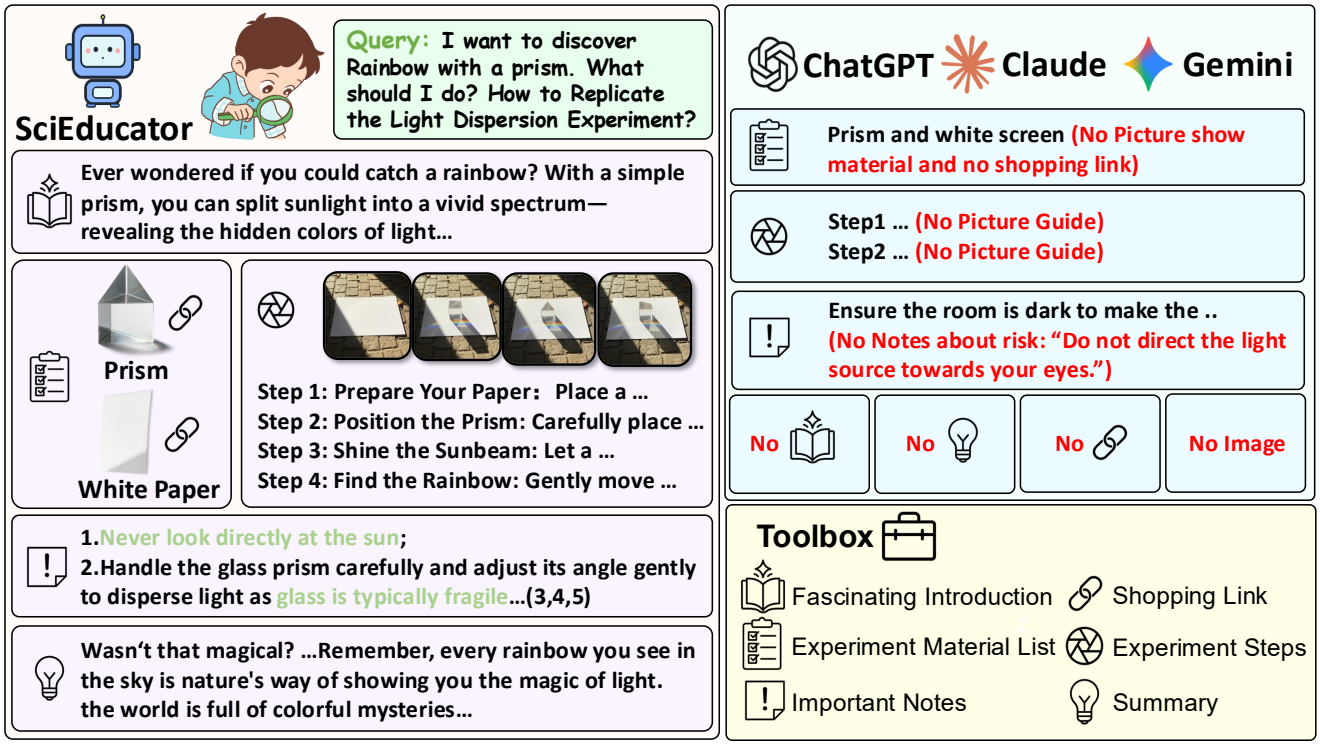


Figure 4. Qualitative comparison between SciEducator and MLLMs in Education E-booklet Generation. The left is our generated e-booklet with comprehensive contents and a well-organized structure, while other popular MLLMs (see the right) all fail in such generation. These examples demonstrate SciEducator's ability to generate more comprehensive, better-structured, and more attractive Education E-booklet.

The goal is to evaluate whether the answer provides misleading or irrelevant information. The scoring uses three discrete values: 0 for irrelevant responses, 0.5 for partially relevant, and 1 for fully relevant.

- **Accuracy (Acc)**: We then use Qwen3-Max [2] to analyze the semantic similarity between the generated answer and the reference answer, scoring the correctness of the model's response. Similarly, this metric employs three discrete scores: 0 for completely incorrect answers, 0.5 for partially correct, and 1 for fully correct.

Detailed scoring strategies and prompts can be found in the supplementary materials. The relevance and accuracy metrics for each model are calculated as the average scores across all questions and are presented in the results table as percentages, with the percentage symbols omitted.

Evaluation Metrics for Education. To assess the educational performance of SciEducator and various MLLMs, we decoupled this part from the understanding benchmark. All models were provided with the scientific terminology of the current experiment in a video and asked to generate procedures and precautions for replicating it. Although SciEducator can generate more modalities than MLLMs, we focused specifically on comparing their shared textual modality for quantitative comparison.

We employed four metrics and uniformly used Qwen-VL-Plus [2] to evaluate all model responses in a comparative setting. Qwen-VL-Plus [2] was 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.

Detailed scoring strategies and prompts can be found in the supplementary materials. We evaluate 40 videos (Education Subset). For each video, all model responses were collectively anonymized and input into the VLM for evaluation. The best response for each metric was selected, and

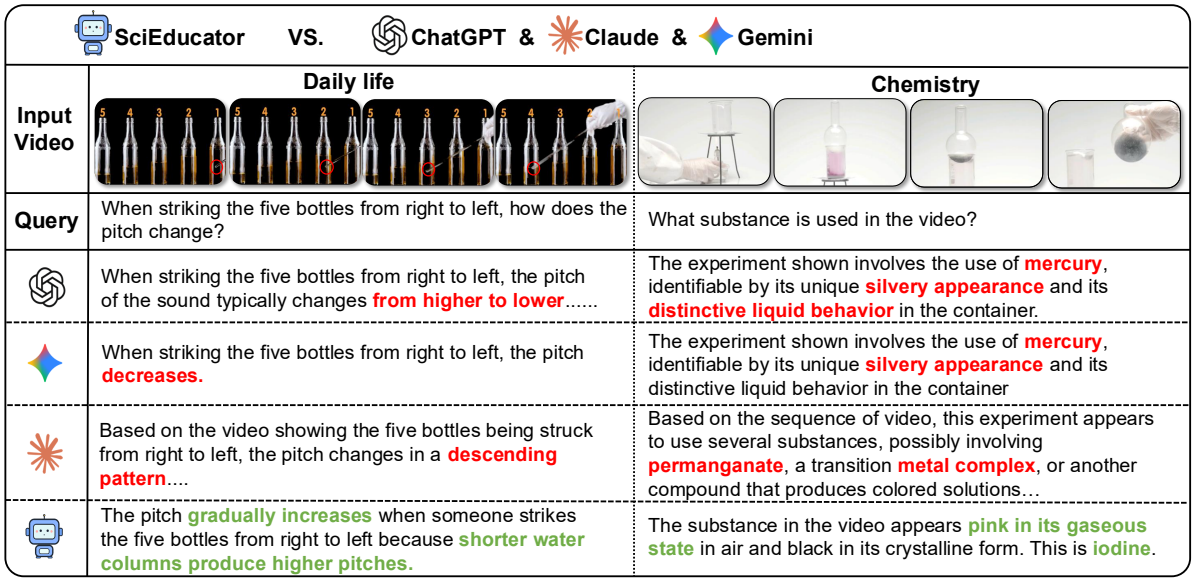


Figure 5. Qualitative comparison between SciEducator and MLLMs. These examples demonstrate SciEducator's ability to generate more comprehensive, better-structured, and more logically coherent answers than the other MLLMs.

Table 1. Quantitative Comparison between SciEducator and popular MLLMs and MASs for scientific video understanding on SciVBench. Compared with these SOTA models, our model exhibits higher relevance and accuracy.

Model	Physics		Chemistry		Daily Life	
	Rel	Acc	Rel	Acc	Rel	Acc
Commercial APIs						
GPT-4o [42]	47.50	34.69	39.86	31.42	30.73	27.86
Gemini 2.0 flash [35]	52.81	38.75	46.96	36.15	34.64	31.25
Claude 3.7 Sonnet [1]	44.06	31.88	40.20	31.76	31.77	28.65
Multi-Agent Systems						
VideoAgent[52]	49.06	36.56	45.61	34.80	30.47	27.34
videoagent[8]	46.25	35.31	46.62	37.16	31.51	28.13
SciEducator (Ours)	**81.88**	**65.31**	**73.97**	**64.86**	**64.58**	**62.24**

final win rates were calculated and presented in tables.

4.3. Experimental Results

Quantitative Analysis. To verify SciEducator's effectiveness, we evaluate on SciVBench against three closed-source MLLMs (Claude 3.7 Sonnet [1], GPT-4o [42], Gemini 2.0 Flash [35]) and two state-of-the-art multi-agent systems (MASs) (VideoAgent [52] and videoagent [8]), and report results in Tab. 1 for scientific video understanding and Tab. 2 for educational content generation.

For scientific video understanding, Tab. 1 shows that SciEducator consistently surpasses all baselines on the relevance and accuracy metrics across the physics, chemistry, and daily-life tracks, indicating tighter adherence to the visual evidence and underlying scientific principles than general-purpose MLLMs and prior MASs.

Tab. 2 reports results for the scientific educational content generation task, comparing SciEducator with the closed-source MLLMs (MASs baselines do not support this task). SciEducator demonstrates a clear advantage in creating high-quality instructional materials.

Table 2. Quantitative Comparison between SciEducator and popular MLLMs for scientific education on SciVBench Education Subset. Each metric is reported by win rate (%).

Model	Education Subset			
	Relevance	IQ	Attractiveness	EV
Gemini 2.0 flash [35]	10.00	2.50	0.00	5.00
GPT-4o [42]	7.50	5.00	2.50	7.50
Claude 3.7 Sonnet [1]	5.00	5.00	0.00	5.00
SciEducator(Ours)	**77.50**	**87.50**	**97.50**	**82.50**

Qualitative Analysis. As illustrated in Fig. 5, for specific user queries, SciEducator provides responses that are more comprehensive, detailed, and clearly explained compared to MLLMs such as GPT-4o [42]. This advantage stems from two core capabilities of SciEducator: first, the integration of multiple powerful search tools significantly expands the system's knowledge capacity; second, its robust planning and replanning capabilities enable it to better analyze complex video content and acquired information, resulting in more organized and higher-quality answers.

As shown in Fig. 1, SciEducator can generate e-learning materials of four modalities: text, images, hyperlinks, and audio, featuring rich colors and aesthetically pleasing layouts. The text is structured with sequential sections including a title, introduction, experimental procedures and precautions, and a summary. As illustrated in Fig. 4, this structural arrangement effectively guides readers through understanding the principles and completing the experiment. The

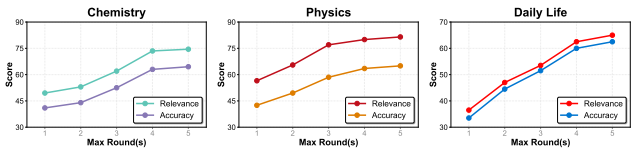


Figure 6. Ablations on different iteration rounds for scientific video understanding. Performance rises with larger Max Rounds for three categories, showing the benefit of iterative PDSA cycles.

Table 3. Ablations on different maximum iteration rounds for education. Each metric is reported by win rate (%).

Model	Max Round(s)	Education Subset			
		Relevance	IQ	Attractiveness	EV
SciEducator	1	2.50	0	2.50	15.00
SciEducator	3	7.50	7.50	32.50	35.00
SciEducator	5	**90.00**	**92.50**	**65.00**	**50.00**

visual elements include images of required materials and model-generated diagrams illustrating experimental steps, providing clearer guidance than text alone and highlighting our significant advantage over conventional LLMs.

4.4. Ablation Studies

Effectiveness of PDSA Cycles. To validate the progressive improvement of SciEducator in response relevance and accuracy through PDSA cycles, we configured the maximum number of cycles from 1 to 5. After reaching the cycle limit, the system was compelled to generate the most likely correct answer based on accumulated knowledge. Results in Fig. 6 show that performance improves significantly with more cycles, demonstrating SciEducator's capacity for self-evolution and its ability to iteratively deepen understanding, acquire relevant knowledge, and solve complex problems.

To validate the progressive improvement of SciEducator in generating educational text via PDSA cycles, we compare three variants with maximum cycle counts of 1, 3, and 5 on the Education Subset of SciVBench, with results also presented as win rates across various evaluation aspects and summarized in Tab. 3. Results demonstrate that as the PDSA cycles increase, the model's performance improves, yielding instructional text with stronger relevance and more educational value, which in turn leads to the generation of higher-quality guidance images.

Effectiveness of Evaluator Agent. To validate that each evaluation element in the Evaluator Agent effectively reduces resource consumption and execution rounds, we compared the complete EA with ablated versions lacking these elements on 500 QA pairs. As shown in Tab. 4, metrics include average time, token consumption, number of execution rounds, and final accuracy, with consumption values shown as ratios. Note that the total consumption of the complete EA is normalized to 1.00. The maximum number of execution rounds was set to 5. Experimental results demonstrate that each evaluation element in the Evaluator Agent effectively increases the probability of selecting superior solutions, reduces resource consumption and identifies more feasible options, and consequently generates responses with higher accuracy within limited iteration cycles.

Table 4. Ablations on Evaluator Agent (EA). Note that $\mathcal{E}$ and A_{percep} denote empirical knowledge and perceptual evaluation metrics. Time and token consumption are presented as ratios, with the total value of the complete EA version normalized to 1.00.

Agent	Time↓	Token↓	Average Rounds↓	Acc ↑
EA w/o $\mathcal{E}$	1.20	1.18	4.09	57.50
EA w/o IDF	1.08	1.06	3.99	59.90
EA w/o A_{percep}	1.14	1.13	4.17	54.50
EA	**1.00**	**1.00**	**3.79**	**64.00**

Table 5. Ablations of on the effectiveness of different elements in the Study Stage. K_{new} represents the new knowledge acquired during execution, and F denotes failure analysis.

Model	Physics		Chemistry		Daily Life	
	Rel	Acc	Rel	Acc	Rel	Acc
SciEducator w/o K_{new} & F	59.69	45.94	53.04	45.27	35.94	32.55
SciEducator w/o K_{new}	65.94	50.94	61.82	54.05	38.28	34.64
SciEducator w/o F	71.56	55.63	66.55	57.09	48.95	45.83
SciEducator	**81.88**	**65.31**	**73.97**	**64.86**	**64.58**	**62.24**

Effectiveness of K_{new} & F. To validate the effectiveness of new knowledge K_{new} and failure analysis F of the Study Stage, which updates the solution pool and improves the quality of candidate solutions, we conducted ablations by removing these elements, as presented in Tab. 5. In the variant where both K_{new} and F are removed, no feedback is available, and updates only discard executed solutions. The experimental results indicate that K_{new} and F are critical for optimizing existing solutions and generating new ones. Within the PDSA process, these elements substantially improve the quality of the new solution pool, thereby enhancing both the relevance and accuracy of the results.

5. Conclusion

In this work, we present SciEducator, the first iterative self-evolving multi-agent system for scientific video comprehension and education. By reformulating the Deming Cycle's *Plan–Do–Study–Act* philosophy into a self-evolving reasoning and feedback mechanism, the system progressively enhances its interpretation of complex scientific activities to deliver accurate and reliable results. To ensure comprehensive performance assessment, we construct the SciVBench benchmark. Extensive experiments show that SciEducator significantly outperforms current popular models, demonstrating remarkable superiority and application potential, and is expected to offer meaningful inspiration for future multi-agent systems and broader domains.

6. Acknowledgments

This work is supported by the National Science and Technology Major Project (Grant No. 2023ZD0120803), the National Natural Science Foundation of China (Grant No. 62502317), and the Fundamental Research Funds for the Central Universities (Grant No. 226-2025-00080).

References

[1] Anthropic. Claude 3.7 sonnet and claude code, 2025. Official model announcement and capabilities overview. 2, 7

[2] 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. 5, 6

[3] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1

[4] Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24507–24517, 2025. 2

[5] Runpeng Dai, Run Yang, Fan Zhou, and Hongtu Zhu. Breach in the shield: Unveiling the vulnerabilities of large language models. *arXiv preprint arXiv:2504.03714*, 2025.

[6] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[7] Francesco Cardinale et al. Isr. https://github.com/idealo/image-super-resolution, 2018.

[8] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer, 2024. 2, 3, 7

[9] Wanlong Fang, Tianle Zhang, and Alvin Chan. To align or not to align: Strategic multimodal representation alignment for optimal performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[10] Xiang Fang and Wanlong Fang. Disentangling adversarial prompts: A semantic-graph defense for robust llm security. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[11] Xiang Fang and Yuchong Hu. Double self-weighted multi-view clustering via adaptive view fusion. *arXiv preprint arXiv:2011.10396*, 2020.

[12] Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Oliver Wu. V3h: View variation and view heredity for incomplete multiview clustering. *IEEE Transactions on Artificial Intelligence*, 1(3):233–247, 2020.

[13] Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Wu. Animc: A soft approach for autoweighted noisy and incomplete multiview clustering. *IEEE Transactions on Artificial Intelligence*, 3(2):192–206, 2021.

[14] Xiang Fang, Yuchong Hu, Pan Zhou, and Dapeng Oliver Wu. Unbalanced incomplete multi-view clustering via the scheme of view evolution: Weak views are meat; strong views do eat. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(4):913–927, 2021.

[15] Xiang Fang, Daizong Liu, Pan Zhou, and Yuchong Hu. Multi-modal cross-domain alignment network for video moment retrieval. *IEEE Transactions on Multimedia*, 25:7517–7532, 2022.

[16] Xiang Fang, Daizong Liu, Wanlong Fang, Pan Zhou, Yu Cheng, Keke Tang, and Kai Zou. Annotations are not all you need: A cross-modal knowledge transfer network for unsupervised temporal sentence grounding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8721–8733, 2023.

[17] Xiang Fang, Daizong Liu, Pan Zhou, and Guoshun Nan. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2448–2460, 2023.

[18] Xiang Fang, Daizong Liu, Pan Zhou, Zichuan Xu, and Ruixuan Li. Hierarchical local-global transformer for temporal sentence grounding. *IEEE Transactions on Multimedia*, 2023.

[19] Xiang Fang, Arvind Easwaran, and Blaise Genest. Uncertainty-guided appearance-motion association network for out-of-distribution action detection. In *IEEE International Conference on Multimedia Information Processing and Retrieval*, 2024.

[20] Xiang Fang, Wanlong Fang, Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Renfu Li, Zichuan Xu, Lixing Chen, Panpan Zheng, et al. Not all inputs are valid: Towards open-set video moment retrieval using language. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 28–37, 2024.

[21] Xiang Fang, Daizong Liu, Wanlong Fang, Pan Zhou, Zichuan Xu, Wenzheng Xu, Junyang Chen, and Renfu Li. Fewer steps, better performance: Efficient cross-modal clip trimming for video moment retrieval using language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1735–1743, 2024.

[22] Xiang Fang, Zeyu Xiong, Wanlong Fang, Xiaoye Qu, Chen Chen, Jianfeng Dong, Keke Tang, Pan Zhou, Yu Cheng, and Daizong Liu. Rethinking weakly-supervised video temporal grounding from a game perspective. In *European Conference on Computer Vision*. Springer, 2024.

[23] Xiang Fang, Arvind Easwaran, and Blaise Genest. Adaptive multi-prompt contrastive network for few-shot out-of-distribution detection. In *International Conference on Machine Learning*, 2025.

[24] Xiang Fang, Arvind Easwaran, Blaise Genest, and Ponnuthurai Nagaratnam Suganthan. Adaptive hierarchical graph cut for multi-granularity out-of-distribution detection. *IEEE Transactions on Artificial Intelligence*, 2025.

[25] Xiang Fang, Arvind Easwaran, Blaise Genest, and Ponnuthurai Nagaratnam Suganthan. Your data is not perfect: Towards cross-domain out-of-distribution detection in class-imbalanced data. *Expert Systems with Applications*, 2025.

[26] Xiang Fang, Wanlong Fang, Wei Ji, and Tat-Seng Chua. Turing patterns for multimedia: Reaction-diffusion multi-modal fusion for language-guided video moment retrieval. In *ACM International Conference on Multimedia*, 2025.

[27] Xiang Fang, Wanlong Fang, and Changshuo Wang. Hierarchical semantic-augmented navigation: Optimal transport and graph-driven reasoning for vision-language navigation. In *Advances in Neural Information Processing Systems*, 2025.

[28] Xiang Fang, Wanlong Fang, Changshuo Wang, Daizong Liu, Keke Tang, Jianfeng Dong, Pan Zhou, and Beibei Li. Multi-pair temporal sentence grounding via multi-thread knowledge transfer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[29] Xiang Fang, Wanlong Fang, Changshuo Wang, Daizong Liu, Keke Tang, Jianfeng Dong, Pan Zhou, and Beibei Li. Multi-pair temporal sentence grounding via multi-thread knowledge transfer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2915–2923, 2025.

[30] Xiang Fang, Wanlong Fang, and Changshuo Wang. Cogniverse: Revolutionizing multi-modal retrieval-augmented generation with cognitive reflection and geometric reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[31] Xiang Fang, Wanlong Fang, and Changshuo Wang. Unveiling the fragility of vision-language models: Multi-modal adversarial synergy via texture-constrained perturbations and cross-modal optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[32] Xiang Fang, Wanlong Fang, Changshuo Wang, Xiaoye Qu, and Daizong Liu. Rethinking video-language model from the language input perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. 2

[33] Xiang Fang, Wanlong Fang, Changshuo Wang, Keke Tang, Daizong Liu, Siyi Wang, and Wei Ji. Towards unified vision-language models with incomplete multi-modal inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. 2

[34] Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640*, 2023. 3

[35] Google DeepMind. Introducing gemini 2.0: our new ai model for the agentic era, 2024. Official announcement; includes Gemini 2.0 Flash details. 2, 7

[36] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024. 2

[37] Kunal Jha, Wilka Carvalho, Yancheng Liang, Simon S Du, Max Kleiman-Weiner, and Natasha Jaques. Cross-environment cooperation enables zero-shot multi-agent coordination. *arXiv preprint arXiv:2504.12714*, 2025. 2

[38] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

[39] Guanghao Li, Wenhao Jiang, Mingfeng Chen, Yan Li, Hao Yu, Shuting Dong, Tao Ren, Ming Tang, and Chun Yuan. Scout: Teaching pre-trained language models to enhance reasoning via flow chain-of-thought. *arXiv preprint arXiv:2505.24181*, 2025. 2

[40] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[41] Ronald Moen, Clifford Norman, et al. Evolution of the pdca cycle, 2006. 2

[42] OpenAI. Gpt-4o system card. https://arxiv.org/abs/2410.21276, 2024. System Card describing capabilities, limitations, and safety evaluations of GPT-4o. 2, 7

[43] Wei Pang, Kevin Qinghong Lin, Xiangru Jian, Xi He, and Philip Torr. Paper2poster: Towards multimodal poster automation from scientific papers. *arXiv preprint arXiv:2505.21497*, 2025. 2

[44] Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 6(3):1, 2023. 2

[45] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, pages 29–48. New Jersey, USA, 2003. 4

[46] Jiayuan Rao, Zifeng Li, Haoning Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Multi-agent system for comprehensive soccer understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3654–3663, 2025.

[47] Yufei Shi, Beijia Lu, Jia-Wei Liu, Ming Li, Si Yong Yeo, and Mike Zheng Shou. Colonnerf: High-fidelity neural reconstruction of long colonoscopy. *Neurocomputing*, page 131445, 2025. 3

[48] Yufei Shi, Weilong Yan, Gang Xu, Yumeng Li, Yucheng Chen, Zhenxi Li, Fei Yu, Ming Li, and Si Yong Yeo. Pvchat: Personalized video chat with one-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23321–23331, 2025. 2

[49] Xiu Su, Yuemin Wu, Zhongze Wu, Yitian Long, Yichao Cao, Yue Liao, Xi Lin, Jun Long, Shuo Jiang, Shan You, et al. A lightweight physics-aware framework for multi-scale marine heatwaves forecasting. *npj Climate and Atmospheric Science*, 2026. 2

[50] Michael J Taylor, Chris McNicholas, Chris Nicolay, Ara Darzi, Derek Bell, and Julie E Reed. Systematic review of the application of the plan–do–study–act method to improve quality in healthcare. *BMJ quality & safety*, 23(4):290–298, 2014. 2

[51] Oguzhan Topsakal and Tahir Cetin Akinci. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International conference on applied engineering and natural sciences*, pages 1050–1056, 2023.

[52] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024. 2, 3, 7

[53] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 1

[54] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024. 1, 2

[55] Zongxin Yang, Guikun Chen, Xiaodi Li, Wenguan Wang, and Yi Yang. Doraemongpt: Toward understanding dynamic scenes with large language models (exemplified as a video agent). *arXiv preprint arXiv:2401.08392*, 2024. 3

[56] Zhenhao Yang, Xin Liu, Deqiang Ouyang, Guiduo Duan, Dongyang Zhang, Tao He, and Yuan-Fang Li. Towards open-vocabulary HOI detection with calibrated vision-language models and locality-aware queries. In *ACM International Conference on Multimedia (ACM MM)*, pages 1495–1504, 2024. 2

[57] Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*, 2023. 2

[58] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*, 2024. 2

[59] Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. Multi-agent architecture search via agentic supernet. *arXiv preprint arXiv:2502.04180*, 2025.

[60] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023. 2

[61] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and A benchmark. In *CVPR*, pages 20228–20237, 2022. 2

[62] Jinkai Zheng, Xinchen Liu, Shuai Wang, Lihao Wang, Chenggang Yan, and Wu Liu. Parsing is all you need for accurate gait recognition in the wild. In *ACM MM*, pages 116–124, 2023. 2

[63] Zhuo Zhi, Qiangqiang Wu, Wenbo Li, Yinchuan Li, Kun Shao, Kaiwen Zhou, et al. Videoagent2: Enhancing the llm-based agent system for long-form video understanding by uncertainty-aware cot. *arXiv preprint arXiv:2504.04471*, 2025. 2, 3

[64] Sifan Zhou, Yichao Cao, Jiahao Nie, Yuqian Fu, Ziyu Zhao, Xiaobo Lu, and Shuo Wang. Comptrack: Information bottleneck-guided low-rank dynamic token compression for point cloud tracking. *arXiv preprint arXiv:2511.15580*, 2025. 2