上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

ExpVid: A Benchmark for Experiment Video Understanding & Reasoning

Outline

01 Background

Q: Why do we need a benchmark for "experimental video understanding"?

A: Most multimodal large models are trained on general video datasets and focus on surface descriptions.

They lack precision in recognizing materials, concentrations, tools, and fine-grained operations, which limits their reproducibility and safety, preventing them from acting as reliable lab assistants.

Existing benchmarks fail to evaluate MLLMs' ability to understand experimental videos.

- They focus on generic actions or medical scenes, lacking systematic evaluation of real laboratory experiments.
- Experimental videos pose unique challenges—micropipetting, small occluded tools, fine-grained material states, and long-term temporal dependencies.
- Thus, an experiment-oriented benchmark is needed to provide quantitative metrics for scientific-assistant applications.

🧠 Caption Generated by MLLM (Multimodal Large Language Model)

Observed experiment scene:

A person wearing purple gloves uses a small plastic tube to transfer white powder from one container labeled "CHEMICAL."

This process appears to be part of an optimization experiment involving different solution concentrations.

🎬 Transcript Caption (Original Video Subtitle)

125 mg of ethylene oxide–propylene oxide copolymer Poloxamer 188 was dissolved in water to prepare a 1% solution.

🔍 Similarity Analysis

Comparison Dimension	Analysis
Operation / Action	Both captions describe **adding or transferring solid substances** (white powder) into a container — operation consistent. ✅
Experimental Purpose / Context	MLLM infers an **"optimization experiment"**, while the transcript explicitly states **preparing a 1% surfactant solution** — semantically related. ✅
Specific Substance	MLLM fails to recognize the substance name; only mentions "white powder." Transcript specifies **125 mg Poloxamer 188**. ❌
Container Information	MLLM describes a container **labeled "CHEMICAL"**, referring to general labeling context — relevant but not identical. 🟣
Handling Tools	MLLM mentions a **"small plastic tube"** as the transfer tool — consistent. ✅
Linguistic Style	MLLM generates a **descriptive narrative**, emphasizing scene clarity and intent — reasonable. 🟣
Inference / Assumption	MLLM uses speculative phrases like *"appears to..."*, **adding interpretive bias** beyond observation. ⚠️

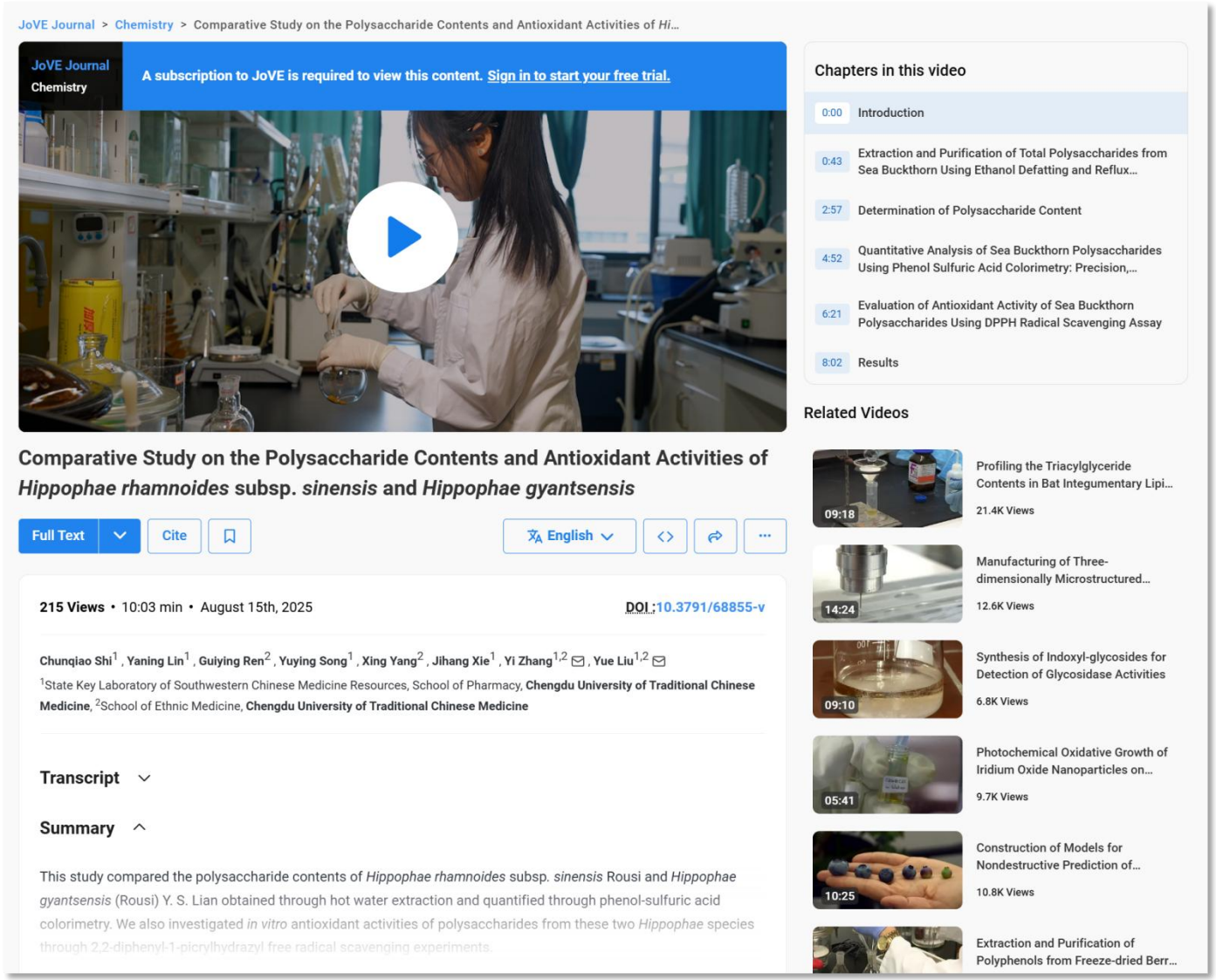
02 Data Source

Journal of Video Experiment

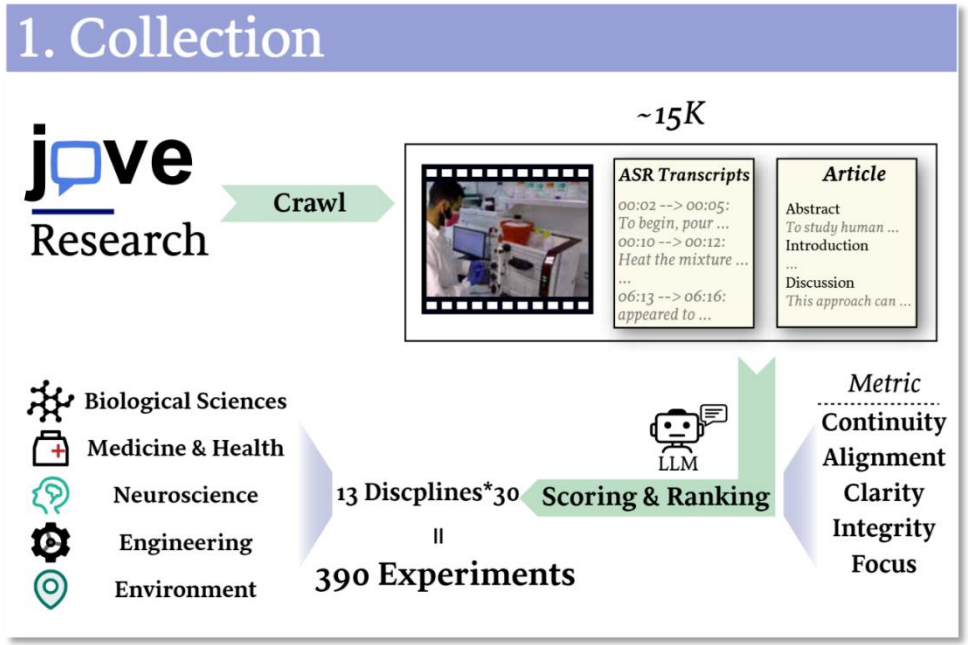
JoVE is a peer-reviewed academic journal focused on visualized experiments.

It presents scientific procedures through high-quality videos, enhancing reproducibility and educational value.

- **Peer-Reviewed Reliability**: Each video corresponds to a reviewed article ensuring trustworthy content.
- **Fine-Grained & Long-Term Processes**: Videos capture the full workflow, ideal for assessing micro-operations and temporal reasoning.
- **Broad Disciplinary Coverage**: Spans life sciences, chemistry, and engineering for diverse benchmark data.
- **Multimodal Completeness**: Each video includes narration, subtitles, and text, offering well-aligned multimodal annotations.

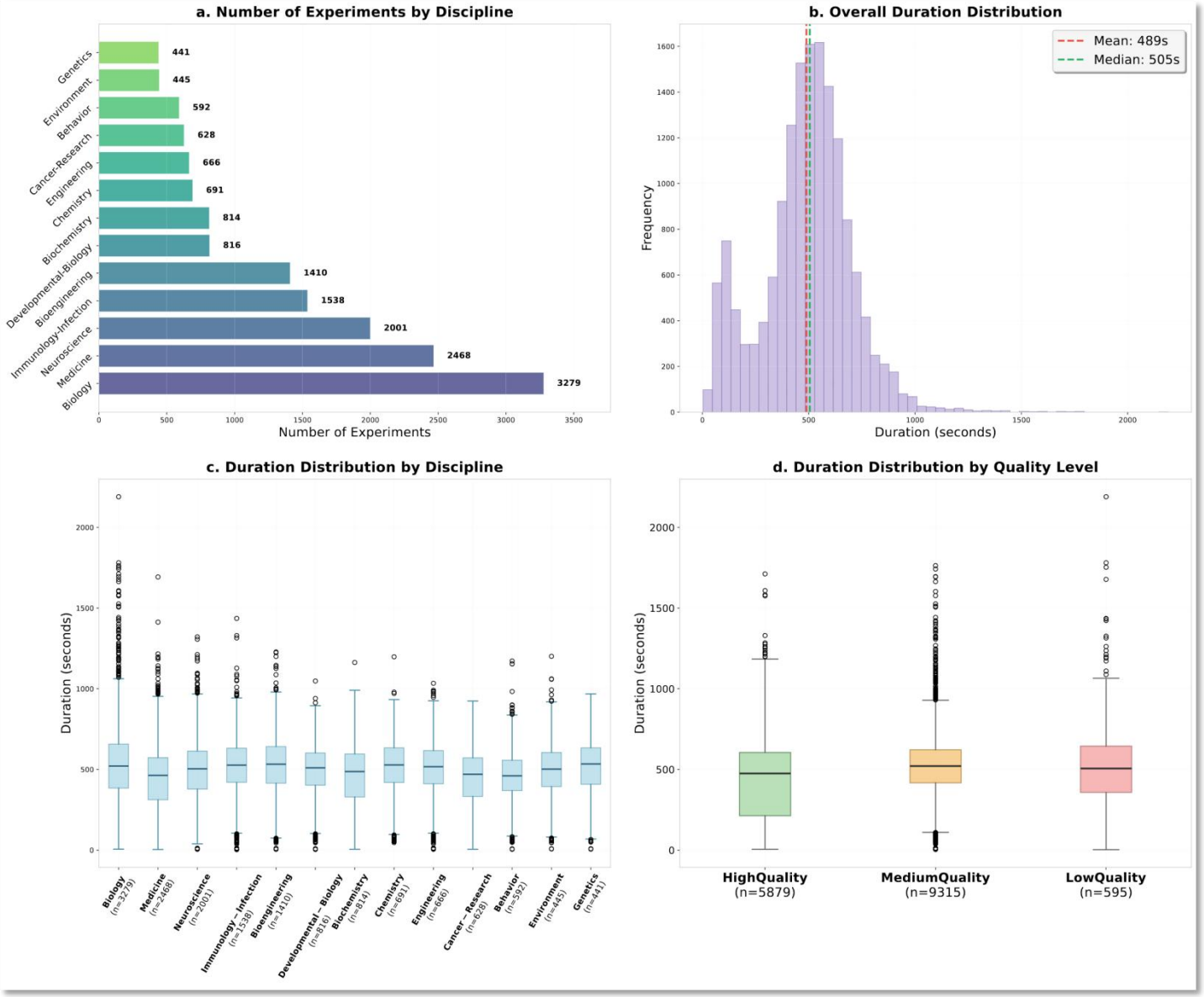


02 Data Source



We crawled about **15K** experiments from JoVE Research section (including videos, ASR transcripts, and paired papers).

Based on LLM scoring of experimental descriptions and manual filtering, we retained **390** experiments across 13 disciplines (30 per discipline).



03 Task Design

Three-Level Evaluation

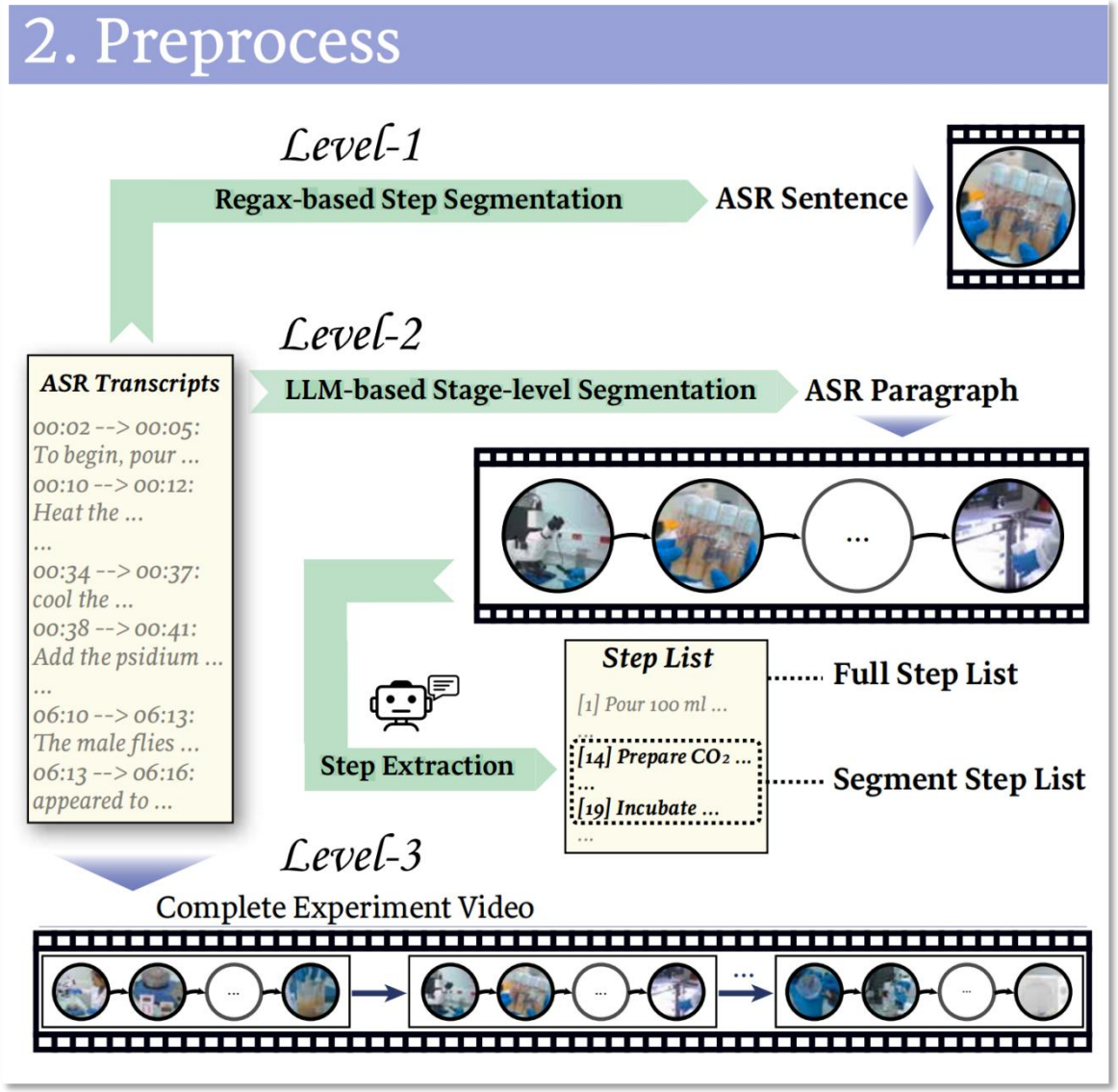
To systematically assess model capability, experimental videos are divided into three granular levels:

- Level-1 (Short clips): Fine-grained perception
- Level-2 (Medium segments): Step-level understanding
- Level-3 (Full videos): Scientific reasoning

These levels cover the full spectrum—from perceiving experimental elements, to understanding procedural logic, to high-level reasoning.

Preprocessing and Segmentation

- Level-1: Sentence/step-level segmentation aligned with subtitles, yielding single-step video clips.
- Level-2: LLM-based stage segmentation and step extraction aligned with subtitles, producing step lists for each clip.
- Level-3: Retains complete experiments to evaluate cross-stage temporal dependencies and overall reasoning ability.



03 Task Design

Level-1: Fine-grained Perception

I. Material

Question
What material is presented?

A. Drosophila pupae
B. Zebrafish embryos
C. Drosophila larvae
D. Tribolium pupae

II. Tool

Question
What tool is being used?

A. Hamilton syringe
B. Standard syringe
C. Micropipette
D. Gas-tight syringe

III. Quantity

Question
What's ethanol's conc.?

A. 70%
B. 50%
C. 80%
D. 90%

IV. Operation

Question
What happend to the tape?

A. Remove tape from ...
B. Attach clips to the ...
C. Secure both Petri ...
D. Apply tape only to ...

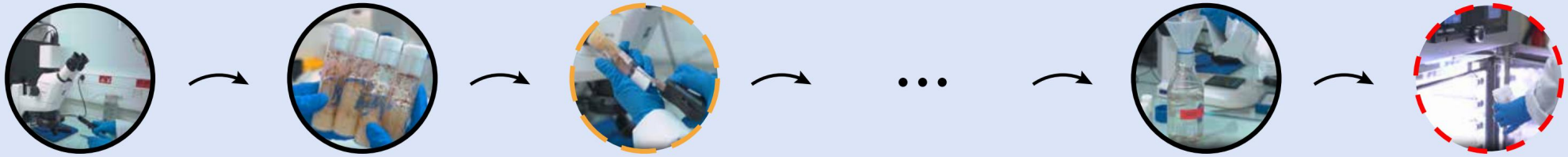
Objective: to evaluate whether MLLMs can visually ground essential elements in short clips of individual experimental steps through four Multi-Choice Question (MCQ) tasks:

- **Material Recognition**: Distinguish the target experimental material and distinguish it from other plausible substances commonly encountered in laboratory settings.
- **Tool Recognition**: Identify the appeared tools from the scene and reject visually or functionally similar distractors.
- **Quantity Recognition**: Choose the correct numerical attribute (e.g., dosage, temperature) by visually interpreting scales, amounts, or counts.
- **Operation Recognition**: Recognize the specific action being performed in the video and differentiate it from confusable but incorrect operations in the similar setup (e.g., Insert → Attach).

03 Task Design

Level-2: Procedural Understanding

Fly vial preparation, adult removal, and incubation



Step List

[1] Pour 100 ml ...
...
[14] Prepare CO_2 ...
[15] Select vials ...
...
[57] Analyze ...

V. Step Ordering

Question
What is the correct sequence?

A. Drop → Incube → Prepare → Select → ...
B. Prepare → Select → ... → Drop → Incube
C. Select → ...→ Prepare → Drop → Incube
D. Prepare → Drop → Incube → ...→ Select

VI. Completeness Verification

Question
Given the {segment step list}, which step was not performed in the video?
: Clipped Step

Correct Answer
Invert vial and insert needle.

VII. Sequence Generation

Question
Based on the {full step list}, determine the steps performed in the video.

Correct Answer
1. Prepare, 2. Select, 3. Inject, ...

VIII. Step Prediction

Question
Based on the {full step list}, infer the step about to take place in the video.
: Clipped Step

Correct Answer
Incubate vial at 25 °C, 60% RH.

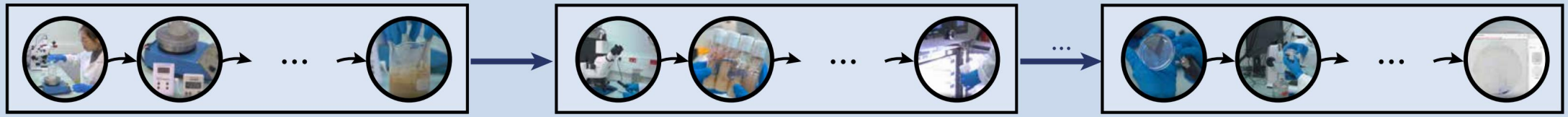
Objective: to valuate models on their reasoning about logical and temporal order across multiple steps within stage-level clips, including:

- **Step Ordering**: Select the correct step execution order when the original sequence is perturbed into plausible but incorrect arrangements.
- **Sequence Generation**: Given the candidates, find out the ordered steps that appear in the clip.
- **Completeness Verification**: Given the candidates, detect the missing step in the clip.
- **Step Prediction**: Given the first n − 1 steps of an experiment stage, predict the next step n.

03 **Task Design**

Level-3: Scientific Reasoning

Fecal Deposit Analysis of Drosophila for the Assessment of Antidiarrheal Drugs and Plant Extracts



Preparation and storage of food mixture. | Fly vial preparation and incubation. | Post-rearing cleanup.

IX. Experimental Analysis

Relative to normal food, flies fed ***Psidium guajava extract*** showed significantly fewer ***fecal deposits***, smaller ***total deposit area***, and lower ***IOD***, and this effect was seen in both virgin males and females.

X. Scientific Discovery

In a standardized Drosophila fecal deposit assay, feeding flies with ***Psidium guajava extract*** demonstrated that the platform can be used for ***antidiarrheal*** screening, since reductions in fecal parameters were achieved without affecting ***solid-food intake***.

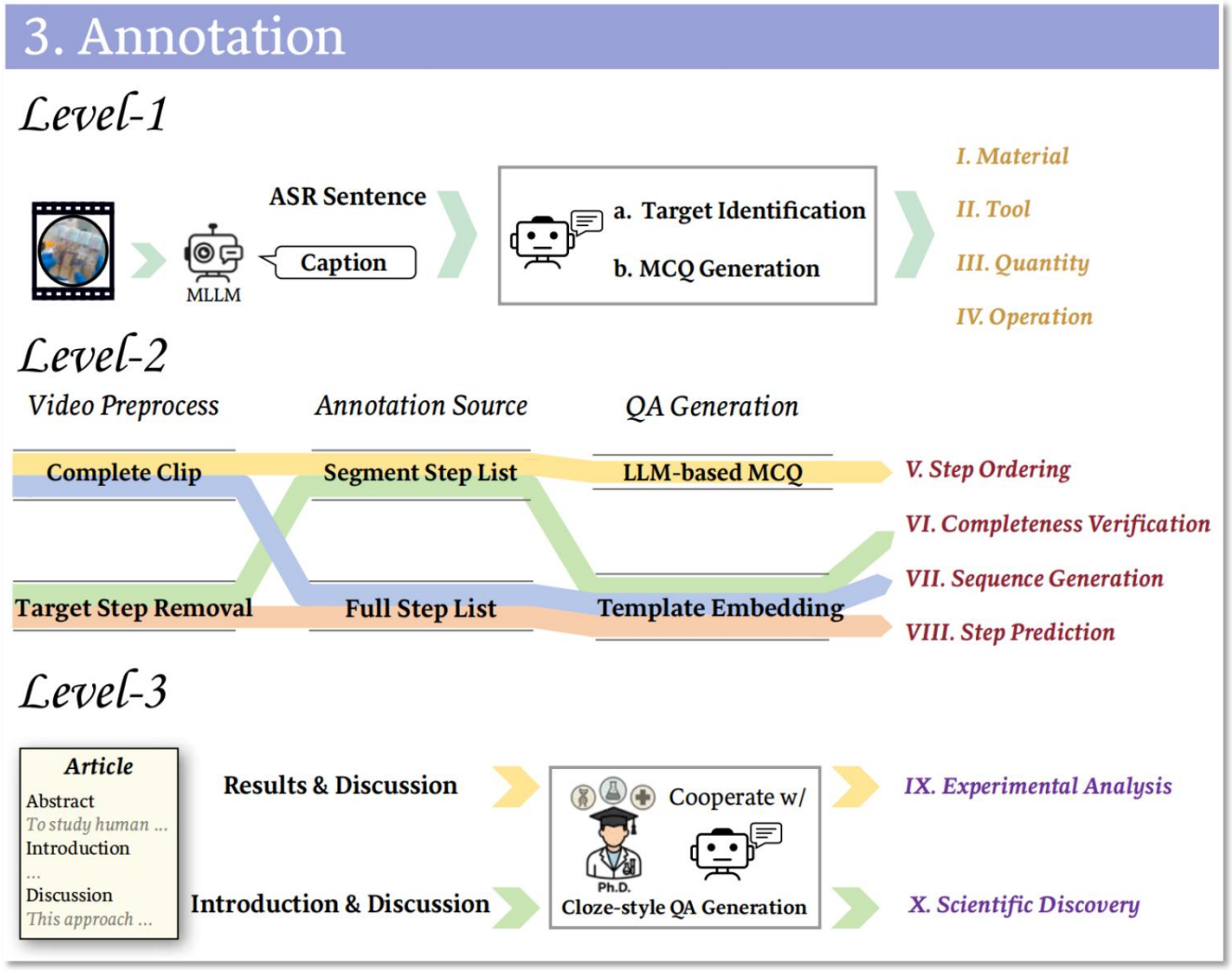
Objective: to evaluate model's scientific reasoning ability over visual experimental process. It has two tasks that require models to integrate visual experiment processes with domain knowledge to draw conclusions, in the form of fill-in-the-blank questions:

- **Experimental Analysis**: Infer crucial conclusions from experimental data, e.g. compare current results with existing studies, highlight new findings, and explain the corresponding mechanisms.
- **Scientific Discovery**: Reason over the entire experiment video, move beyond current outcomes, and abstract broader insights, such as linking results or innovations to larger scientific phenomena, interpreting the significance in filling blanks of which domain or potential application values, and proposing improved solutions for the current limitations and new directions for this area.

04 Annotation

Key Principle — **"Vision-centric"**:
All questions must be answerable only by watching the video.
Tasks avoid relying on text hints or common sense answers.

- **Level-1**: Use subtitles and MLLM-generated captions as visual anchors. LLMs extract materials, tools, quantities, and actions, creating visually similar distractors to ensure answers depend on the video.
- **Level-2**: Based on step lists for video segments and full experiments, construct four task types: ordering, completeness verification, sequence generation, and step prediction.
- **Level-3**: Build fill-in tasks referencing paper Results/Discussion. Questions provide minimal context, hiding key entities or numbers, requiring evidence directly grounded in the video.



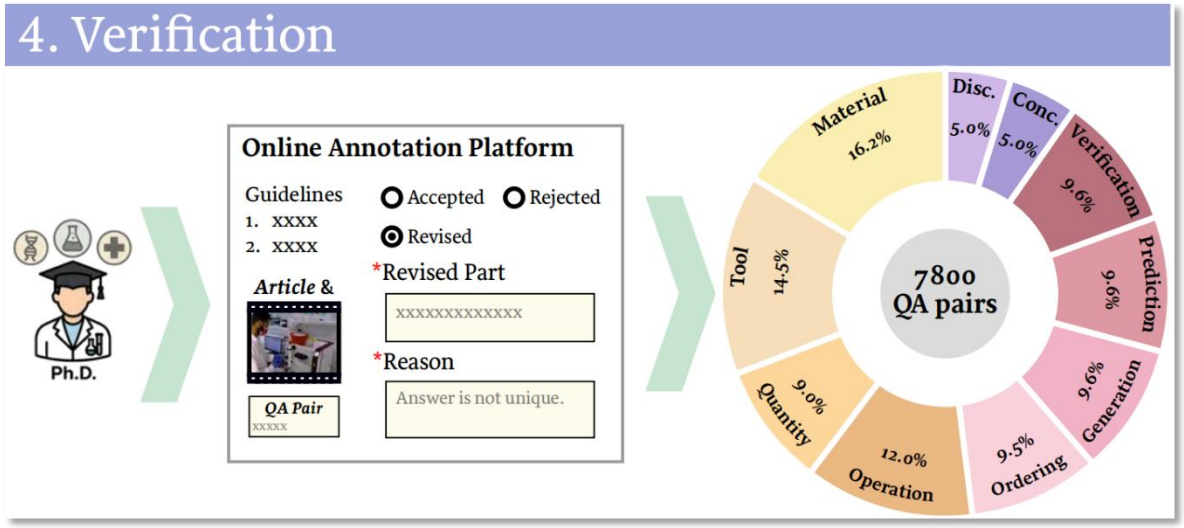
04 Annotation

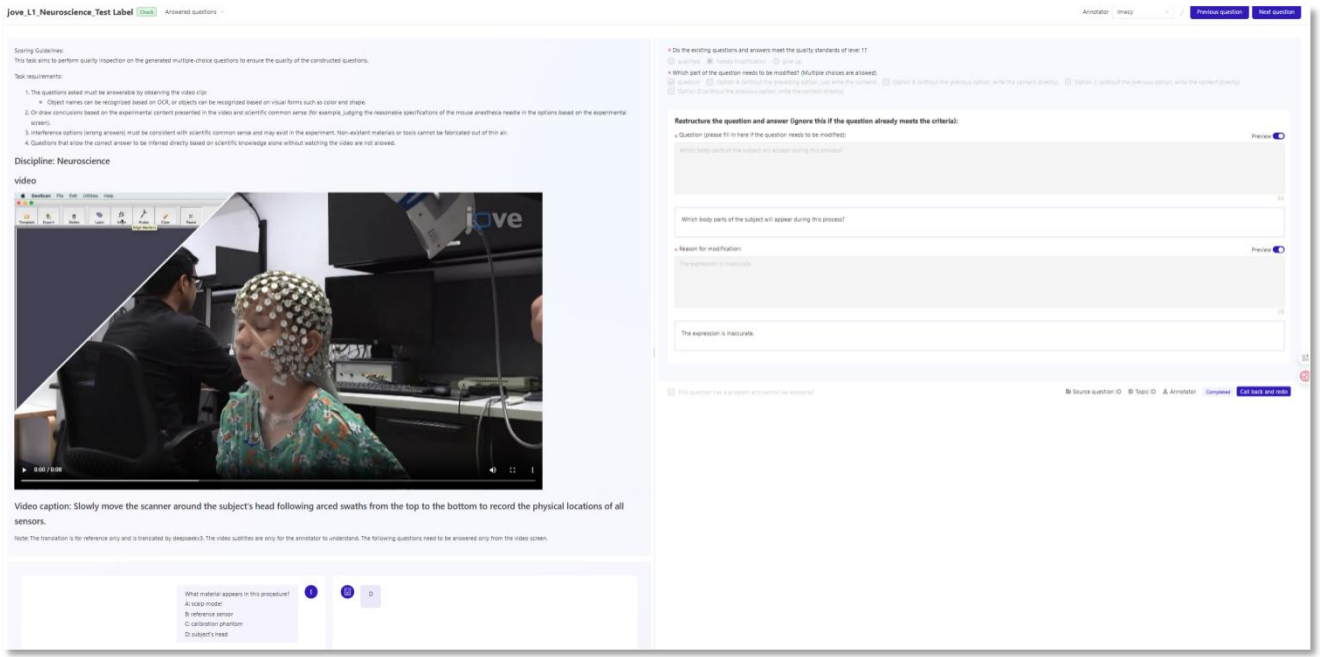
Expert Review Process

In collaboration with data center PhD-level experts in biology, medicine, and chemistry, each task undergoes multi-round review and refinement until meeting benchmark standards.

- **Level-1**: Verify target visibility and clarity in clips; remove visually ambiguous or scientifically unreasonable items.
- **Level-2**: Check consistency between step lists and video segments; delete unsynchronized or incomplete actions and refine vague descriptions.
- **Level-3**: Review fill-in tasks to ensure complete process reasoning; answers must be clear, unique, and textually consistent.

All levels follow the rule: responses must be strictly grounded in corresponding video evidence, with invalid or ambiguous cases revised or discarded.





04

Annotation

Statistics

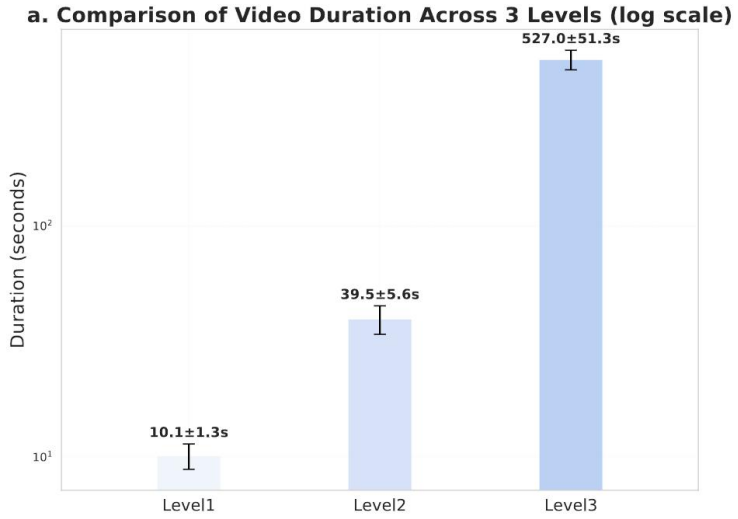
Video Duration

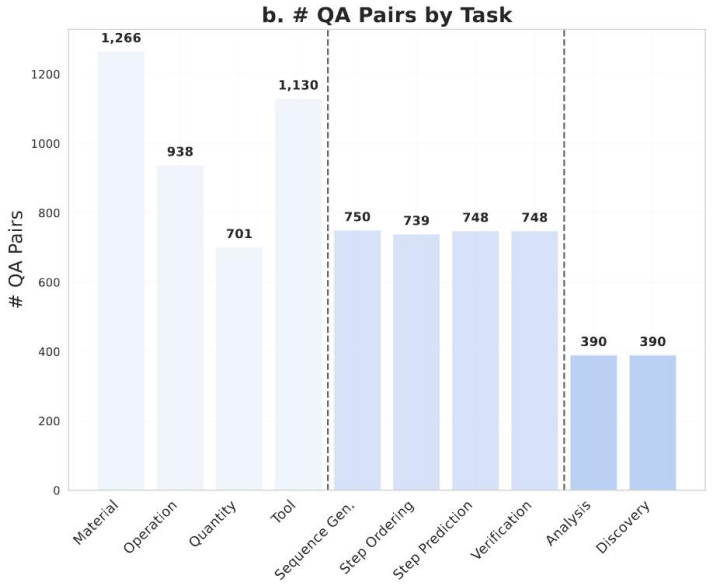
- Level-1: Avg. 10.1s
- Level-2: Avg. 39.5s
- Level-3: Avg. 527.0s

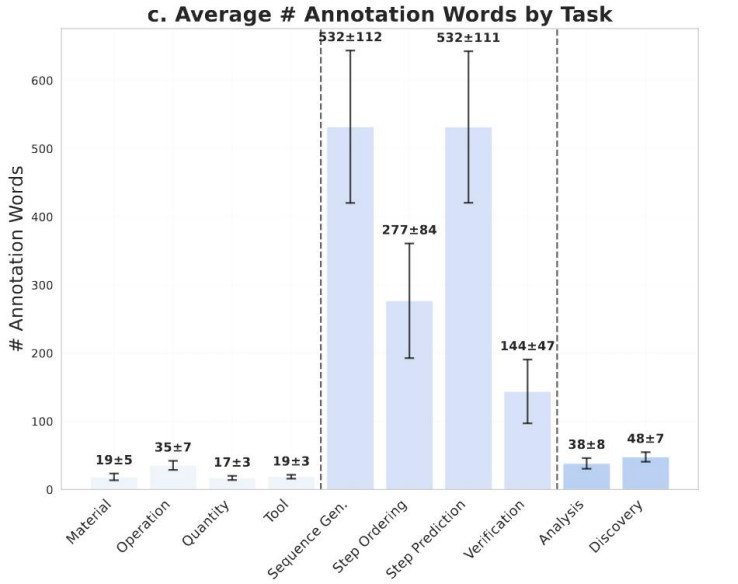
QA Annotation Volume

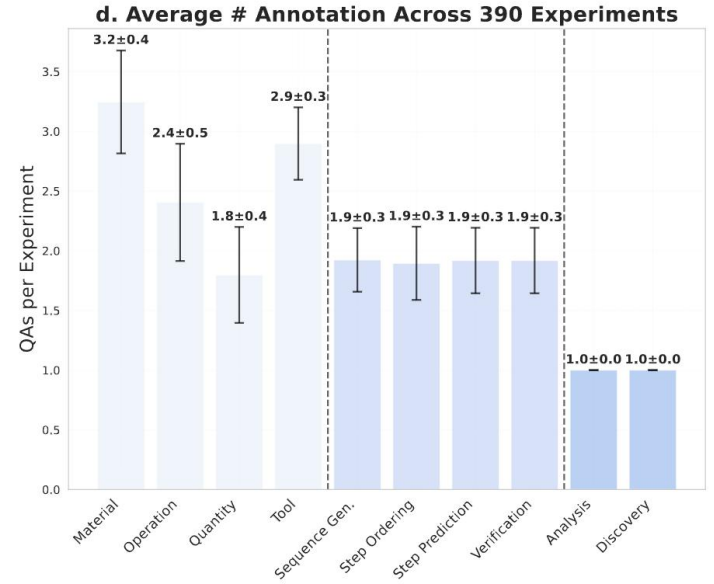
Total: 7,800 QAs

- Level-1: 4 types — 4,035 QAs
- Level-2: 4 types — 2,985 QAs
- Level-3: 2 types — 780 QAs









Level 1 (Fine-grained Perception) Level 2 (Procedural Understanding) Level 3 (Scientific Reasoning)

05 Experiments

Findings:

Closed-source models generally lead across all levels, and the gap widens with task complexity.

- **L1 Perception**: Gemini-2.5-Flash (think) 60.2 vs. InternVL3-78B 50.9 and Intern-S1 49.9. → Open-source models still competitive in basic visual perception, though a gap remains.
- **L2 Process Understanding**: GPT-5 57.5, Gemini-2.5-Pro 54.3, vs. InternVL3-78B 41.9. The gap widens notably. InternVL3-78B excels in step ordering (87.1) but lags in sequence generation (45.5) and step prediction (15.5), showing weaknesses in structured reasoning and generative prediction.
- **L3 Scientific Reasoning**: GPT-5 leads (avg. 56.4, up to 57.4) while the best open-source model, Intern-S1, scores only 39.6. → Closed-source models retain clear advantages in high-level reasoning — a key target for open-source improvement.

Model	Think	Level-1					Level-2					Level-3		
		Tool	Mat.	Quan.	Oper.	Avg.	Ord.	Gen.	Veri.	Pred.	Avg.	Anal.	Disc.	Avg.
Human Performance		17.5	15.9	61.3	55.5	37.6	69.8	31.2	45.6	21.8	42.1	–	–	–
Open-source MLLMs														
Qwen2.5-VL-7B-Instruct	×	32.0	33.9	49.0	62.4	42.6	56.2	20.8	20.7	1.3	24.6	25.2	21.4	23.3
MiMo-VL-7B-RL	×	34.2	33.7	44.2	62.4	42.4	43.9	28.5	18.5	11.4	27.4	28.7	25.9	27.3
MiMo-VL-7B-RL	✓	36.1	29.1	53.6	67.8	44.3	64.8	32.3	24.9	15.6	34.3	29.3	27.3	28.3
InternVL3-8B	×	27.5	31.0	38.8	65.6	39.4	43.4	20.4	20.2	3.9	23.9	29.2	25.3	27.2
InternVL3.5-8B	×	27.3	30.8	45.5	64.8	40.3	82.3	25.8	23.7	4.8	34.0	22.6	18.4	20.5
Intern-S1-mini	✓	33.3	31.2	52.5	61.4	42.5	73.6	14.3	16.8	8.3	28.1	33.5	28.3	30.9
Keye-VL-8B-Preview	✓	16.6	22.4	38.9	60.8	32.6	25.4	12.4	19.1	1.7	14.6	9.5	6.7	8.1
Keye-VL-1.5-8B	✓	21.0	23.4	51.3	64.0	37.0	56.7	9.5	20.0	2.8	22.1	8.4	6.1	7.2
GLM-4.1V-9B	✓	30.8	29.8	47.5	59.6	40.1	64.1	18.2	25.0	7.4	28.6	28.1	26.5	27.3
GLM-4.5V	✓	35.5	33.6	61.5	62.3	45.6	71.9	34.9	27.2	12.9	36.6	33.3	32.5	32.9
Kimi-VL-A3B-Thinking	✓	34.6	32.6	40.7	59.5	40.8	32.3	18.2	23.3	6.2	20.0	24.6	21.8	23.2
InternVL3.5-38B	✓	35.9	34.0	46.7	65.3	44.0	65.8	36.7	23.0	19.0	36.0	33.1	30.8	31.9
InternVL3-78B	✓	35.1	34.3	**73.2**	75.8	50.9	**87.1**	45.5	19.8	15.5	41.9	40.3	35.3	37.7
Qwen2.5-VL-72B-Instruct	×	30.5	34.7	54.5	64.5	43.9	86.3	34.1	23.8	0.3	35.9	31.9	29.3	30.6
Intern-S1	✓	38.9	35.2	58.9	73.8	49.9	82.2	45.0	24.1	15.4	36.0	43.0	36.3	39.6
Closed-source MLLMs														
Seed-VL-1.5	✓	32.9	24.6	43.9	69.2	40.7	73.9	48.6	19.8	27.9	42.5	32.0	29.4	30.7
Claude-Sonnet-4	×	25.6	31.2	54.3	61.9	40.8	78.7	37.6	16.5	11.6	36.0	29.1	30.1	29.6
Gemini-2.5-Flash	×	52.7	50.1	65.2	72.6	58.6	86.0	50.5	24.1	40.2	50.1	47.2	41.1	44.1
Gemini-2.5-Flash	✓	52.7	**50.7**	71.9	73.3	**60.2**	85.1	54.3	22.3	38.0	49.8	44.8	41.3	43.0
Gemini-2.5-Pro	×	**53.1**	45.9	64.3	**80.8**	59.2	83.7	61.3	**26.8**	49.6	53.8	50.6	45.2	47.9
Gemini-2.5-Pro	✓	51.3	44.3	63.8	74.4	56.7	84.2	59.9	**26.8**	46.9	54.3	50.1	44.8	47.4
GPT-5	✓	51.6	37.8	59.5	71.9	53.3	85.1	**66.9**	**26.8**	**51.8**	**57.5**	**55.4**	**57.4**	**56.4**

06

Discussion

Q&A

Paper

https://arxiv.org/abs/2510.11606

Hugging Face Dataset

https://huggingface.co/datasets/OpenGVLab/ExpVid

Github

https://github.com/OpenGVLab/ExpVid

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Thanks for your attention!

Email:yxu040@e.ntu.edu.sg
Wechat: linghan199