MMVU: Measuring Expert-Level Multi-Discipline Video Understanding

Yilun Zhao Haowei Zhang Lujing Xie Tongyan Hu Guo Gan Yitao Long
Zhiyuan Hu Weiyuan Chen Chuhan Li Zhijian Xu Chengye Wang
Ziyao Shangguan Zhenwen Liang Yixin Liu Chen Zhao Arman Cohan

Yale NLP MMVU Team

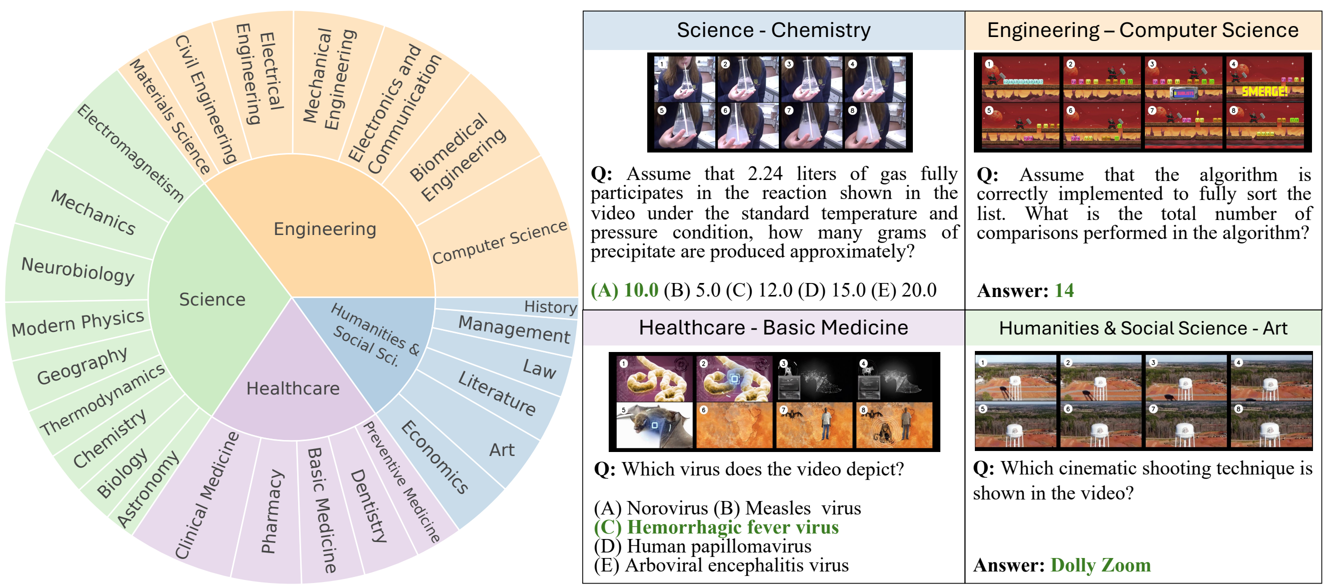


Figure 1. Overview of our benchmark. MMVU includes 3,000 expert-annotated examples, covering 27 subjects across four core disciplines. It is designed to assess foundation models in expert-level, knowledge-intensive video understanding and reasoning tasks.

Abstract

We introduce MMVU, a comprehensive expert-level, multi-discipline benchmark for evaluating foundation models in video understanding. MMVU includes 3,000 expert-annotated questions spanning 27 subjects across four core disciplines: Science, Healthcare, Humanities & Social Sciences, and Engineering. Compared to prior benchmarks, MMVU features three key advancements. First, it challenges models to apply domain-specific knowledge and perform expert-level reasoning to analyze specialized-domain videos, moving beyond the basic visual perception typically assessed in current video benchmarks. Second, each example is annotated by human experts from scratch. We implement strict data quality controls to ensure the high quality of the dataset. Finally, each example is enriched with expert-annotated reasoning rationals and relevant domain knowledge, facilitating in-depth analysis. We conduct an extensive evaluation of 36 frontier multimodal foundation models on MMVU. The latest System-2-capable models, o1 and Gemini 2.0 Flash Thinking, achieve the highest performance among the tested models. However, they still fall short of matching human expertise. Through in-depth error analyses and case studies, we offer actionable insights for future advancements in expert-level, knowledge-intensive video understanding for specialized domains.

1. Introduction

Foundation models have demonstrated remarkable capabilities in reasoning across various domains, yet their ability to handle expert-level knowledge remains a critical area of

Dataset	QA Type	Data Source	College Level?	Detailed Solution Rational?	Detailed Solution Knowledge?
		Text			
MMLU [60]	MC	Exam, course, textbook	✓	✗	✗
MMLU-Pro [147]	MC	Datasets → Human & LLM augment	✓	✗	✗
C-Eval [65]	MC	Exam	✓	✗	✗
SciEval [132]	MC, Open	Internet, datasets → LLM rewrite	✓	✗	✗
TheoremQA [21]	MC, T/F, Open	Internet, exam → Human rewrite	✓	✗	✓
SciKnowEval [42]	MC, T/F, Open	Textbooks, database, other datasets → LLM rewrite	✓	✗	✓
		Text + Image			
VisScience [72]	MC, Open	Internet, exam, textbook	✗	✗	✗
EXAMS-V [32]	MC	Exam	✗	✗	✗
ScienceQA [103]	MC	Internet, course	✗	✓	✗
SceMQA [95]	MC, Open	Internet, exam	✗	✓	✗
CharXiv [149]	Open	arXiv paper → Human annotate	✓	✗	✗
MMSci [94]	MC	Scientific paper → LLM generate	✓	✗	✗
OlympicArena [67]	MC, T/F, Open	Olympic competitions	✓	✓	✗
MMMU [161]	MC, Open	Internet, exam, textbook	✓	17.6%	✗
CMMMU [164]	MC, T/F, Open	Internet, exam, textbook	✓	2.1%	✗
MMMU-Pro [162]	MC	MMMU → Human & LLM augment	✓	15.4%	✗
		Text + Video			
MMWorld [58]	MC	Human experts (24%) / LLM-gen (76%)	39.5%	✗	✗
MMVU (ours)	MC, Open	Human experts annotate from scratch	✓	✓	✓

Table 1. Comparison between MMVU and existing multi-discipline benchmarks for evaluating foundation models. In the "QA Type" column, "MC" denotes Multiple-Choice questions, "Open" denotes Open-ended questions, and "T/F" denotes True-False questions.

evaluation [60, 161]. In recent years, researchers have developed numerous benchmarks to assess these models' proficiency in specialized domains, primarily focusing on text-based reasoning [42, 60, 132, 147] and image-based contexts [94, 104, 161, 162, 164]. However, as capabilities of foundation models expand across multiple modalities, there is a significant gap in evaluating expert-level reasoning over specialized-domain *videos*. This gap is particularly concerning as video is one of the most information-rich and naturalistic modalities, and is widely used to convey complex, dynamic information in specialized fields like healthcare, engineering, and scientific research [58]. Unlike static text or images, expert-level videos often capture temporal dynamics, procedural knowledge, and complex interactions that are essential in many specialized domains. For example, in science, expert-level and knowledge-intensive reasoning might involve analyzing a chemical reaction video (Figure 1). A model must identify key reaction stages based on subtle visual cues like color changes or the formation of precipitates, which requires integrating chemical knowledge in addition to recognizing visual patterns.

To bridge this gap, we introduce MMVU, a comprehensive benchmark measuring **M**ultimodal foundation models in expert-level, **M**ulti-discipline **V**ideo **U**nderstanding and reasoning. MMVU consists of 3,000 expert-annotated QA examples over 1,529 specialized-domain videos, spanning 27 subjects across four key disciplines: Science, Healthcare, Humanities & Social Sciences, and Engineering. To ensure both the breadth of domain knowledge and the depth of reasoning required for MMVU, we implement a textbook-guided data annotation process. Expert annotators first locate key concepts from textbooks in their fields, then source relevant videos and create corresponding questions that require domain knowledge and expert-level reasoning to comprehend the videos. Each example also includes expert-annotated reasoning rationale and relevant domain knowledge, facilitating fine-grained evaluation of model performance. Thorough data quality controls are implemented to ensure high quality of MMVU.

We conduct an extensive evaluation on MMVU, covering 36 frontier multimodal foundation models from 17 organizations. Notably, the latest o1 model demonstrates the highest performance among all tested models, approaching the expertise of human experts. Despite this progress, other models still fall noticeably short of human-level capabilities. For instance, GPT-4o achieves a score of 66.7%, which is substantially lower than the benchmark set by human experts (*i.e.*, 86.8%) in the open-book setting. Our analysis highlights the effectiveness of CoT reasoning, which generally enhances model performance compared to directly generating final answers without intermediate reasoning steps. To deepen understanding of the current models' limitations, we perform an in-depth error analysis of frontier models, including numerous case studies reviewed by human experts. These insights provide valuable guidance for future advancements in the field.

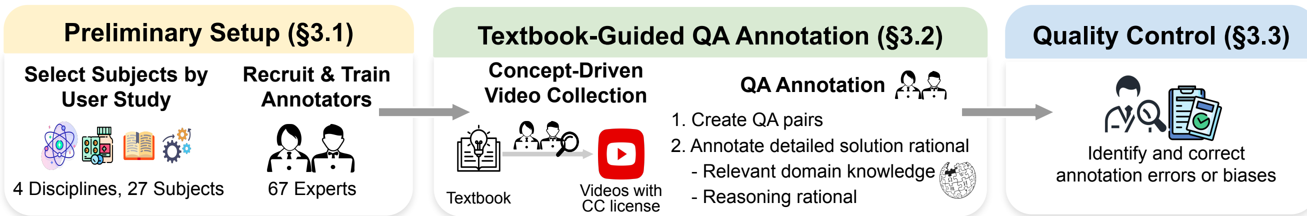


Figure 2. An overview of the MMVU benchmark construction pipeline.

2. Related Work

Video Understanding Benchmark. Existing video understanding benchmarks primarily focus on *general-purpose* video comprehension tasks, such as action recognition [34, 59, 100, 130], captioning and description [81, 89, 134, 153, 157], grounding [23, 74, 85, 145], temporal reasoning [17, 29, 70, 75, 91, 101, 129], and long video understanding [7, 39, 112, 144, 166]. The rise of video-based foundation models [41, 66, 135, 165] has driven the development of new benchmarks that include diverse video comprehension tasks for more comprehensive evaluation [48, 76, 90, 92, 115, 156, 160]. However, these benchmarks remain predominantly focused on natural scenes and general-purpose tasks. A significant gap persists in benchmarks targeting *expert-level* and *knowledge-intensive reasoning* over specialized-domain videos, where both visual perception and domain-specific expertise are required—especially in critical fields like healthcare, engineering, and science [58].

Multi-discipline Evaluation Benchmark. The rapid development of foundation models has significantly enhanced expert-level reasoning across various disciplines [53, 71, 118, 136, 159]. Early benchmarks focused on domain-specific tasks for textual domains, establishing a foundation for assessing the models' strengths and limitations in expert reasoning [21, 27, 60, 133, 147, 150, 168, 169]. More recently, benchmarks have evolved to include multimodal tasks [94, 104, 149, 161, 162, 164], emphasizing visual perception and advanced reasoning with domain knowledge. However, these efforts remain largely limited to *static* images. Developing a high-quality, multidisciplinary video benchmark presents greater challenges than those for text or image-based tasks due to the scarcity of suitable resources (*e.g.*, textbooks or exam questions). This leaves the critical modality of videos and video-based expert-level reasoning significantly underexplored. Recent work, MMWorld [58], has made pioneering strides by incorporating videos across multiple disciplines. However, only a limited portion of its dataset (39.5%) requires domain-specific expertise, and 76.4% of the examples are generated by the GPT-4V model. Moreover, most existing benchmarks provide only the ground-truth answer, restricting researchers' ability to conduct a fine-grained evaluation. To address this limitation, MMVU includes expert-annotated reasoning rationales and relevant domain knowledge for each example, enabling a more nuanced assessment of expert-level reasoning. Table 1 further distinguishes the difference between MMVU and existing multi-discipline benchmarks.

3. MMVU Benchmark

We present MMVU, a comprehensive evaluation benchmark that focuses on measuring progress on knowledge-intensive, expert-level reasoning in the video modality. MMVU has the following key features: (1) **Breadth of Domain Knowledge**: We employ a textbook-guided QA annotation pipeline to ensure the wide coverage of domain knowledge within each subject (§3.2). (2) **Depth of Expert-level Reasoning**: Each example in MMVU requires models to comprehend specialized-domain video context, applying expert knowledge and reasoning (§3.2). (3) **True Visual Understanding**: Recent studies [20, 162, 167] have shown that visual content is unnecessary for many examples in current multimodal benchmarks. To alleviate this issue, each example in MMVU is carefully validated by human experts to confirm that video comprehension is required for accurate answering (§3.3). (4) **Support of Fine-grained Evaluation**: We provide expert-annotated solutions and the requisite knowledge for each example (§3.2), enabling more comprehensive analysis for future research (§4.3). Figure 2 provides an overview of the three stages involved in constructing MMVU, which is detailed in the following subsections.

3.1. Preliminary Setup

We first discuss the preliminary setup for data construction.

Subject Selection. To ensure a broad and accurate representation of expert-level video understanding across diverse disciplines, we conduct a user study involving 133 college and graduate students for subject selection. We ask them to curate two QA examples requiring expert-level video understanding in subjects relevant to their field of study, and provide feedback on their experiences during the curation

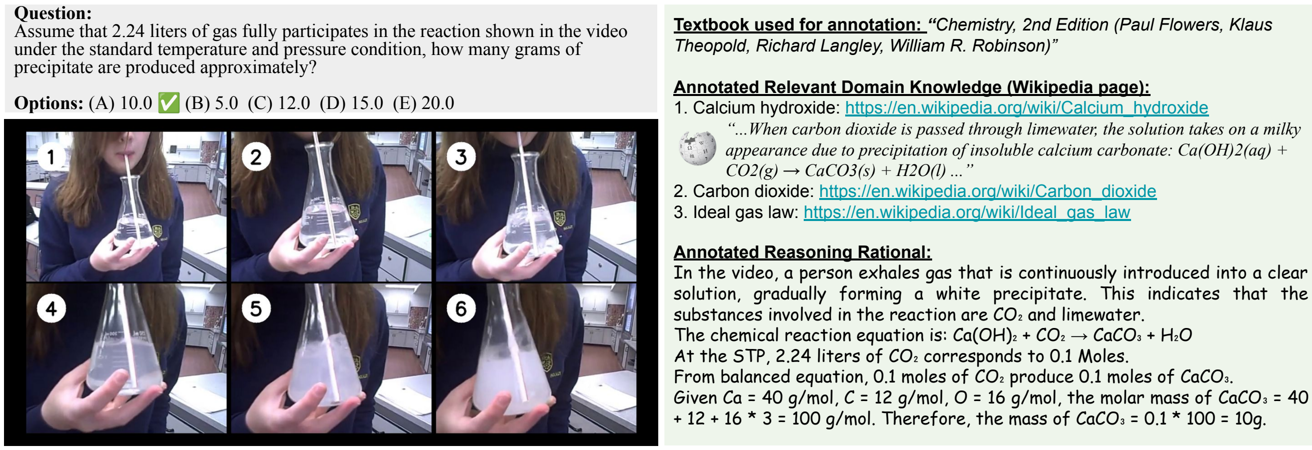


Figure 3. A dataset example from MMVU with the discipline of chemistry. Each example in MMVU includes expert annotation of relevant domain knowledge and step-by-step reasoning rational.

process. Such a user study-guided approach helps us identify subjects within each discipline that may not be obvious from a top-down selection process. It also offers insights into the challenges of designing expert-level video examples, helping us design and refine the textbook-guided QA annotation process (detailed in §3.2). The authors manually analyze the collected examples and select **27 subjects** (as listed in Figure 1) across four disciplines that align best with our benchmark's construction desiderata discussed earlier.

Expert Annotator Recruitment and Training. For each subject, we assign at least two annotators with relevant expertise. We include 67 expert annotators (detailed biographies are presented in Appendix A.1), comprising 22 third- or fourth-year undergraduate students, 36 graduate students, and nine of the authors. All the annotators also participated in our initial user study. Each annotator is required to finish a training session to learn the annotation protocol (detailed in Appendix A.3) before official annotation.

3.2. Textbook-Guided QA Example Annotation

Constructing a high-quality, expert-level, multi-disciplinary benchmark for video-based tasks is more challenging than the ones for text- or image-based, as there is no existing resources (*e.g.*, textbooks or exam questions) that can adapted from and each example has to be curated from scratch. Therefore, it is crucial to establish a structured approach that ensures the quality and comprehensiveness of the benchmark. We employ a textbook-guided example annotation pipeline designed to capture both the *breadth of knowledge* and *depth of reasoning*. In brief, annotators first identify key concepts from the textbook and locate relevant videos that align with these concepts. The textbooks for each subject (listed in Appendix A.2) are selected by expert annotators and are recognized as authoritative references in their respective fields. Annotators then curate QA examples and detailed solution rationales. We detail the annotation procedure as follows:

Concept-Driven CC-Licensed Video Collection. Annotators are instructed to first review each chapter of the textbook to identify key concepts that inherently require dynamic visual representation, such as experimental procedures in science or mechanical operations in engineering. They then search for related videos on YouTube having Creative Commons (CC) license that effectively illustrate the selected concept. The CC license enables reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. We use YouTube Data API v3[1] to verify the license type. Existing video benchmarks typically utilize YouTube videos, yet do not confine their selections to content with CC licenses, introducing potential copyright concerns. We recognize that by restricting our selection to CC-licensed content, we are compelled to forgo coverage of certain subjects (*e.g*. sports), where CC-licensed videos is scarce. To ensure the collected videos effectively challenge the model's visual reasoning capabilities, the video should be vision-intensive, requiring models to focus solely on visual information for comprehension. To this end, we ensure that audio tracks are excluded to eliminate potential shortcuts models might exploit through auditory cues; and the video should contain minimal on-screen text, as an overabundance of text may detract from the core visual understanding task. Consequently, videos such as lecture recordings, which typically include slides or text-based explanations that simplify the task of answering associated questions, are excluded.

[1] https://developers.google.com/youtube/v3

QA Annotation. After identifying suitable videos, annotators are required to create two or three questions, either multiple-choice or open-ended. Each question is designed to test the model's expert-level reasoning by applying domain-specific knowledge to interpret the video content and derive a solution. Annotators are also required to specify the start and end timestamps of the video clip relevant to answering each question. For annotating multi-choice question, the annotators are required to carefully craft the four distractor options to reflect common misconceptions or plausible alternatives, ensuring that models cannot easily eliminate incorrect options without reasoning over video content. Once the five options are finalized, the annotation interface randomly shuffles them.

Solution Rationale Annotation. For each annotated question, annotators must also provide detailed solution for the correct answers. As shown in Figure 3, the solution comprises two key components: (1) *relevant domain knowledge*, which includes a list of domain-specific concepts or keywords necessary for answering the question, with each concept linked to its corresponding Wikipedia page. (2) *reasoning rationale*, which details the step-by-step reasoning process to reach the correct answer. These solution annotations are critical for enhancing transparency in the evaluation process and facilitating future research focused on understanding model failure modes.

3.3. Data Quality Control

We next discuss our methods to ensure high data quality.

Time-Based Annotation Compensation. As discussed earlier, annotating examples for MMVU can be particularly time-intensive, especially when there is limited availability of videos with Creative Commons licenses in the required subjects. To accommodate this and ensure a high-quality benchmark, we compensate annotators based on the time they spend rather than the number of examples completed, preventing them from rushing through tasks (See Appendix A.5 for annotation compensation details). On average, annotating one example takes 20 minutes and 17 seconds, while validation requires 4 minutes and 12 seconds.

Human Expert Validation. To ensure that the final dataset remains high-quality and meets expert-level standards without introducing unnecessary biases, each example in MMVU undergoes expert review by one of the authors or top-performing annotators to verify the accuracy of its annotations. Recent studies [20, 129, 162, 167] have shown that visual content is unnecessary for many examples in current multimodal benchmarks. To address this concern, each example in MMVU is carefully validated by human experts to ensure that video comprehension is required for accurate answering. If an example is determined to be answerable solely through the textual components of the question, a single video frame, or if it contains annotation errors, evaluators first attempt to revise the example. If revision is not feasible, detailed feedback is provided to the original annotator, who then revises and submits it for a second iteration. A total of 523 examples were revised during the data validation process. Among them, 72 examples were still found to be misaligned with our design criteria and were excluded from the final benchmark. Overall, $1 - \frac{523}{3{,}000+72} = 83.0\%$ of the initial examples met our design criteria without requiring revisions, indicating the high quality of initial annotation.

Statistics	**Value**
Total Questions	3,000
Validation Set	1,000
Test Set	2,000
Unique Videos	1,529
Video Length (Seconds, `avg/max`)	51.4 / 228
Number of Disciplines	4
Number of Subjects	27
Multiple Choice Questions	1,858
Question Length (`avg/max`)	16.8 / 70
Single Choice Length (`avg/max`)	7.6 / 42
Number of Choices per Question	5
Open-ended Questions	1,142
Question Length (`avg/max`)	16.4 / 39
Ground-truth Answer Length (`avg/max`)	1.5 / 7
Number of Required **Knowledge** per Question (`avg/max`)	4.3 / 7
Solution Rationale Length (`avg/max`)	56.6 / 193
Total Number of Unique Knowledge (*i.e.*, Wikipedia pages)	4,770

Table 2. Key statistics of the MMVU benchmark.

3.4. MMVU Benchmark Analysis

Data Statistics. Table 2 presents the key statistics of MMVU. It consists of 3,000 examples, which are randomly divided into two subsets: validation and test. The validation set contains 1,000 examples, and is intended for model development and validation. The test set, comprising the remaining 2,000 examples, is strictly reserved for standard evaluation to prevent data contamination [35, 50, 69]. To further promote fair benchmarking, the test set remains hidden. We are developing an online evaluation pipeline on a public platform, enabling researchers to benchmark their models and participate in a public leaderboard.

Human Performance. To provide a rough but informative estimate of human-level performance on MMVU, we randomly sampled 30 questions per discipline from the test set, resulting in a total of 120 questions for evaluation. Five participants—three graduate students specializing in biology, anesthesiology, and East-Asian literature, along with two of the authors—individually answered these questions. The evaluation proceeded in three phases: (1) **Closed-book Setting**: In the first phase, participants had 3.5 hours to answer questions without access to external resources. The average accuracy across the four participants was 49.7%. (2) **Open-book Setting**: In the second phase, participants were permitted to use external resources (*e.g.*, internet and textbooks) to review answers they felt uncertain about. They were not informed of the correctness of their initial responses, and a 4-hour time limit was set. This open-book approach led to an increase in average accuracy to 86.8%. (3) **Oracle Setting**: Finally, participants were required to revise each incorrect answer based on ground-truth domain knowledge and self-sourced online resources. The average accuracy after this final revision was 95.3%.

4. Experiments

This section discusses experiment setup and key findings.

4.1. Experiment Setup

Evaluated Multimodal Foundation Models. To establish a comprehensive understanding of the challenges posed by MMVU and provide reference points for future research, we evaluate a broad range of frontier multimodal foundation models that support *video* or *multiple images* as input. Specifically, we evaluate **19 series of open-source models**, including InternVL-2 & 2.5 [22, 24], Qwen2-VL [143, 159], LLaVA-NeXT [99], Pixtral [111], DeepSeek-VL2 [154], H2OVL Mississippi [49], Idefics2 [84], Aria [87], LLaVA-NeXT-Video [88], LLaVA-OneVision [86], Llama-3.2-Vision [37], Phi-3.5-Vision [1], InternVideo2 & 2.5 [146, 148], VideoChat-Flash [93], and VideoLLaMA 2 & 2.1 & 3 [25, 163]. We also evaluate **eight series of proprietary models**, including OpenAI o1 [117] and GPT-4o [118], Gemini-1.5 & 2 and Gemini-Thinking [53], GLM-4V-Plus [51, 63], Grok-2-Vision [155], and Claude-3.5 [4]. For open-source models, we prioritize the vLLM pipeline [83] for model inference; otherwise, we use the Transformers pipeline [152]. We use the official API service for proprietary models. For models without native video support, following VideoMME [48], we provide visual input using the maximum number of images that fits within the model's context window. §B.1 details the parameter settings and model configurations. We evaluate the models with both **Direct Answer** and **Chain-of-Thought** prompts (presented in Appendix B.2), which is adapted from the versions used in MMMU-Pro [162].

Accuracy Evaluation. We use accuracy as the primary metric to evaluate model performance on MMVU. Following recent benchmarks for foundation model evaluation [58, 104, 149], we employ GPT-4o to assess accuracy. Specifically, given a question, its ground truth answer, and the model's response, GPT-4o is instructed to extract the final answer from the model response and determine its correctness. The evaluation prompts for both multiple-choice and open-ended questions are presented in Appendix B.3.

4.2. Main Findings

Section 4.1 presents the evaluated models' CoT performance on MMVU, while Figure 4 illustrates a comparison between the model performance in CoT reasoning and direct answering. Our key findings are as follows:

MMVU presents substantial challenges for current multimodal foundation models. Even the top-performing model falls well short of human expert performance. For instance, GPT-4o achieves 66.7% accuracy with CoT prompting, significantly lower than the 86.8% accuracy achieved by human experts in an open-book setting. Notably, while GPT-4o has narrowed the performance gap with human experts in text-based expert-level reasoning on MMLU (88.7% vs 89.8% [60]) and image-based expert-level reasoning on MMMU (69.1% vs 82.6% [161]), the gap remains large on MMVU. This disparity underscores MMVU's critical role in advancing and evaluating multimodal foundation models' capabilities in video-based expert reasoning across specialized domains.

Performance of open-sourced models. As for open-source multimodal foundation models, they still lag behind the proprietary models. However, the Qwen2-VL-72B and DeepSeek-VL2 models have achieved performance levels that exceed human benchmarks in closed-book settings and are approaching the performance of leading proprietary models. These advancements highlight the significant progress being made in open-source model development.

CoT reasoning generally improves model performance compared to directly outputting the answer. However, the degree of improvement varies across different foundation models. For instance, Claude 3.5 Sonnet demonstrated a remarkable enhancement, achieving a notable performance gain of 11.0%, as corroborated by the findings in MMMU-Pro [162]. Conversely, models like GPT-4o exhibited only marginal improvements. These results indicate that the impact of CoT reasoning is not uniformly beneficial across all models on MMVU.

System-2 thinking demonstrates effectiveness. Models capable of System-2 thinking and employing long CoT demonstrate significant performance advantages. Notably,

	Release	Test Set				Avg. Validation	Avg. Test
		Science	Healthcare	Human. & Social Sci.	Engineering		
Human Performance							
Human Oracle		95.3	93.3	96.0	96.7	95.3	
Human Open-book		86.7	84.7	92.7	83.3	86.8	
Human Closed-book		54.7	42.7	44.7	56.7	49.7	
Proprietary Models							
o1	2024-12	76.5	80.1	80.9	71.9	75.5	76.1
Gemini 2.0 Flash Thinking	2024-12	69.3	71.2	73.4	67.3	69.1	69.5
GPT-4o	2024-08	67.2	71.8	72.0	61.6	67.4	66.7
Gemini 2.0 Flash	2024-12	70.8	62.7	71.6	63.0	65.9	66.5
Gemini 1.5 Pro	2024-09	67.2	68.1	67.0	62.8	65.4	65.8
Claude 3.5 Sonnet	2024-10	60.5	64.0	70.9	64.5	65.2	64.1
Grok-2-Vision	2024-12	60.6	72.5	72.0	57.4	62.7	63.4
GPT-4o-mini	2024-07	60.3	60.9	70.6	59.3	61.6	61.5
Gemini 1.5 Flash	2024-09	56.8	57.3	66.3	58.2	58.8	58.8
GLM-4V-Plus	2025-01	52.2	57.3	64.9	55.4	56.2	56.2
Open-sourced Models							
Qwen2-VL-72B	2024-09	48.0	53.6	61.7	53.9	53.0	53.2
DeepSeek-VL2	2024-12	50.3	53.4	58.9	48.6	52.1	51.5
InternVL2.5-38B	2024-11	50.3	45.6	52.8	52.8	50.5	50.7
Aria	2024-11	46.8	43.3	61.0	49.9	49.3	49.3
InternVideo2.5-8B	2025-01	47.6	50.0	54.3	44.9	48.3	48.0
Llama-3.2-90B-Vision	2024-09	46.5	43.5	53.9	48.1	47.1	47.6
VideoLLaMA3-7B	2025-01	46.5	47.9	57.4	43.5	45.0	47.2
DeepSeek-VL2-Small	2024-12	47.5	48.7	47.5	45.1	46.9	46.9
VideoChat-Flash-7B	2025-01	43.6	50.8	50.7	41.5	45.1	45.2
Qwen2-VL-7B-Instruct	2024-08	43.6	42.5	43.6	41.2	42.1	42.5
InternVL2.5-8B	2024-11	39.2	36.8	47.2	42.3	41.1	41.0
VideoLLaMA2.1-7B	2024-10	35.3	38.9	45.4	41.6	39.5	39.8
VideoLLaMA3-2B	2025-01	40.0	42.7	47.5	34.6	38.7	39.6
Llama-3.2-11B-Vision	2024-09	40.5	39.4	44.0	35.7	38.9	39.0
Phi-3.5-Vision	2024-08	38.3	29.5	45.4	41.1	38.1	38.7
LLaVA-OneVision-7B	2024-09	34.3	38.6	40.8	38.8	37.9	37.7
Qwen2-VL-2B	2024-08	32.6	40.9	40.4	35.7	36.5	36.5
InternVL2-8B	2024-06	36.7	32.9	36.9	37.2	36.3	36.2
Idefics3-8B	2024-08	37.0	35.5	44.0	31.2	35.3	35.6
VideoLLaMA2-7B	2024-06	32.3	27.7	44.3	35.7	34.4	34.4
DeepSeek-VL2-Tiny	2024-12	34.3	33.4	35.8	30.1	33.0	32.8
Pixtral-12B	2024-09	36.1	24.6	37.9	30.8	32.3	32.2
LLaVA-NeXT-Video-34B	2024-06	31.8	24.6	35.8	30.3	30.5	30.4
InternVideo2-8B	2024-08	29.6	31.1	37.2	26.5	29.9	29.9
H2OVL Mississippi-2B	2024-10	29.1	29.5	29.4	28.0	29.1	28.8
LLaVA-NeXT-Video-7B	2024-06	27.0	31.1	27.3	29.5	28.6	28.7

Table 3. Accuracy of evaluated foundation models on the MMVU validation and test sets using CoT prompts. Model performance is ranked based on overall results on the test set. *: For o1, as the API access for its multimodal version has not been granted, we randomly sampled 100 examples from the validation set and 200 examples (50 for each core discipline) from the test set.

the o1 and Gemini 2.0 Flash Thinking models achieved the top two results on MMVU, illustrating that increasing test-time compute and applying long CoT can significantly enhance model performance in expert-level video reasoning tasks. These results highlight the potential of developing open-source models designed to facilitate and advance System-2 thinking capabilities.

4.3. Qualitative Analysis

To gain a deeper understanding of the capabilities and limitations of frontier models on MMVU, we perform comprehensive case studies and error analysis by humans. The inclusion of expert-annotated reasoning rationales and domain knowledge for each example in MMVU facilitate a more effective analysis compared to datasets that provide

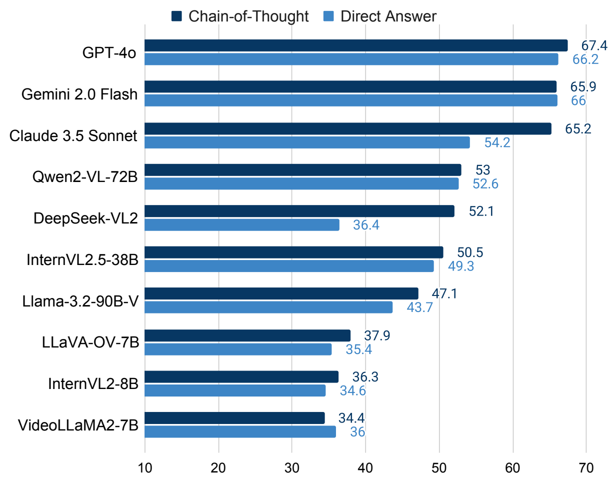


Figure 4. Comparison of model performance between CoT and direct answering on validation set. Full results are shown in §C.1.

only answers. We focus on four top-performing models, GPT-4o, Qwen2-VL-72B, Llama-3.2-90B-Vision, and DeepSeek-VL2, for human evaluation. From the MMVU validation set, we randomly sample 50 error cases for each model. These cases are analyzed by the authors using ground-truth features (*i.e*. expert-annotated reasoning rationales and required domain knowledge) as references. We identify the following six primary errors, with illustrative examples provided in Appendix C:

Visual Perception Error (18%): The model fails to accurately interpret spatial, temporal, or semantic aspects of visual information within a video. Additionally, it might "hallucinate", detecting objects or events that are not actually present in the video. Figure 16 is a typical instance where the model fails to correctly perceive the traversal order of binary tree. Figure 18 shows that the model mistakenly identifies the device shell in the video as water, leading to completely wrong reasoning about the device's function.

Misuse or Lack Domain Knowledge in Visual Perception (20%): The model fails to apply the domain-specific expertise required to accurately interpret specialized concepts or elements within the video. For example, in a medical video, it may identify objects but fail to recognize their technical terms or misunderstand their importance within the procedure being demonstrated. Moreover, as shown in Figure 20, the model correctly perceives the ascending numbers (array indices), but misuses its pretrained knowledge and misidentifies them as the numbers to be sorted. It leads to the wrong conclusion that the video demonstrates a sorting algorithm. This limitation underscores a gap in model's ability to integrate domain knowledge with visual perception effectively.

Misuse or Lack Domain Knowledge in Reasoning (27%): The model fails to effectively recall and apply domain knowledge during its reasoning processes. For instance, when addressing questions over chemistry videos, it may fail to correctly apply relevant chemical equations, leading to errors in computing the reaction mass. A notable example is Figure 23, where the model misuses the domain knowledge that bats often live in unsanitary environments and makes the wrong inference that poor hygiene conditions are the cause of virus outbreaks. Besides, in Figure 26, the model lacks the domain knowledge about relevant chemical equations, so that it cannot correctly answer the question. This limitation underscores the model's inability to integrate domain knowledge into its reasoning processes effectively.

Heavy Reliance on Textual Information (20%): The model predominantly depends on textual information for problem-solving, especially when addressing multiple-choice questions, as it evaluates each option individually without leveraging the actual video content. For instance, Figure 27 shows the model ignores the video information about the reason of the disease and overly focuses on the textual question. Similar limitations have been observed in other multimodal benchmarks [48, 161]. Future work could enhance multimodal reasoning by more effectively incorporating non-textual content into the reasoning process.

Logical Reasoning Error (6%): The model exhibits inconsistencies between its reasoning process and final answer, leading to self-contradiction. As depicted in Figure 29, the analysis of one specific option contradicts with the other reasoning steps, which is a typical self-contradiction logical error.

Other Error (9%): This includes refusing to answer a question due to insufficient context or safety concerns, exceeding the output limit, generating repetitive information, or making incorrect math computation.

5. Conclusion

We introduce MMVU, a high-quality, multi-disciplinary benchmark designed to assess the expert-level, knowledge-intensive reasoning capabilities of multimodal foundation models on specialized-domain videos. We employ a textbook-guided example annotation pipeline designed to capture both the breadth of knowledge and depth of reasoning. In our evaluation of 36 frontier multimodal foundation models, we find that while the latest o1 model achieves the highest performance among all tested models—approaching human expert-level proficiency—a notable performance gap remains between other models and human experts. Additionally, models employing CoT reasoning consistently outperform those that generate final answers directly. Through comprehensive error analysis and case studies, we identify persistent challenges of MMVU, offering valuable insights for advancing foundation models' capabilities to achieve expert-level video understanding.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 6th edition, 2014.

[3] Phillip E Allen and Douglas R Holberg. *CMOS analog circuit design*. Elsevier, 2011.

[4] Anthropic. Introducing the next generation of claude, 2024.

[5] Mumtaz Anwar, Riyaz Ahmad Rather, and Zeenat Farooq. *Fundamentals and advances in medical biotechnology*. Springer, 2022.

[6] Steven Ascher and Edward Pincus. *The Filmmaker's Handbook: A Comprehensive Guide for the Digital Age*. Plume, Penguin Random House, 5th edition, 2012.

[7] Kirolos Ataallah, Chenhui Gou, Eslam Abdelrahman, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. Infinibench: A comprehensive benchmark for large multimodal models in very long video understanding, 2024.

[8] Peter William Atkins, Julio De Paula, and James Keeler. *Atkins' physical chemistry*. Oxford university press, 2023.

[9] Eugene A. Avallone, Theodore Baumeister, and Ali M. Sadegh. *Marks' Standard Handbook for Mechanical Engineers*. McGraw-Hill Education, 12th edition, 2018.

[10] Ashwani Bedi and Ramsey Dabby. *Structure for Architects: A Case Study in Steel, Wood, and Reinforced Concrete Design*. Routledge, 1st edition, 2019.

[11] Fred G Bell. *Engineering geology and construction*. CRC Press, 2004.

[12] Olivier Blanchard. *Macroeconomics*. Pearson, 9th edition, 2024.

[13] David S. Bright, Anastasia H. Cortes, et al. *Principles of Management*. OpenStax, Rice University, 2019. Available at https://openstax.org/details/books/principles-management.

[14] Theodore L. Brown, H. Eugene LeMay, Bruce E. Bursten, Catherine J. Murphy, Patrick M. Woodward, and Matthew E. Stoltzfus. *Chemistry: The Central Science*. Pearson, 15th edition, 2023.

[15] Laurence L. Brunton, Randa Hilal-Dandan, and Bjorn Knollman. *Goodman & Gilman's: The Pharmacological Basis of Therapeutics*. McGraw-Hill Education, 13th edition, 2017.

[16] Randal E Bryant and David Richard O'Hallaron. *Computer systems: a programmer's perspective*. Prentice Hall, 2011.

[17] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models, 2024.

[18] William D Callister Jr and David G Rethwisch. *Materials science and engineering: an introduction*. John wiley & sons, 2020.

[19] Krishan K. Chawla. *Composite Materials: Science and Engineering*. Springer, 3rd edition, 2012.

[20] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024.

[21] Wenhu Chen, Ming Yin, Max Ku, Pan Lu, Yixin Wan, Xueguang Ma, Jianyu Xu, Xinyi Wang, and Tony Xia. TheoremQA: A theorem-driven question answering dataset. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7901, Singapore, 2023. Association for Computational Linguistics.

[22] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[23] Zhihong Chen, Ruifei Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Advancing visual grounding with scene knowledge: Benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15039–15049, 2023.

[24] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng

Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites, 2024.

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

[26] Mary Ann Clark, Jung Choi, and Matthew Douglas. *Biology*. OpenStax, Rice University, 2nd edition, 2018. Available at https://openstax.org/details/books/biology-2e.

[27] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[28] Jonathan Clayden, Nick Greeves, and Stuart Warren. *Organic chemistry*. Oxford University Press, USA, 2012.

[29] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Tvbench: Redesigning video-language evaluation, 2024.

[30] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022.

[31] Braja M. Das. *Principles of Geotechnical Engineering*. Cengage Learning, 9th edition, 2017.

[32] Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models, 2024.

[33] Mackenzie L. Davis and David A. Cornwell. *Introduction to Environmental Engineering*. McGraw-Hill Education, 5th edition, 2012.

[34] Andong Deng, Taojiannan Yang, and Chen Chen. A large-scale study of spatiotemporal representation learning with a new benchmark on action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20519–20531, 2023.

[35] Chunyuan Deng, Yilun Zhao, Yuzhao Heng, Yitong Li, Jiannan Cao, Xiangru Tang, and Arman Cohan. Unveiling the spectrum of data contamination in language model: A survey from detection to remediation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16078–16092, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[36] Avi Domb, Boaz Mizrahi, and Shady Farah. *Biomaterials and Biopolymers*. Springer, 2023.

[37] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, More, and Zhiwei Zhao. The llama 3 herd of models, 2024.

[38] John D. Enderle and Joseph D. Bronzino. *Introduction to Biomedical Engineering*. Academic Press, 4th edition, 2017.

[39] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding, 2024.

[40] Adam Feather, David Randall, and Mona Waterhouse. *Kumar and Clark's Clinical Medicine E-Book: Kumar and Clark's Clinical Medicine E-Book*. Elsevier Health Sciences, 2020.

[41] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos, 2024.

[42] Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. Sciknoweval: Evaluating multi-level scientific knowledge of large language models, 2024.

[43] Harry L Field and John M Long. *Introduction to agricultural engineering technology: a problem solving approach*. Springer, 2018.

[44] Paul Flowers, Klaus Theopold, Richard Langley, and William R. Robinson. *Chemistry*. OpenStax, Rice University, 2nd edition, 2019. Available at https://openstax.org/details/books/chemistry-2e.

[45] Erin H Fouberg and Alexander B Murphy. *Human Geography: People, Place, and Culture*. John Wiley & Sons, 2020.

[46] Fabrizio Frigeni. *Industrial Robotics Control: Mathematical Models, Software Architecture, and Electronics Design*. Springer, 2022.

[47] Victoria Fromkin, Robert Rodman, and Nina Hyams. *An Introduction to Language*. Cengage Learning, 11th edition, 2017.

[48] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024.

[49] Shaikat Galib, Shanshan Wang, Guanshuo Xu, Pascal Pfeiffer, Ryan Chesler, Mark Landry, and Sri Satish Ambati. H2ovl-mississippi vision language models technical report, 2024.

[50] Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2024.

[51] Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

[52] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[53] Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[54] Nikolai V. Gorbunov. *Tissue Barriers in Disease, Injury and Regeneration*. Elsevier, 1st edition, 2022.

[55] Steven A. Greenlaw, David Shapiro, and Daniel MacDonald. *Principles of Economics*. OpenStax, Rice University, 3rd edition, 2023. Available at https://openstax.org/details/books/principles-economics-3e.

[56] David J Griffiths. *Introduction to electrodynamics*. Cambridge University Press, 2023.

[57] Allan R Hambley. *Electrical Engineering: Principles and Applications*. Pearson London, UK, 2018.

[58] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, Kevin Lin, William Yang Wang, Lijuan Wang, and Xin Eric Wang. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos, 2024.

[59] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015.

[60] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

[61] Darrel Hess and Tom L. McKnight. *McKnight's Physical Geography: A Landscape Appreciation*. Pearson, 13th edition, 2021.

[62] HLTCOE@JHU. Turkle: A web-based tool for managing annotation tasks. `https://github.com/hltcoe/turkle`, 2024. Accessed: 2024-11-01.

[63] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

[64] Paul Horowitz and Winfield Hill. The art of electronics, 2015.

[65] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models, 2023.

[66] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, 2024.

[67] Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, and Pengfei Liu. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai, 2024.

[68] Peter Huber and Alastair Mullis. *The CISG: A new textbook for students and practitioners*. Sellier de Gruyter, 2009.

[69] Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5084, Singapore, 2023. Association for Computational Linguistics.

[70] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017.

[71] Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023.

[72] Zhihuan Jiang, Zhen Yang, Jinhao Chen, Zhengxiao Du, Weihan Wang, Bin Xu, Yuxiao Dong, and Jie Tang. Visscience: An extensive benchmark for evaluating k12 educational multi-modal scientific reasoning, 2024.

[73] Eric R. Kandel, James H. Schwartz, Thomas M. Jessell, Steven A. Siegelbaum, and A.J. Hudspeth. *Principles of Neural Science*. McGraw-Hill Education, 6th edition, 2021.

[74] Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and Erkut Erdem. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models, 2023.

[75] Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, and Erkut

Erdem. ViLMA: A zero-shot benchmark for linguistic and temporal grounding in video-language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[76] Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Jameel Hassan, Muzammal Naseer, Federico Tombari, Fahad Shahbaz Khan, and Salman Khan. How good is my video lmm? complex video reasoning and robustness evaluation suite for video-lmms, 2024.

[77] Richard R. Kibbe, Roland O. Meyer, John E. Neely, and Warran T. White. *Machine Tool Practices*. Pearson, 11th edition, 2019.

[78] David R. Klein. *Organic Chemistry as a Second Language: First Semester Topics*. John Wiley & Sons, 2024.

[79] Fred S. Kleiner. *Art Through the Ages: A Global History, Volume I*. Cengage Learning, 16th edition, 2020.

[80] Ann Kordas, Ryan J. Lynch, et al. *World History Volume 1*. OpenStax, Rice University, 2022. Available at https://openstax.org/details/books/world-history-volume-1.

[81] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017.

[82] Vinay Kumar, Abul K. Abbas, and Jon C. Aster. *Robbins and Cotran Pathologic Basis of Disease*. Elsevier, 10th edition, 2020.

[83] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[84] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024.

[85] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium, 2018. Association for Computational Linguistics.

[86] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024.

[87] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, Chongyan Zhu, Xiaoyi Ren, Chao Li, Yifan Ye, Peng Liu, Lihuan Zhang, Hanshu Yan, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model, 2025.

[88] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

[89] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024.

[90] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[91] Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models, 2024.

[92] Xinhao Li, Zhenpeng Huang, Jing Wang, Kunchang Li, and Limin Wang. Videoeval: Comprehensive benchmark suite for low-cost evaluation of video foundation model, 2024.

[93] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024.

[94] Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, Linda Ruth Petzold, Stephen D. Wilson, Woosang Lim, and William Yang Wang. Mmsci: A dataset for graduate-level multi-discipline multimodal scientific understanding, 2024.

[95] Zhenwen Liang, Kehan Guo, Gang Liu, Taicheng Guo, Yujun Zhou, Tianyu Yang, Jiajun Jiao, Renjie Pi, Jipeng Zhang, and Xiangliang Zhang. Scemqa: A scientific college entrance level multimodal question answering benchmark, 2024.

[96] Samuel J. Ling, Jeff Sanny, and William Moebs. *University Physics Volume 1*. OpenStax, Rice University, 2016. Available at https://openstax.org/details/books/university-physics-volume-1.

[97] Samuel J. Ling, Jeff Sanny, and William Moebs. *University Physics Volume 2*. OpenStax, Rice University, 2016. Available at https://openstax.org/details/books/university-physics-volume-2.

[98] Samuel J. Ling, Jeff Sanny, and William Moebs. *University Physics Volume 3*. OpenStax, Rice University, 2016. Available at https://openstax.org/details/books/university-physics-volume-3.

[99] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[100] Jiaying Liu, Sijie Song, Chunhui Liu, Yanghao Li, and Yueyu Hu. A benchmark dataset and comparison study for multi-modal human action analytics. *ACM Trans. Multimedia Comput. Commun. Appl.*, 16(2), 2020.

[101] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[102] William Lowrie and Andreas Fichtner. *Fundamentals of geophysics*. Cambridge university press, 2020.

[103] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, pages 2507–2521. Curran Associates, Inc., 2022.

[104] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024.

[105] Liqun Luo. *Principles of neurobiology*. Garland Science, 2020.

[106] Marina MacKay. *The Cambridge introduction to the novel*. Cambridge University Press, 2010.

[107] Upamanyu Madhow. *Introduction to communication systems*. Cambridge University Press, 2014.

[108] PK Mallick. Fiber-reinforced composites: Materials, manufacturing, and design, 2007.

[109] Gregory N. Mankiw. *Principles of Microeconomics*. Cengage Learning, 9th edition, 2020.

[110] Kenneth Fuller Maxcy, Milton Joseph Rosenau, John M Last, Robert B Wallace, Neal Kohatsu, and Ross Brownson. *Maxcy-Rosenau-Last public health & preventive medicine*. McGraw-Hill, 2008.

[111] MistralAI. Announcing pixtral 12b, 2024.

[112] Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, Mikhail Sirotenko, Yukun Zhu, and Tobias Weyand. Neptune: The long orbit to benchmarking long video understanding. 2024.

[113] Jill Nelmes, editor. *Introduction to Film Studies*. Routledge, 5th edition, 2012.

[114] Donald A Nield and Adrian Bejan. *Convection in Porous Media*. Springer, 2017.

[115] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models, 2023.

[116] Katsuhiko Ogata. *Modern Control Engineering*. Prentice Hall, 5th edition, 2010.

[117] OpenAI. Openai o1 system card. 2024.

[118] OpenAI. Hello gpt-4o, 2024.

[119] Judith A. Owen, Jenni Punt, and Sharon A. Stranford. *Kuby Immunology*. W.H. Freeman, 8th edition, 2018.

[120] David A. Patterson and John L. Hennessy. *Computer organization and design: The hardware/software interface*. Elsevier, 6th edition, 2022.

[121] Onno Rudolf Pols. *Stellar structure and evolution*. Astronomical Institute Utrecht NY, 2011.

[122] Dale Purves, GJ Augustine, David Fitzpatrick, WC Hall, AS LaMantia, RD Mooney, ML Platt, and LE White. Neuroscience (sixth edit), 2018.

[123] C Gonzalez Rafael and E Woods Richard. *Digital Image Processing*. Pearson Education, 2018.

[124] Colin Renfrew and Paul Bahn. *Archaeology: Theories, Methods, and Practice*. Thames & Hudson, 7th edition, 2016.

[125] Robert E. Ricklefs. *The Economy of Nature*. W.H. Freeman, 7th edition, 2013.

[126] Barbara Ryden and Bradley M Peterson. *Foundations of astrophysics*. Cambridge University Press, 2020.

[127] Daniel V. Schroeder. *An introduction to thermal physics*. Oxford University Press, 2020.

[128] Robert Sedgewick and Kevin Wayne. Algorithms (4th edn). *Google Scholar Google Scholar Digital Library Digital Library*, 2011.

[129] Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models, 2024.

[130] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016.

[131] Abraham Silberschatz, Peter B. Galvin, and Greg Gagne. *Operating System Concepts*. John Wiley & Sons, 10th edition, 2018.

[132] Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19053–19061, 2024.

[133] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada, 2023. Association for Computational Linguistics.

[134] Rikito Takahashi, Hirokazu Kiyomaru, Chenhui Chu, and Sadao Kurohashi. Abstractive multi-video captioning: Benchmark dataset construction and extensive evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 57–69, Torino, Italia, 2024. ELRA and ICCL.

[135] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023.

[136] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao,

Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023.

[137] Chris Turner. *Contract law*. Routledge, 2013.

[138] Ray Turner. *Arbitration awards: a practical approach*. John Wiley & Sons, 2008.

[139] G Cornelis Van Kooten. *Land resource economics and sustainable development: economic policies and the common good*. UBC Press, 2011.

[140] Hal R. Varian. *Intermediate Microeconomics: A Modern Approach*. W.W. Norton & Company, 8th edition, 2010.

[141] William R Wagner, Shelly E Sakiyama-Elbert, Guigen Zhang, and Michael J Yaszemski. *Biomaterials Science: An Introduction to Materials in Medicine*. Elsevier, 2020.

[142] Jinfeng Wang. *Intelligent Manufacturing System and Intelligent Workshop*. Springer.

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

[144] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2024.

[145] Yuxuan Wang, Difei Gao, Licheng Yu, Weixian Lei, Matt Feiszli, and Mike Zheng Shou. Geb+: A benchmark for generic event boundary captioning, grounding and retrieval. In *European Conference on Computer Vision*, pages 709–725. Springer, 2022.

[146] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding, 2024.

[147] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024.

[148] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling, 2025.

[149] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms, 2024.

[150] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

[151] Edward J. Wing and Fred J. Schiffman. *Cecil Essentials of Medicine*. Elsevier, 10th edition, 2021.

[152] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics.

[153] Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[154] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024.

[155] xAI. Grok-2 beta release, 2024.

[156] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, 2021.

[157] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[158] John A Yagiela, Frank J Dowd, Bart Johnson, Angelo Mariotti, and Enid A Neidle. *Pharmacology and Therapeutics for Dentistry-E-Book: Pharmacology and Therapeutics for Dentistry-E-Book*. Elsevier Health Sciences, 2010.

[159] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren

Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024.

[160] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces, 2024.

[161] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567, 2024.

[162] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2024.

[163] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025.

[164] Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, Haoran Zhang, Xingwei Qu, Junjie Wang, Ruibin Yuan, Yizhi Li, Zekun Wang, Yudong Liu, Yu-Hsuan Tsai, Fengji Zhang, Chenghua Lin, Wenhao Huang, and Jie Fu. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark, 2024.

[165] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore, 2023. Association for Computational Linguistics.

[166] Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. Movqa: A benchmark of versatile question-answering for long-form movie understanding, 2023.

[167] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024.

[168] Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. Financemath: Knowledge-intensive math reasoning in finance domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12841–12858, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[169] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico, 2024. Association for Computational Linguistics.

Appendix Contents

A. MMVU Preliminary Setup

A.1. Annotator Biography

The detailed biographies of the annotators involved in MMVU construction are presented in Table 4. All annotators are from universities ranked in the Top 500 of the 2024 QS Global Rankings[2] and are fluent in English.

ID	Year	Major	Assigned Subject(s)	Author?	Validator?
1	1st year Master	Biomedical Engineering	Biomedical Engineering Computer Science Electrical Engineering	✗	✗
2	1st year Master	Bioinformatics	Biomedical Engineering	✗	✗
3	1st year Master	Biological Engineering	Biomedical Engineering	✗	✗
4	2nd year Master	Biomedical Engineering	Biomedical Engineering Electronics and Communication	✗	✗
5	5th year PhD	Agricultural and Biosystems Engineering	Biomedical Engineering	✗	✗
6	2nd year Master	Architecture	Civil Engineering	✗	✗
7	3rd year PhD	Civil Engineering	Civil Engineering Mechanical Engineering	✗	✗
8	–	–	–	✓	✓
9	3rd year Undergraduate	Electrical Engineering	Computer Science Electrical Engineering	✗	✗
10	2nd year Master	Electrical Engineering	Computer Science Electronics and Communication	✗	✗
11	2nd year Master	Electrical Engineering	Computer Science Mechanical Engineering	✗	✗
12	3rd year Undergraduate	Software Engineering	Computer Science	✗	✗
13	2nd year Master	Computer Science	Computer Science	✗	✗
14	–	–	–	✓	✗
15	1st year PhD	Electrical Engineering	Electrical Engineering Computer Science Electronics and Communication	✗	✗
16	1st year PhD	Electrical Engineering	Electrical Engineering	✗	✗
17	–	–	–	✓	✓
18	1st year Master	Electrical Engineering	Electrical Engineering Mechanical Engineering	✗	✗
19	1st year PhD	Electrical Engineering	Electronics and Communication	✗	✗
20	3rd year PhD	Food Science	Mechanics	✗	✗
21	4th year PhD	Materials Science	Materials Science	✗	✗
22	4th year Undergraduate	Aerospace Engineering	Materials Science	✗	✗
23	4th year Undergraduate	Mechanical Engineering	Mechanical Engineering Materials Science Mechanical Engineering	✗	✓
24	2nd year PhD	Mechanical Engineering	Mechanical Engineering	✗	✗
25	1st year PhD	Mechanical Engineering	Mechanical Engineering	✗	✗
26	1st year Master	Medicine	Basic Medicine Clinical Medicine	✗	✗
27	1st year Master	Radiology	Basic Medicine Clinical Medicine	✗	✗
28	1st year Master	Dentistry	Basic Medicine Dentistry	✗	✗
29	1st year PhD	Nursing	Basic Medicine Pharmacy	✗	✗
30	3rd year Undergraduate	Epidemiology	Basic Medicine Preventive Medicine	✗	✗

Table 4. Biographies of 73 annotators and validators involved in MMVU construction (Author biographies are hidden to protect identity confidentiality). Note that some authors only participated in the data validation stage.

[2] https://www.topuniversities.com/world-university-rankings

ID	Year	Major	Assigned Subject(s)	Author?	Validator?
31	3rd year Undergraduate	Medicine	Clinical Medicine	✗	✗
32	–	–	–	✓	✓
33	2nd year PhD	Medicine	Clinical Medicine Pharmacy	✗	✗
34	4th year PhD	Dentistry	Dentistry	✗	✗
35	3rd year Undergraduate	Dentistry	Dentistry	✗	✗
36	4th year PhD	Dentistry	Dentistry	✗	✗
37	1st year PhD	Public Health	Pharmacy Preventive Medicine	✗	✗
38	4th year Undergraduate	Pharmacy	Pharmacy	✗	✗
39	3rd year PhD	East Asian Studies	Art	✗	✗
40	4th year PhD	Literature	Art History Literature	✗	✗
41	–	–	– History	✓	✗
42	1st year PhD	Economics	Economics	✗	✗
43	4th year Undergraduate	Accounting	Economics Law	✗	✗
44	4th year PhD	Finance	Economics	✗	✗
45	3rd year PhD	Public Administration	Law Management	✗	✗
46	1st year Master	Literature	Literature	✗	✗
47	5th year PhD	Linguistics	Literature	✗	✗
48	3rd year Undergraduate	Public Administration	Management	✗	✗
49	5th year PhD	Astronomy	Astronomy	✗	✗
50	–	–	–	✓	✓
51	2nd year Master	Astronomy	Astronomy	✗	✗
52	–	–	– Geography	✓	✗
53	3rd year PhD	Biology	Biology	✗	✗
54	1st year PhD	Biology	Biology Neurobiology	✗	✗
55	3rd year PhD	Marine Biology	Biology Chemistry	✗	✗
56	–	–	–	✓	✗
57	1st year PhD	Chemistry	Chemistry	✗	✗
58	3rd year Undergraduate	Chemistry	Chemistry	✗	✗
59	1st year PhD	Physics	Electromagnetism	✗	✗
60	4th year Undergraduate	Physics	Electromagnetism Thermodynamics	✗	✗
61	4th year PhD	Physics	Electromagnetism	✗	✗
62	1st year PhD	Physics	Electromagnetism Mechanics Thermodynamics	✗	✗
63	1st year Master	Physics	Thermodynamics Electromagnetism	✗	✗
64	3rd year Undergraduate	Agricultural and Environmental Sciences	Geography	✗	✗
65	4th year PhD	Physics	Thermodynamics Mechanics Modern Physics	✗	✗
66	1st year PhD	Physics	Mechanics	✗	✗
67	3rd year PhD	Physics	Mechanics	✗	✗
68	4th year PhD	Physics	Modern Physics	✗	✗
69	3rd year Undergraduate	Neurobiology	Neurobiology	✗	✗
70	1st year PhD	Neurobiology	Neurobiology	✗	✗
71	–	–	–	✓	✓
72	3rd year Undergraduate	Biology	Neurobiology	✗	✗
73	1st year Master	Biology	Neurobiology	✗	✗

Table 5. Biographies of 73 annotators and validators involved in MMVU construction (Author biographies are hidden to protect identity confidentiality). Note that some authors only participated in the data validation stage.

A.2. Textbook for Each Subject

As discussed in Section 3.2, we design a textbook-guided example annotation pipeline to encompass both the *breadth of knowledge* and the *depth of reasoning*. The textbooks used for each subject are detailed in the following tables. They are selected by expert annotators and are recognized as authoritative references in their respective fields.

Subject	Textbook
Astronomy	1. *Foundations of Astrophysics* [126] 2. *Stellar Structure And Evolution* [121]
Biology	1. *Biology, 2nd Edition* [26] 2. *Introduction to Agricultural Engineering Technology: A Problem Solving Approach, 4th Edition* [43] 3. *Introduction to Environmental Engineering, 5th Edition* [33] 4. *The Economy of Nature, 7th Edition* [125] 5. *The Molecular Biology of the Cell, 6th Edition* [2]
Chemistry	1. *Atkins' Physical Chemistry, 12th Edition* [8] 2. *Chemistry, 2nd Edition* [44] 3. *Chemistry: The Central Science, 15th Edition* [14] 4. *Organic Chemistry As A Second Language* [78] 5. *Organic Chemistry, 2nd Edition* [28]
Electromagnetism	1. *Introduction to Electrodynamics, 4th Edition* [56] 2. *University Physics Volume 2 (Electromagnetism)* [97]
Geography	1. *Fundamentals of Geophysics, 2nd Edition* [102] 2. *Human Geography, 12th Edition* [45] 3. *Physical Geography: A Landscape Appreciation, 10th Edition* [61]
Mechanics	1. *University Physics Volume 1* [96]
Modern Physics	1. *University Physics Volume 3* [98]
Neurobiology	1. *Neuroscience, 6th Edition* [122] 2. *Principles of Neural Science, 6th Edition* [73] 3. *Principles of Neurobiology* [105]
Thermodynamics	1. *An Introduction to Thermal Physics* [127] 2. *University Physics Volume 2 (Thermodynamics)* [97]

Table 6. List of textbooks and corresponding example numbers for the **Science** discipline.

Subject	Textbook
Biomedical Engineering	1. *Biomaterials Science: An Introduction to Materials in Medicine, 4th Edition* [141] 2. *Biomaterials and Biopolymers* [36] 3. *Fundamentals and Advances in Medical Biotechnology* [5] 4. *Introduction to Biomedical Engineering, 4th Edition* [38]
Civil Engineering	1. *Engineering Geology and Construction* [11] 2. *Principles of Geotechnical Engineering, 9th Edition* [31] 3. *Structure for Architects: A Case Study in Steel, Wood, and Reinforced Concrete Design* [10]
Computer Science	1. *Algorithms, 4th Edition* [128] 2. *Computer Organization and Design: The Hardware/Software Interface, 6th Edition* [120] 3. *Computer Systems: A Programmer's Perspective, 3rd Edition* [16] 4. *Deep Learning* [52] 5. *Digital Image Processing, 4th Edition* [123] 6. *Introduction to Algorithms, 4th Edition* [30] 7. *Operating System Concepts, 10th Edition* [131]
Electrical Engineering	1. *Electrical Engineering: Principles and Applications, 7th Edition* [57]
Electronics and Communication	1. *CMOS Analog Circuit Design, 3rd Edition* [3] 2. *Introduction to Communication Systems* [107] 3. *The Art of Electronics, 3rd Edition* [64]
Materials Science	1. *Composite Materials: Science and Engineering, 3rd Edition* [19] 2. *Convection in Porous Media, 5th Edition* [114] 3. *Fiber-Reinforced Composites Materials, Manufacturing, and Design, 3rd Edition* [108] 4. *Materials Science and Engineering: An Introduction, 10th Edition* [18]
Mechanical Engineering	1. *Industrial Automation: An Engineering Approach* 2. *Industrial Robotics Control: Mathematical Models, Software Architecture, and Electronics Design* [46] 3. *Intelligent Manufacturing System and Intelligent Workshop* [142] 4. *Machine Tool Practices, 11th Edition* [77] 5. *Marks' Standard Handbook for Mechanical Engineers, 12th Edition* [9] 6. *Modern Control Engineering, 5th Edition* [116]

Table 7. List of textbooks and corresponding example numbers for the **Engineering** discipline.

Subject	Textbook
Basic Medicine	1. *Kuby Immunology, 8th Edition* [119] 2. *Robbins and Cotran Pathologic Basis of Disease, 10th Edition* [82] 3. *Tissue Barriers in Disease, Injury and Regeneration* [54]
Clinical Medicine	1. *Cecil Essentials of Medicine, 10th Edition* [151] 2. *Kumar and Clark's Clinical Medicine, 10th Edition* [40]
Dentistry	1. *Pharmacology and Therapeutics for Dentistry, 7th Edition* [158]
Pharmacy	1. *The Pharmacological Basis of Therapeutics, 13th Edition* [15]
Preventive Medicine	1. *Public Health and Preventive Medicine, 15th Edition* [110]

Table 8. List of textbooks and corresponding example numbers for the **Healthcare** discipline.

Subject	Textbook
Art	1. *Art Through the Ages: A Global History Volume I, 16th Edition* [79] 2. *Introduction to Film Studies, 5th Edition* [113] 3. *The Filmmaker's Handbook: A Comprehensive Guide for the Digital Age, 5th Edition* [6]
Economics	1. *Intermediate Microeconomics: A Modern Approach, 8th Edition* [140] 2. *Land Resource Economics and Sustainable Development: Economic Policies and the Common Good* [139] 3. *Macroeconomics, 9th Edition* [12] 4. *Principles of Economics, 3rd Edition* [55] 5. *Principles of Microeconomics, 9th Edition* [109]
History	1. *Archaeology: Theories Methods and Practice, 7th Edition* [124] 2. *World History Volume 1: to 1500* [80]
Law	1. *Arbitration Awards: A Practical Approach* [138] 2. *Contract Law* [137] 3. *The CISG: A new textbook for students and practitioners* [68]
Literature	1. *An Introduction to Language, 11th Edition* [47] 2. *The Cambridge Introduction to the Novel* [106]
Management	1. *Principles of Management* [13]

Table 9. List of textbooks and corresponding example numbers for the **Humanities and Social Science** discipline.

A.3. Annotation Guideline and Interface

With the goal of ensure the high quality of data, MMVU adheres to the following four benchmark construction desiderata, we develop the following annotation interface based on Turkle [62], an open-source clone of Amazon's Mechanical Turk:

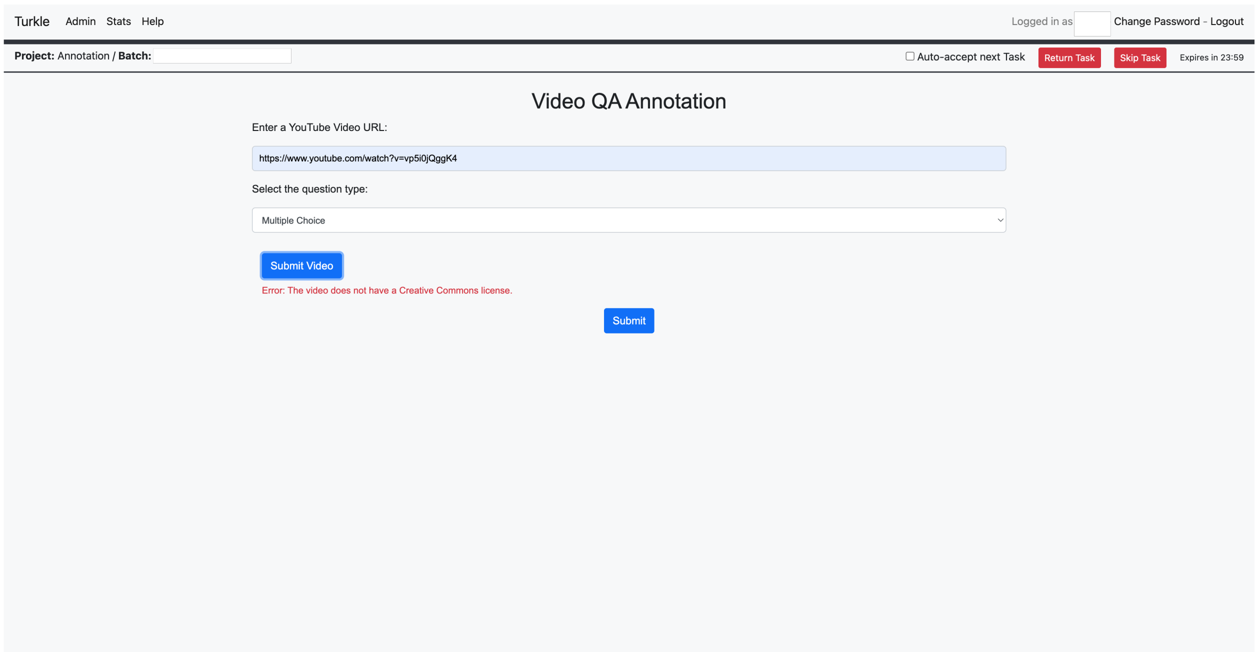


Figure 5. **Annotation Interface - Step 1: Video Collection**. In this step, annotators are required to input the YouTube video URL and select the desired question type. The backend system of the interface will automatically verify whether the provided YouTube video is under a Creative Commons license using the YouTube Data API v3. If the video does not meet this requirement, as shown in the figure, a warning message will be displayed, and the submission will be blocked. Once a valid example is submitted, the annotation interface will proceed to Step 2, which is illustrated in the following two figures.

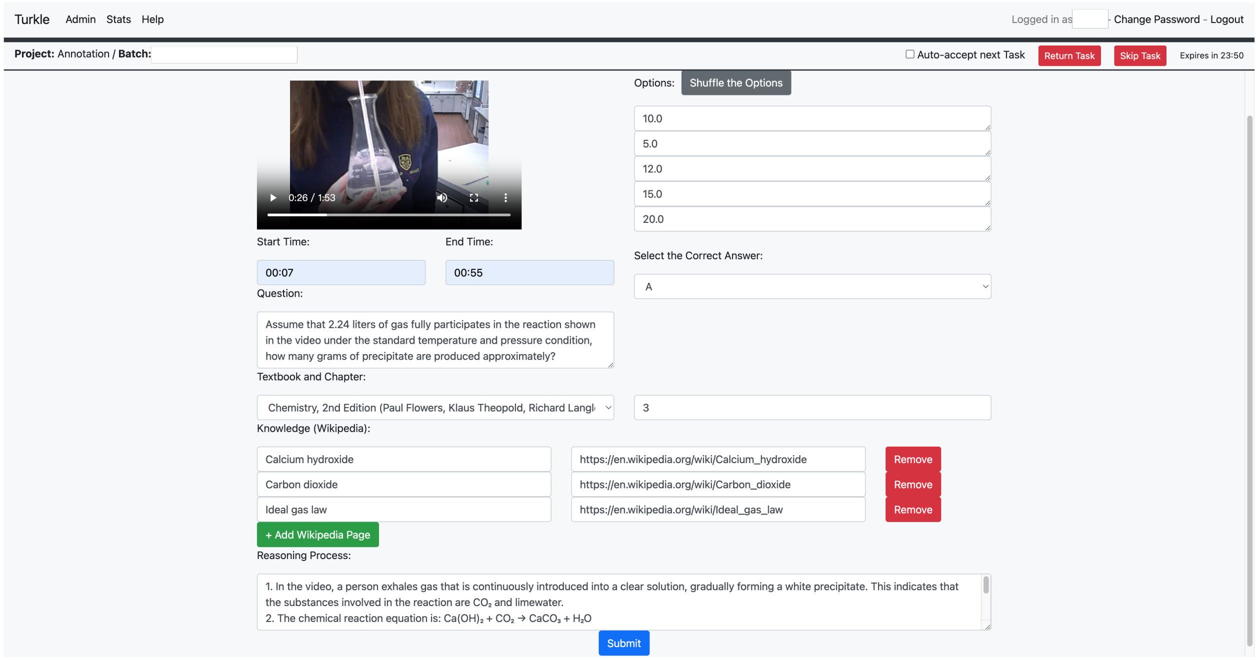


Figure 6. Annotation Interface - Step 2: Multiple-choice Question Annotation.

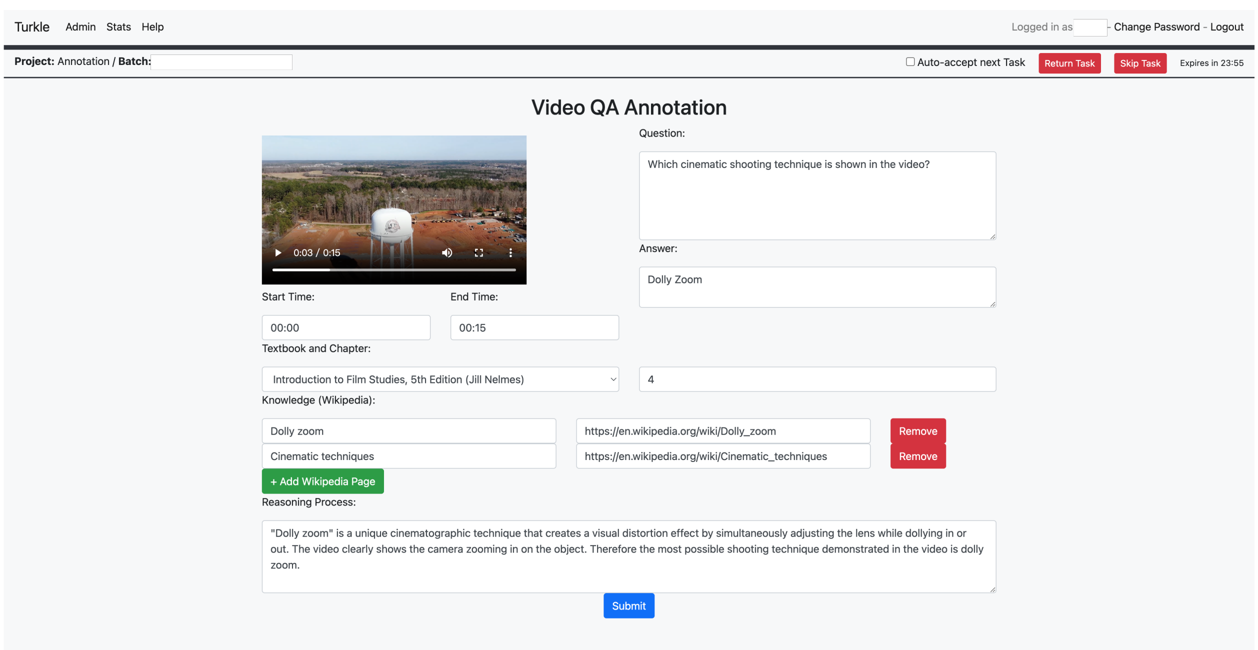


Figure 7. Annotation Interface - Step 2: Open-ended Question Annotation.

A.4. Validation Guideline and Interface

To ensure that the final dataset remains high-quality and meets expert-level standards without introducing unnecessary bias, each example in MMVU undergoes expert review by one of the authors or top-performing annotators to verify the accuracy of its annotations, following the annotation guideline detailed in Appendix A.3. The examples of validation interface are presented as follows:

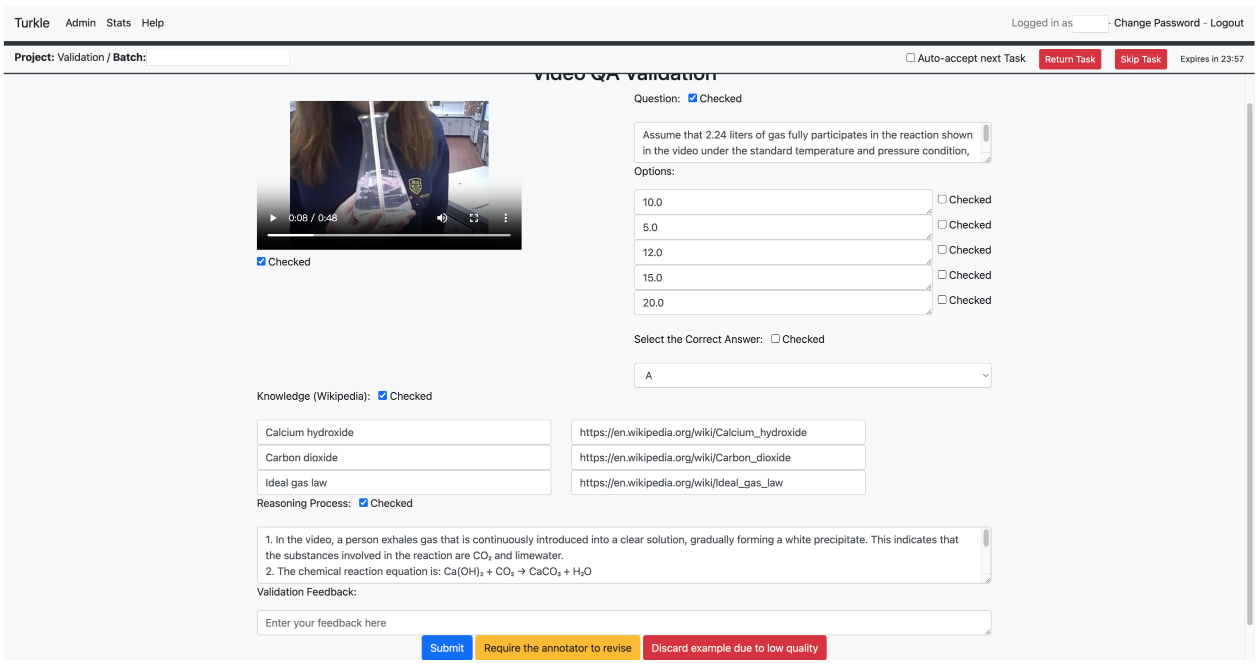


Figure 8. **Validation Interface.** Human validators are required to thoroughly review each annotation feature to ensure alignment with benchmark construction criteria and annotation guidelines. If revisions are not feasible, detailed feedback must be provided to the original annotator, who will then revise and resubmit the annotation for a second review. Additionally, validators may discard examples deemed to be of low quality and unlikely to meet the desired criteria through revision.

A.5. Data Annotation and Validation Payment

The annotation and validation process for MMVU spans three months. As outlined in Section 3.2, annotating examples for MMVU can be particularly time-intensive, especially when there is limited availability of videos with Creative Commons licenses in the required subjects. To accommodate this and ensure a high-quality dataset, we compensate annotators based on the time they spend rather than the number of examples completed, preventing them from rushing through tasks. Annotators are required to record their screens throughout the annotation process, which enables us to verify time reporting accuracy and maintain productivity standards. This also helps us identify any distractions and precisely track the total time spent on each task. We offer a *base rate* of 6 USD per hour for both annotation and validation work, with an additional 2 USD per completed annotation and 0.40 USD per validated example. On average, annotating a single question for MMVU takes 20 minutes and 17 seconds, while validation requires 4 minutes and 12 seconds. This compensation structure ensures that annotators earn wages that are competitive with the average payment for teaching assistants at their respective universities. To reduce pressure and maintain a comfortable pace, we recommended that annotators limit their work to a maximum of 10 QA example annotations or 50 QA example validations per day.

Back to Appendix Table of Contents

B. Experiment Setup

B.1. Configuration of Evaluated Models

Table 10 detail the configuration of each evaluated models. We use the default settings from the official implementation of each model to process vision input. Across all experiments, the temperature is set to 1.0, with a maximum output length of 1024 tokens. However, for Gemini-2-Flash-Thinking, the maximum output length is set as 8192 tokens to accommodate its long CoT reasoning mechanism. All inferences are reproducible on a workstation equipped with two NVIDIA A100-80G GPUs.

Organization	Model	Release	Version	Support Video?	Input Frames	# Inference Pipeline
			Proprietary Models			
OpenAI	o1*	2024-12	`o1-2024-12-17`	✗	32	API
	GPT-4o	2024-8	`gpt-4o-2024-08-06`	✗	32	
	GPT-4o-mini	2024-7	`gpt-4o-mini-2024-07-18`	✗	32	
Google	Gemini 2.0 Flash Thinking	2024-12	`gemini-2.0-flash-thinking-exp-1219`	✗	32	API
	Gemini 2.0 Flash	2024-12	`gemini-2.0-flash-exp`	✗	32	
	Gemini 1.5 Pro	2024-9	`gemini-1.5-pro`	✓	32	
	Gemini 1.5 Flash	2024-9	`gemini-1.5-flash`	✓	32	
Anthropic	Claude-3.5-Sonnet	2024-10	`claude-3-5-sonnet-20241022`	✗	32	API
xAI	Grok-2-Vision	2024-12	`grok-2-vision-1212`	✗	32	API
Zhipu AI	GLM-4V-Plus	2025-1	`glm-4v-plus-0111`	✓	4	API
			Open-source Multimodal Foundation Models			
Mistral AI	Pixtral-12B	2024-9	`Pixtral-12B-2409`	✗	8	vLLM
Microsoft	Phi-3.5-Vision	2024-7	`Phi-3.5-vision-instruct`	✗	16	vLLM
Shanghai AI Lab	InternVL2.5-38B	2024-11	`InternVL2.5-38B`	✗	4	vLLM
	InternVL2.5-8B	2024-11	`InternVL2.5-8B`	✗	4	
	InternVL2-8B	2024-6	`InternVL2-8B`	✗	4	
Alibaba	Qwen2-VL-2B	2024-8	`Qwen2-VL-2B-Instruct`	✓	1fps	vLLM
	Qwen2-VL-7B	2024-8	`Qwen2-VL-7B-Instruct`	✓	1fps	
	Qwen2-VL-72B	2024-9	`Qwen2-VL-72B-Instruct`	✓	1fps	
Meta	Llama-3.2-11B-Vision	2024-9	`Llama-3.2-11B-Vision-Instruct`	✗	8	vLLM
	Llama-3.2-90B-Vision	2024-9	`Llama-3.2-90B-Vision-Instruct`	✗	8	
DAMO	VideoLLaMA2-7B	2024-6	`VideoLLaMA2-7B`	✓	1fps	HF
	VideoLLaMA2.1-7B	2024-10	`VideoLLaMA2.1-7B-16F`	✓	1fps	HF
DeepSeek	DeepSeek-VL2	2024-12	`deepseek-vl2`	✗	2	vLLM
	DeepSeek-VL2-Small	2024-12	`deepseek-vl2-small`	✗	2	vLLM
	DeepSeek-VL2-Tiny	2024-12	`deepseek-vl2-tiny`	✗	2	vLLM
Rhymes	Aria	2024-11	`Aria-Chat`	✗	8	vLLM
Llava Hugging Face	LLaVA-OneVision-7B	2024-9	`llava-onevision-qwen2-7b-ov-chat-hf`	✓	1fps	vLLM
	LLaVA-NeXT-Video-34B	2024-6	`LLaVA-NeXT-Video-34B-hf`	✗	8	vLLM
	LLaVA-NeXT-Video-7B	2024-6	`LLaVA-NeXT-Video-7B-hf`	✓	16	vLLM
HuggingFaceM4	Idefics3-8B	2024-8	`Idefics3-8B-Llama3`	✗	4	vLLM
OpenGVLab	InternVideo2-8B	2024-8	`InternVideo2-Chat-8B`	✓	1fps	HF
H2O	H2OVL Mississippi-2B	2024-10	`h2ovl-mississippi-2b`	✗	4	vLLM

Table 10. Details of the multimodal foundation models evaluated in MMVU. The "Source" column includes URLs for proprietary models and Hugging Face model names for open-source models. The "# Input Frames" column, for those models only support multi-image input, represents the default number of input frames, chosen from 2, 4, 8, 16, 32, based on the maximum value that does not exceed the model's context window. "HF" means "Hugging Face".

B.2. Chain-of-Thought and Direct Answer Prompts

The following figures illustrates the CoT reasoning and Direct Answer prompts applied in this study for answering multiple-choice and open-ended questions, respectively.

```
Question:{question}
A: {option_a}
B: {option_b}
C: {option_c}
D: {option_d}
E: {option_e}

Visual Information: {processed_video}

Answer the given multiple-choice question step by step. Begin by explaining your reasoning process clearly. Conclude by stating the final answer using the following format: “Therefore, the final answer is: $LETTER” (without quotes), where $LETTER is one of the options. Think step by step before answering.
```

Figure 9. CoT reasoning prompt, adopted from MMMU-Pro [162], for answering multiple-choice question.

```
Question:{question}

Visual Information: {processed_video}

Answer the given question step by step. Begin by explaining your reasoning process clearly. Conclude by stating the final answer using the following format: ’Therefore, the final answer is: “Answer: $ANSWER” (without quotes), where $ANSWER is the final answer of the question. Think step by step before answering.
```

Figure 10. CoT reasoning prompt for answering open-ended question.

```
Question:{question}
A: {option_a}
B: {option_b}
C: {option_c}
D: {option_d}
E: {option_e}

Visual Information: {processed_video}

Do not generate any intermediate reasoning process. Answer directly with the option letter from the given choices.
```

Figure 11. Direct Answer prompt, adopted from MMMU-Pro [162], for answering multiple-choice question.

```
Question:{question}

Visual Information: {processed_video}

Do not generate any intermediate reasoning process. Directly output the final answer.
```

Figure 12. Direct Answer prompt for answering open-ended question.

B.3. Prompts for Accuracy Evaluation

[Instruction]
Evaluate whether the model's final answer is correct by comparing it to the ground-truth answer provided for the given question.
You should first extract the final answer from the model's response, and then compare the extracted answer with the ground-truth answer to determine its accuracy. Output your response in the following structured format:
{
    'extracted_answer': // str value "A" "B" "C" "D" "E", should be a single character
    'correct': // boolean value, True if the answer is correct, False otherwise
}

[User]
Question:{question}
A: {option_a}
B: {option_b}
C: {option_c}
D: {option_d}
E: {option_e}

Ground Truth Answer: {ground_truth}

Model Response to the Question: {model_response}

Figure 13. Evaluation prompt used for assessing the accuracy of multi-choice QA.

[Instruction]
Evaluate whether the model's final answer is correct by comparing it to the ground-truth answer provided for the given question. You should first extract the final answer from the model's response, and then compare the extracted answer with the ground-truth answer to determine its accuracy. The final answer generated by the model does not need to match the ground-truth answer word-for-word. However, it should only be considered correct if it demonstrates the exact same technique or concept explicitly and unambiguously equivalent to the ground-truth answer. Output your response in the following structured format:
{
    'extracted_answer': // str value, the short final answer extracted from the model's response, do not hallucinate one that is not present in the response
    'correct': // boolean value, True if the answer is correct, False otherwise
}

[User]
Question:{question}

Ground Truth Answer: {ground_truth}

Model Response to the Question: {model_response}

Figure 14. Evaluation prompt used for assessing the accuracy of open-ended QA.

C. Experiment

C.1. Comparison Between CoT Reasoning and Direct Answering

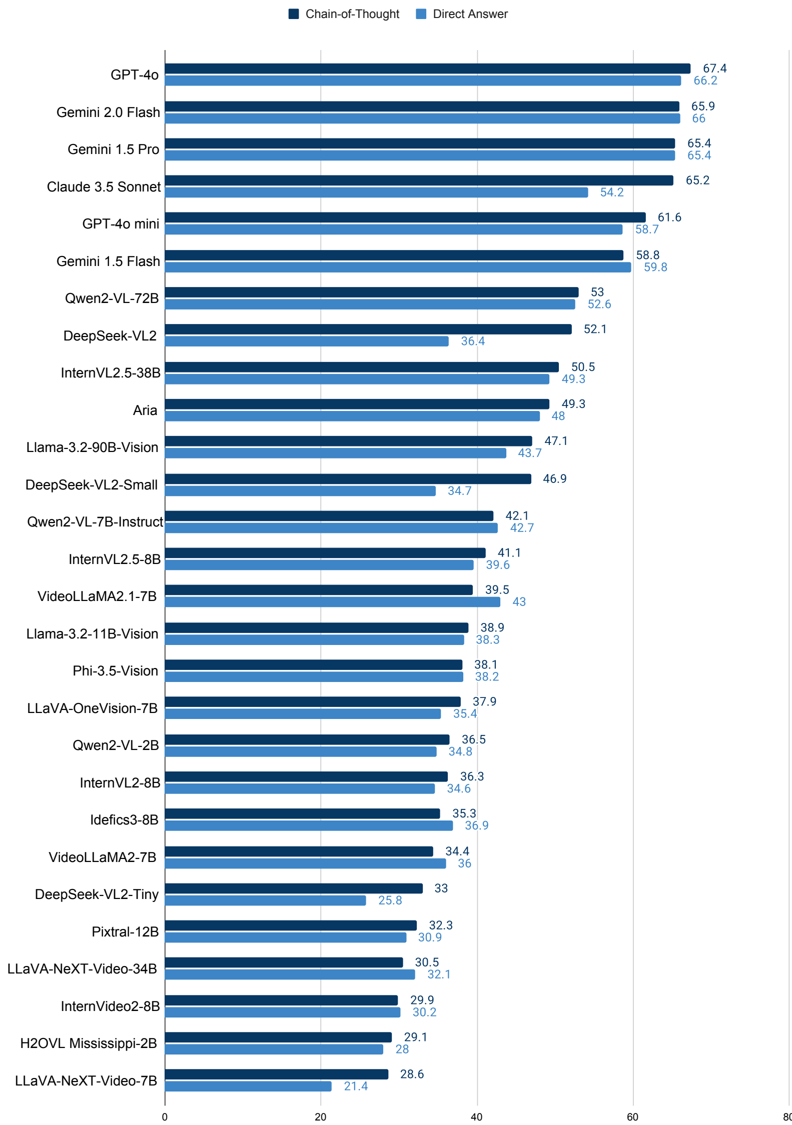
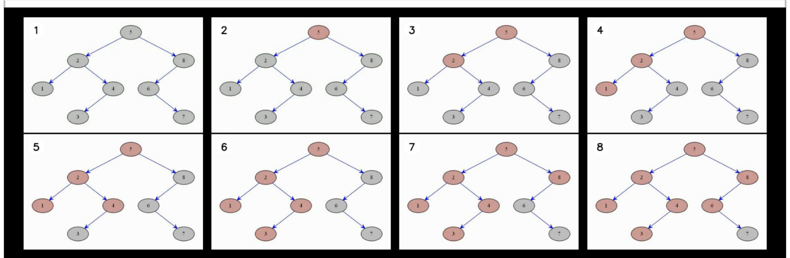


Figure 15. Comparison of model performance between CoT reasoning and direct answering on the validation set.

C.2. Error Case Analysis: Visual Perception Error

Visual Perception Error

Question: What is the algorithm used in the video?



Answer: Depth-First Search (DFS)
Reasoning Rational: In video, the traversal starts from the root node and proceeds downward, first exploring the left subtree and then the right subtree, which aligns with the characteristics of DFS; thus the algorithm used is DFS.

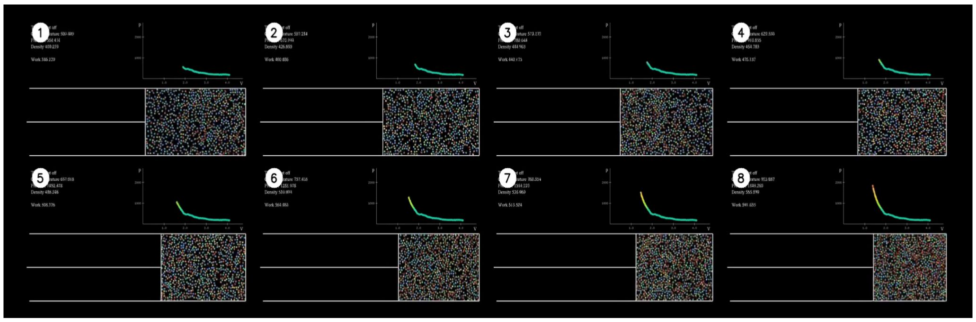
Llama-3.2-90B-Vision:
The algorithm used in the video is Breadth-first search (BFS). This conclusion can be drawn from the following observations: 1. Traversal Order: The traversal starts from the root (node with the value 5) and then moves to its direct children (nodes 2 and 8), and so on, level by level. 2. Pattern: BFS typically traverses a graph level by level, visiting all nodes at one depth before moving to the next deeper level. 3. Coloring of Nodes: The nodes are colored in the order of their traversal depth, starting from the top and moving down. (Human Evalutor's Comment: This observation contradicts with the video content, where the nodes in video are expanded left to right, top to down.) Given these observations, the algorithm used is BFS. Therefore, the final answer is: Breadth-first search (BFS).

Error Reason: The model fails to correctly identify the traversal order of the binary tree in the video, incorrectly interpreting it as "level by level." Thus it produces a wrong answer Breadth-First Search. The model does not accurately capture the temporal sequence information in the video.

Figure 16. An error case of Computer Science.

Thermodynamics

Question: Which process is shown in the animation?
Options:
(A) Isothermal compression (B) Adiabatic compression (C) Isobaric compression
(D) Adiabatic expansion (E) Isothermal expansion



Answer: B
Relevant Knowledge (Wikipedia page title): Adiabatic process, Isothermal process, Isobaric process, Ideal gas law
Reasoning Rational:
From the video, it can be seen that the gas is thermally isolated and returns to its initial state through compression. This physical process is known as adiabatic compression.

Llama-3.2-90B-Vision:
To determine the process shown in the animation, we need to analyze the provided graph and the printed values.

The graph plots pressure (P) against volume (V), and the curve is descending, indicating that the volume is increasing while the pressure decreases. This is characteristic of expansion.

Additionally, the absence of a thermostat indicates that no temperature regulation is involved. This suggests the process is not isothermal (constant temperature), but rather adiabatic (no heat exchange with the surroundings).

Since the volume is increasing, it is likely that the gas is expanding. (Human Evalutor's Comment: This observation contradicts with the video content, where the gas is being compressed.) Adiabatic expansion occurs when a gas expands without gaining or losing heat, and its temperature decreases.

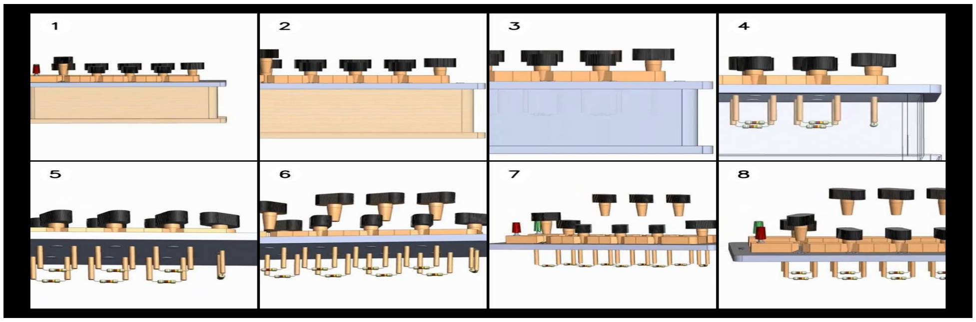
Therefore, the final answer is: D. Adiabatic expansion.

Error Category: Visual perception error
Error Reason: The model wrongly identifies the process of gas compression as the process of gas expansion, which is not consistent with the human observation and leads to the wrong final answer.

Figure 17. An error case of Thermodynamics.

Electromagnetism

Question: Which physical variable is changed in the animation?



Answer: Resistance
Relevant Knowledge (Wikipedia page title): Electrical resistance and conductance, Ohm's law, Electromagnetism
Reasoning Rational:
The video shows that by pulling out the knob, the resistance in the circuit can be changed.

DeepSeek-VL2:
In the animation, a PSP (Pressure Sensor Package) is shown being compressed by water, and its output increases as the depth increases due to the increased pressure and density of the water. (Human Evalutor's Comment: This observation contradicts with the video content, where there is no water.) This indicates that the deformation variable is the one that changes in the animation.

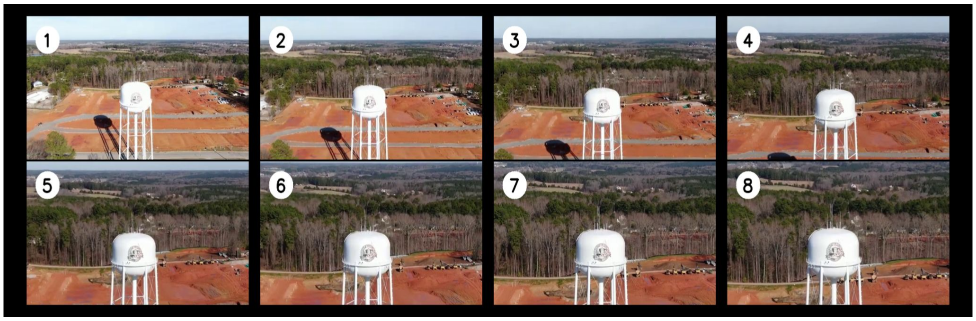
Therefore, the final answer is: Deformation

Error Category: Visual Perception Error
Error Reason: The model incorrectly identifies the device shell, which gradually turns blue and transparent in the video for inside demonstration, as water. The hallucination misleads the reasoning process and results in a totally wrong conclusion.

Figure 18. An error case of Electromagnetism.

Art

Question: Which cinematic shooting technique is shown in the video?



Answer: Dolly Zoom
Relevant Knowledge (Wikipedia page title): Dolly zoom, Cinematic techniques
Reasoning Rational:
"Dolly zoom" is a unique cinematographic technique that creates a visual distortion effect by simultaneously adjusting the lens while dollying in or out. The video clearly shows the camera zooming in on the object. Therefore, the most likely shooting technique demonstrated in the video is the dolly zoom.

Qwen2-VL-72B:
The video shows the panning technique, where the camera moves horizontally, from left to right gradually, (Human Evalutor's Comment: This observation contradicts with the video content, where the camera zooming in on the object, instead of moving horizontally.) to capture a wide view of the scene.

Therefore, the answer is: panning.

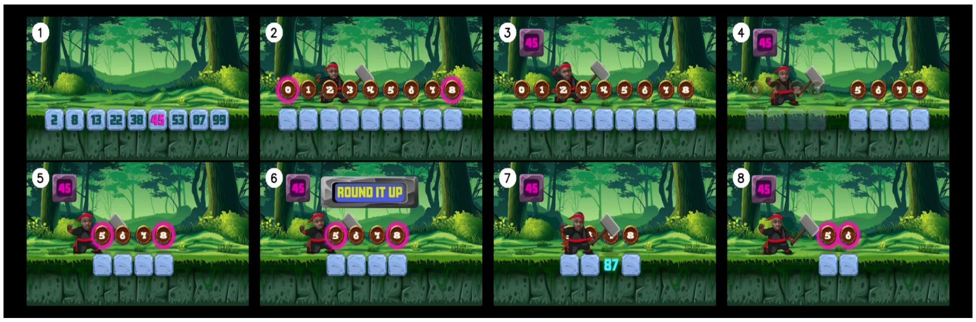
Error Category: Visual Perception Error
Error Reason: The model fails to interpret the dollying-in motion in the video or associate it with a cinematic technique. Instead, it thinks the motion of camera is panning from left to right. Therefore, the model fails to effectively interpret temporal visual data and it infers a wrong answer.

Figure 19. An error case of Art.

C.3. Error Case Analysis: Misuse or Lack Domain Knowledge in Visual Perception

Computer Science

Question: What is the algorithm shown in the video?



Answer: Binary Search
Relevant Knowledge (Wikipedia page title): Binary search, Search algorithm, Computer Science
Reasoning Rational:
The character in the video always locates data from the middle and then discards a part of the data, ultimately finding the target value. From the entire process, we can infer that this sequence is an ordered sequence, which aligns with the process of Binary Search.

Llama-3.2-90B-Vision:
To determine what algorithm is shown in the video, let's analyze the steps and their characteristics:
1. The video starts by showing a set of numbers.
2. It then rearranges the numbers in a specific order.

The observed steps suggest that the algorithm is sorting the numbers in ascending order. The algorithm sorts an array by repeatedly finding the minimum element from the unsorted part and putting it at the beginning. It involves selecting the smallest element and moving it to the correct position, which matches the steps shown in the video.

Based on the analysis, the algorithm shown in the video involves selecting the minimum element and moving it to the correct position repeatedly, which is characteristic of the selection sort algorithm. (Human Evalutor's Comment: Misuse of domain knowledge in visual perception. The indices of the array are mistakenly identified as the values.)

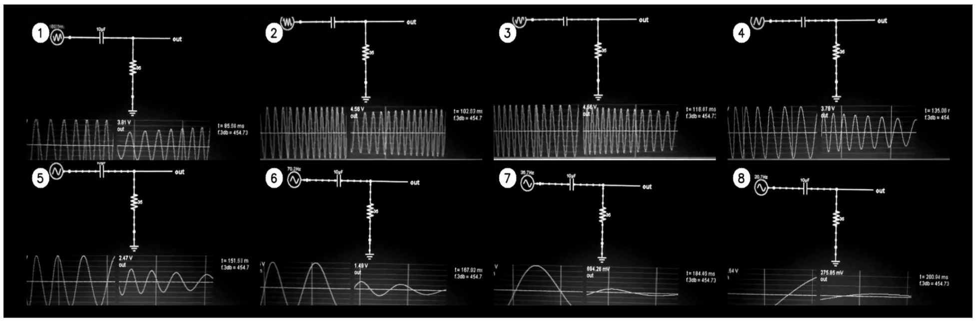
Therefore, the final answer is: Selection Sort.

Error Category: Misuse or Lack Domain Knowledge in Visual Perception
Error Reason: The model wrongly recognized the numerical sequence in the video as values (when in fact they are the indices of the sequence), leading to the incorrect conclusion that the algorithm shown in the video is a sorting algorithm.

Figure 20. An error case of Computer Science.

Electrical Engineering

Question: What kind of filter is the circuit implemented?



Answer: High-pass filter
Relevant Knowledge (Wikipedia page title): High-pass filter, RC circuit, Electronic filter, Signal processing
Reasoning Rational:
From the video, it can be seen that the circuit processes different frequency components of the signal, filtering out low frequencies to produce a new signal. Thus, it is a High-pass filter.

Qwen2-VL-72B:
To determine the type of filter implemented in the circuit, we need to analyze the components and their configuration.

1. Inductor and Capacitor Configuration:
   - The circuit includes an inductor and a capacitor.
   - These components are arranged in a series configuration.
2. Circuit Analysis:
   - The circuit diagram shows a series configuration of an inductor and a capacitor. (Human Evalutor's Comment: Misuse of domain knowledge in visual perception. The resistor is mistakenly identified as an inductor.)
   - This configuration is characteristic of a band-pass filter, where the inductor passes high frequencies and the capacitor passes low frequencies, but together they allow a specific range of frequencies to pass.
3. Confirmation:
   - The given figure and the analysis of the components and their arrangement strongly suggest that this is a band-pass filter.

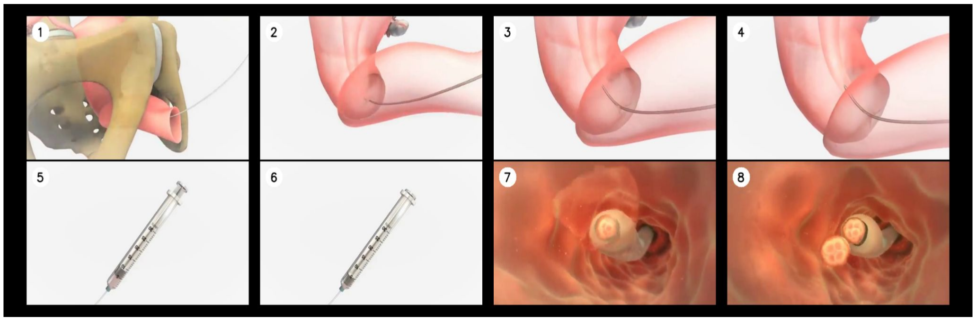
Therefore, the final answer is: Band-pass filter.

Error Category: Misuse or Lack Domain Knowledge in Visual Perception
Error Reason: The model lacks domain knowledge of common circuit component diagrams, leading to the incorrect identification of the resistor in the circuit as an inductor. The wrong identification of circuit components results in the false reasoning of the fitler type.

Figure 21. An error case of Electrical Engineering.

Pharmacy

Question: What process is shown in the video?



Answer: Embryo Transfer
Relevant Knowledge (Wikipedia page title): Embryo transfer, In vitro fertilisation, Childbirth, Female reproductive system
Reasoning Rational:
The video demonstrates the process of embryo transfer in which an embryo is implanted into a woman's uterus following in vitro fertilization. It features specific visuals of the fertilized egg, syringe, and the injection procedure.

DeepSeek-VL2:
The video shows a syringe injecting a cell into a tube, leading to the creation of baby DNA and cells multiplying, eventually forming a dark blue circular mass with an umbilical cord attached to it. This process is illustrated with illustrations of babies inside, suggesting fetal development. (Human Evalutor's Comment: Misuse of domain knowledge in visual perception. The uterus of a woman where the process happens is mistakenly identified as the inside of a fetus.) Therefore, the final answer is: Fetal Development Process.

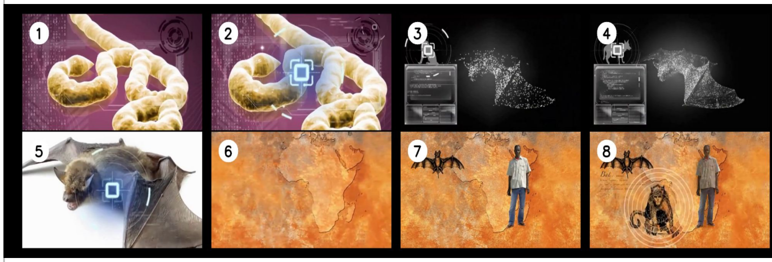
Error Category: Misuse or Lack Domain Knowledge in Visual Perception
Error Reason: The model misuses its domain knowledge of fetal development, wrongly identifying the uterus shown in the video as a fetus. Then the model draws the incorrect conclusion that the video demonstrates the fetal development process.

Figure 22. An error case of Pharmacy.

C.4. Error Case Analysis: Misuse or Lack Domain Knowledge in Reasoning

Misuse or Lack Domain Knowledge in Reasoning

Question: Which of the following virus infections does it belong to?
Options: (A) Norovirus (B) Measles virus (C) Hemorrhagic fever virus ✅
(D) Human papillomavirus (E) Arboviral encephalitis virus



Reasoning Rational:
The video displays a simplified structure of the Ebola virus (ssRNA) and its origin (bats), which suggests that the virus is the Ebola Virus Disease (EVD) that has had a significant impact on the globe. The Ebola virus is classified under regional hemorrhagic fever viruses.

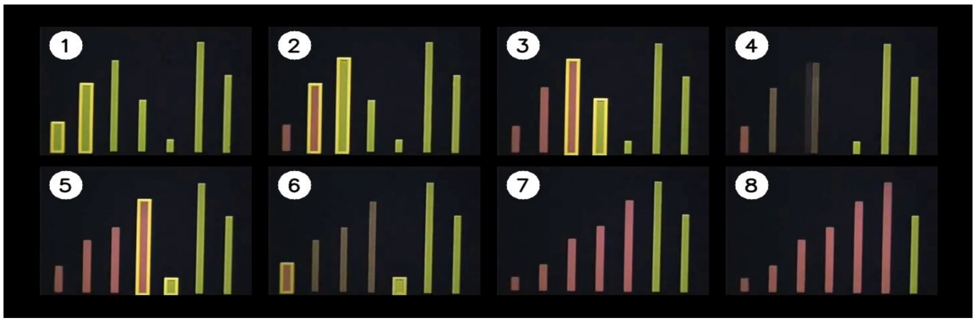
Qwen2-VL-72B:
Norovirus primarily infects areas with high population density and poor sanitation conditions. The worms and bats in the video indicate that the environment has poor sanitation conditions (Human Evalutor's Comment: Misuse of domain knowledge in reasoning. The bats represents the origin and the mode of transmission of the virus.). So the answer is (A) Norovirus.

Error Reason: The model misuses the domain knowledge that bats are found in some places with poor hygiene to reason that this is a norovirus outbreak occurring in an unhygienic location. However, the model should recall in its reasoning that "bats are significant sources and transmission vectors for certain diseases."

Figure 23. An error case of Basic Medicine.

Computer Science

Question: The algorithm demonstrated in the video is which common algorithm?
Options:
(A) Bubble Sort (B) Merge Sort (C) Insertion Sort (D) Selection Sort (E) Heap Sort



Answer: C
Relevant Knowledge (Wikipedia page title): Insertion sort, Bubble sort, Merge sort, Selection sort, Heapsort
Reasoning Rational:
The video ultimately sorts these blocks in ascending order, thus it is clearly a sorting algorithm. Observing the video process, there is no need to scan the sorting process from beginning to end, and there is an obvious insertion action, which allows us to eliminate Selection Sort, indicating it is Insertion Sort.

DeepSeek-VL2:
The given video shows computer bars that increase and decrease according to their size. The algorithm being demonstrated is a sorting algorithm, as it reorganizes the bars in order of their size. Among the given options, options (B) and (C) are not suitable, as Merge Sort and Insertion Sort are not typically demonstrated through visual bar representations. Option (F) is also not suitable, as Heap Sort is not typically demonstrated through visual bar representations. (Human Evaluator's Comment: Biased misuse of pre-trained domain knowledge in reasoning. The sorting algorithms can be demonstrated through visual bar representations and the form of presentation should not be the main focus.) Although option (D) Selection Sort is a possibility, the given algorithm does not repeatedly find the minimum element of the unsorted part and put it at the front, which characterizes Selection Sort. Therefore, the given algorithm is most likely a Bubble Sort, as this algorithm uses visual bar representations.

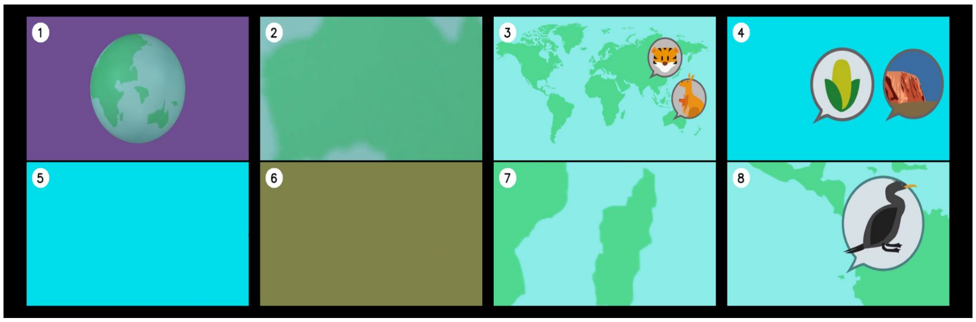
Therefore, the final answer is: A. Bubble Sort

Error Category: Misuse or Lack Domain Knowledge in Reasoning
Error Reason: The model considers Insertion Sort not to be the correct answer because it believes that Insertion Sort is not typically demonstrated through visual bar representations. This is a typical case of reasoning errors caused by the misuse of pretrained domain knowledge.

Figure 24. An error case of Computer Science.

Biology

Question: What biological process is shown in the video?
Options:
(A) Biodegradation (B) Community succession (C) Genetic drift (D) Migration of organism (E) Species formation



Answer: E
Relevant Knowledge (Wikipedia page title): Speciation, Reproductive isolation, Allopatric speciation, Charles Darwin, On the Origin of Species
Reasoning Rational:
In the video, the species undergo geographical isolation, which prevents gene flow with other populations. Over time, this leads to changes in allele frequencies within the gene pool. When significant differences arise between two populations' gene pools causing reproductive isolation, a new species is formed.

Qwen2-VL-72B:
The video shows a girl floating over a virtual map and selecting various animals. This image is used to demonstrate the process of migration of organisms. The girl's selection of animals and their placement on the map suggests the migration of organisms to different locations. (Human Evalutor's Comment: Misuse of domain knowledge in reasoning. The movement of organisms is due to the continental drift, not a spontaneous migration.) Additionally, the video features a tree, a fish, a koala, a giraffe, and a lion on the edges of the map, which further supports the notion of migration. Based on this analysis, the biological process shown in the video is the migration of organisms.

Therefore, the final answer is: {D. Migration of organisms}

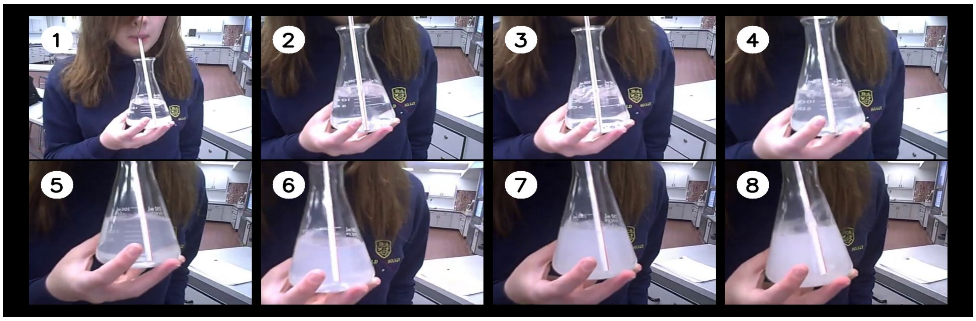
Error Category: Misuse or Lack Domain Knowledge in Reasoning
Error Reason: The model misuses pretrained knowledge and incorrectly infers that this is a case of spontaneous organism migration. In fact, the organism movement and isolation is caused by continental drift.

Figure 25. An error case of Biology.

Chemistry

Question: Assume that 2.24 liters of gas fully participates in the reaction shown in the video under the standard temperature and pressure condition, how many grams of precipitate are produced approximately?

Options: (A) 10.0 (B) 5.0 (C) 12.0 (D) 15.0 (E) 20.0



Answer: A
Relevant Knowledge (Wikipedia page title): Calcium hydroxide, Carbon dioxide, Ideal gas law
Reasoning Rational:
1. In the video, a person exhales gas that is continuously introduced into a clear solution, gradually forming a white precipitate. This indicates that the substances involved in the reaction are CO_2 and limewater.
2. The chemical reaction equation is: $Ca(OH)_2 + CO_2 \rightarrow CaCO_3 + H_2O$
3. At the STP, 2.24 liters of CO_2 corresponds to 0.1 Moles.
4. From balanced equation, 0.1 moles of CO_2 produce 0.1 moles of $CaCO_3$.
5. Given Ca = 40 g/mol, C = 12 g/mol, O = 16 g/mol, the molar mass of $CaCO_3$ = 40 + 12 + 16 * 3 = 100 g/mol. Therefore, the mass of $CaCO_3$ = 0.1 * 100 = 10g.

DeepSeek-VL2:
To determine the amount of precipitate produced, we need to identify the chemical reaction involved and use the stoichiometry of the reaction. However, the video does not provide enough information to directly identify the reaction (Human Evalutor's Comment: Lack of domain knowledge in reasoning. The model fails to recall and apply relevant chemical equations.). Assuming the reaction is a typical precipitation reaction involving a gas and a solution, we can use the ideal gas law (PV=nRT) to estimate the volume of gas produced under standard temperature and pressure (STP).

Given:
- Volume of gas produced = 2.24 liters
- STP conditions: P = 1 atm, T = 273 K, R = 0.0821 L·atm/(mol·K)

First, we calculate the number of moles of gas produced using the ideal gas law:
n = PV/RT = (1 atm * 2.24 L) / (0.0821 (L·atm)/(mol·K) * 273 K) ≈ 0.1 mol

Assuming the reaction produces a precipitate with a known stoichiometry, we can calculate the mass of the precipitate. However, without knowing the specific reaction, we cannot accurately determine the mass of the precipitate.

Given the options, we need to make an educated guess based on typical reactions. If we assume the reaction produces a precipitate with a molar mass of around 50 g/mol (a common value for many precipitates), we can estimate the mass of the precipitate as follows:
Mass of precipitate = n * Molar mass = 0.1 mol * 50 g/mol ≈ 5 g

Therefore, the correct answer is (B) 5.0 g.

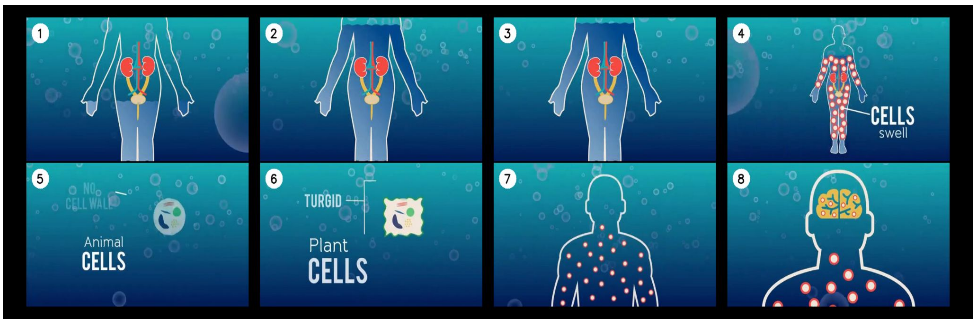
Error Category: Misuse or Lack Domain Knowledge in Reasoning
Error Reason: The model fails to recall and apply the correct relevant chemical equations. Instead, it merely makes a vague statement and then guesses a possible ratio between the products and reactants, which comes to a wrong conclusion at last.

Figure 26. An error case of Chemistry.

C.5. Error Case Analysis: Heavy Reliance on Textual Information

Clinical Medicine

Question: When the phenomenon in the video occurs, which of the following options is the most important to adjust?
Options:
(A) Renin-angiotensin system
(B) Hypothalamic-pituitary-thyroid axis
(C) Adrenaline and norepinephrine
(D) Melatonin secretion
(E) Sodium-potassium pump activity



Answer: A
Relevant Knowledge (Wikipedia page title): Renin–angiotensin system, Homeostasis, Assessment of kidney function, Osmoregulation
Reasoning Rational:
The issue presented in the video is related to the imbalance of water within the body, causing cells to swell and resulting in symptoms such as headaches. Among the options provided, the regulation of water balance is primarily managed through the renin-angiotensin system.

Llama-3.2-90B-Vision:
The phenomenon in the video likely refers to the effects of overhydration on the body. To adjust this, the most important physiological response would be to regulate the body's sodium level, which is primarily managed by the sodium-potassium pump activity. (Human Evalutor's Comment: The model ignores the video information about water balance.) This pump plays a crucial role in maintaining the balance of electrolytes and fluids within the body, which is crucial for the proper functioning of cells.

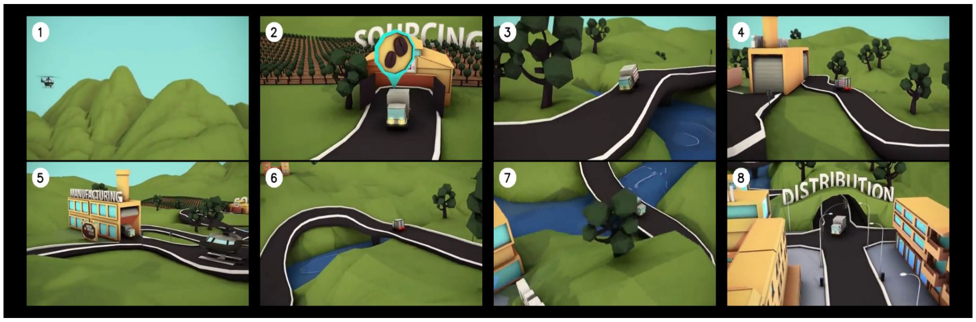
Therefore, the final answer is: {E. Sodium-potassium pump activity}.

Error Category: Heavy Reliance on Textual Information
Error Reason: The model ignores the video information about water excretion in the human body. It relies solely on the text and chooses the incorrect option E simply because cellular ion balance is crucial to the body's internal environment, thereby missing the actual disorder of fluid overload mentioned in the video.

Figure 27. An error case of Clinical Medicine.

Management

Question: For the company shown in the video, what is one of the methods to handle associated risks?
Options:
(A) diversification (B) vertical integration (C) horizontal integration (D) outsourcing (E) speculation



Answer: B
Relevant Knowledge (Wikipedia page title): Vertical integration, Horizontal integration, Diversification, Risk management, Supply chain management
Reasoning Rational:
According to the content of the video, the company being discussed is a coffee company (the bubbles above the truck represent coffee beans) along with its upstream supply chain (costs of raw coffee beans, labor costs, etc.). A common practice to mitigate the price risk of upstream products is vertical integration, which involves establishing the coffee brand's own coffee plantation.

Qwen2-VL-72B:
To handle the risk associated with a company, diversification is one of the methods that can be used. Diversification involves spreading investments across various sectors or industries to reduce risk. Vertical integration and horizontal integration are strategies to gain control over the supply chain or expand market dominance, respectively, but they do not directly address risk management (Human Evalutor's Comment: The model ignores the video information about the features of the company.). Inventory management and outsourcing are operational strategies that can influence risk, but they are not specifically methods for handling risk. Therefore, the correct answer is Option A. Therefore, the final answer is: A. diversification.

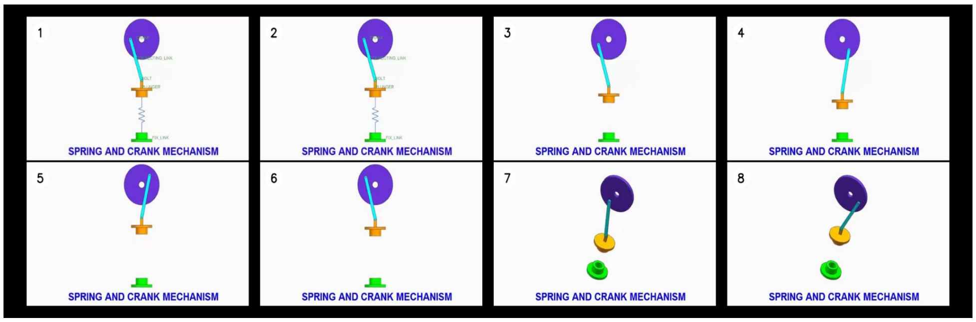
Error Category: Heavy Reliance on Textual Information
Error Reason: The model ignores the features of the company demonstrated in the video. Instead, the model jumps to the conclusion that (A) diversification is the most appropriate method of risk management, solely based on the question and the given options.

Figure 28. An error case of Management.

C.6. Error Case Analysis: Logical Reasoning Error

Mechanical Engineering

Question: What kind of mechanical transformation is not represented in the video?
Options:
(A) It may convert gravitational energy into elastic energy.
(B) It may convert one force to another force.
(C) It may convert one speed to another speed.
(D) It may convert linear motion into rotational motion.
(E) It may convert gravitational energy to kinetic energy.



Answer: D
Relevant Knowledge (Wikipedia page title): Mechanical energy, Kinematics, Mechanical Engineering
Reasoning Rational:
In the video, the rotation of the disk carries the vertical movement of the animal block, and the speed of the block changes during the start and stop. The gravity of the mass and the elasticity of the spring transform each other. So the choice is D.

Llama-3.2-90B-Vision:
To determine which type of mechanical transformation is not represented in the video, we need to consider the actions depicted in the video and the various forms of energy and motion transformation mechanisms typically found in mechanical systems.

1. Linear Motion into Rotational Motion: The video shows a mechanical linkage where a linear motion is converted into rotational motion.
2. Rotational Motion into Linear Motion: The video also shows the reverse process, where rotational motion is converted into linear motion, which is visible in the reciprocal motion of the device.
3. Kinetic Energy Transformation: The video shows a transformation of kinetic energy from one part of the mechanism to another as parts move and interact.
4. Elastic Energy Transformation: The presence of a spring in the mechanism suggests the conversion of kinetic energy into elastic potential energy and vice versa.

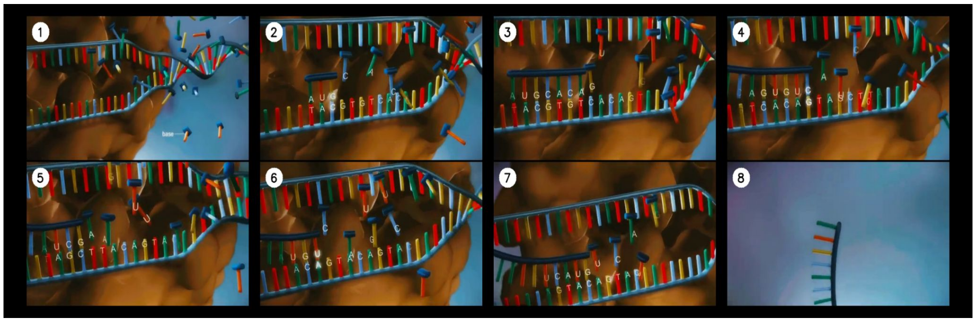
Now, let's evaluate the options to see which transformation is not explicitly shown in the video:
Option A: It may convert gravitational energy into elastic energy. - This transformation is not directly shown in the video. (Human Evalutor's Comment: The reasoning here contradicts with the other reasoning.) The video focuses on the interactions between mechanical parts and their resulting motions rather than showing the specific conversion of gravitational energy into elastic energy.
Option B: It may convert one force to another force. - This is generally true as mechanical systems often convert forces between different parts.
Option C: It may convert one speed to another speed. - Mechanical systems often include gears and linkages that change speeds.
Option D: It may convert linear motion into rotational motion. - This is explicitly shown in the video.
Option E: It may convert gravitational energy to kinetic energy. - This transformation could be inferred based on the interactions of the parts, but it is not the main focus of the video.
After careful analysis, the final answer is: {A}.

Error Category: Logical Reasoning Error
Error Reason: There is self-contradiction in the entire reasoning process. The model believes that kinetic energy can be converted into elastic potential energy, and that gravitational potential energy can be converted into kinetic energy. According to this logic, gravitational potential energy can also be converted into elastic potential energy, which aligns with the statement in option A. However, the model incorrectly assumes that there is no mechanical transformation mentioned in option A in the video, resulting in a contradictory reasoning process.

Figure 29. An error case of Mechanical Engineering.

Clinical Medicine

Question: What could the brown stuff in the video be?
Options:
(A) peptidyltransferase (B) RNA polymerase (C) DNA polymerase (D) Topoisomerase (E) Spliceosome complex



Answer: B
Relevant Knowledge (Wikipedia page title): RNA polymerase, Transcription (biology), Peptidyltransferase, DNA polymerase, Topoisomerase, Spliceosome
Reasoning Rational:
The video demonstrates the process of DNA unwinding and transcription of RNA under the action of enzymes, specifically highlighting the role of RNA polymerase.

DeepSeek-VL2:
The brown structure in the video likely represents a biological process, as it is associated with DNA, RNA, and various molecular structures. The brown stuff could correspond to an enzyme that replicates DNA or RNA, or it could represent a complex that splices a strand of RNA. DNA polymerase synthesizes DNA strands. RNA polymerase synthesizes RNA from DNA, and peptidyltransferase catalyzes the formation of peptide bonds between two amino acids. Topoisomerase adjusts the tension or catenation in DNA by introducing or removing supercoils. Spliceosomes are complexes involved in pre-mRNA splicing.

Given the context, it seems most plausible that the brown stuff represents an enzyme involved in the replication of RNA (RNA polymerase) or an enzyme involved in cutting and splicing an mRNA strand (spliceosome complex). Therefore, we can eliminate options A, B, C, and D based on the actions shown in the video. (Human Evalutor's Comment: Self-contradiction.) The most likely final answer is: E, Spliceosome complex.

Error Category: Logical Reasoning Error
Error Reason: The model thinks that the brown stuff represents a RNA polymerase or a spliceosome complex. But when the model comes to the conclusion, it excludes the correct answer spliceosome complex without any further reason, which contradicts with its previous logical reasoning.

Figure 30. An error case of Clinical Medicine.