

Video-MMMU: Evaluating Knowledge Acquisition from Multidisciplinary Professional Videos

Kairui Hu¹, Penghao Wu¹, Fanyi Pu¹, Wang Xiao¹,
Xiang Yue², Bo Li¹, Yuanhan Zhang¹, Ziwei Liu^{1,✉}

¹S-Lab, Nanyang Technological University, ²Carnegie Mellon University
{kairui.hu, ziwei.liu}@ntu.edu.sg, fpu001@e.ntu.edu.sg

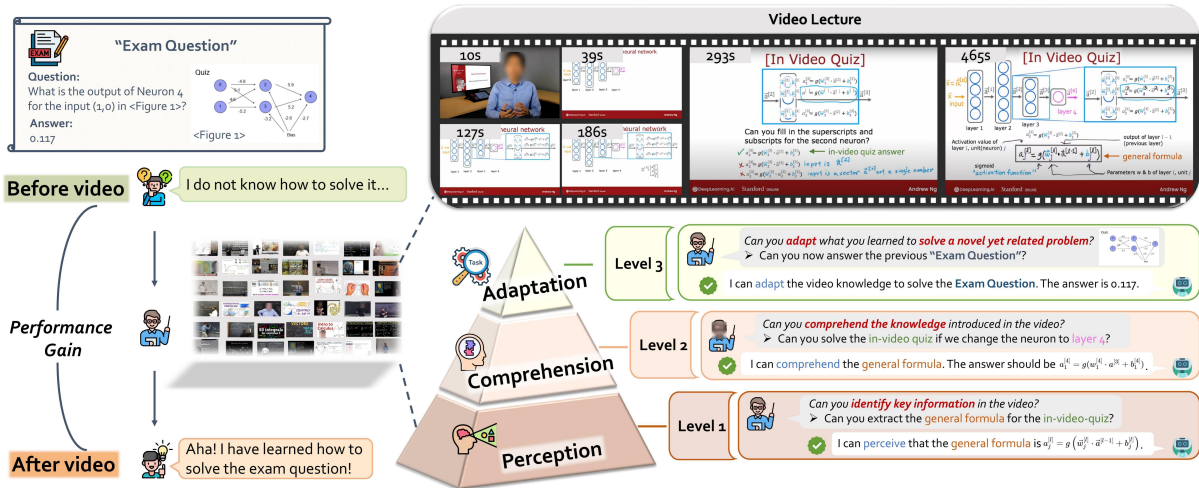


Figure 1: **Overview of Video-MMMU:** a) From visual understanding towards **knowledge acquisition**: Video-MMMU evaluates LMMs’ ability to acquire knowledge from educational videos through three cognitive stages: **1) Perception**: identify key information; **2) Comprehension**: understand the underlying concepts; **3) Adaptation**: adapt the knowledge to novel scenarios. b) From absolute performance towards **learning gain**: Video-MMMU does not only measure the final accuracy but also the performance gain: how effectively the model learns from video to solve an initially unsolvable problem.

Abstract

Humans acquire knowledge through three cognitive stages: perceiving information, comprehending knowledge, and adapting knowledge to solve novel problems. Videos serve as an effective medium for knowledge acquisition, facilitating a natural progression through these learning stages. However, existing video benchmarks fail to evaluate the *knowledge acquisition* capabilities of Large Multimodal Models (LMMs). To address this gap, we introduce Video-MMMU, a multi-modal, multi-discipline, multi-track benchmark that evaluates LMMs’ ability to acquire knowledge from college-level, educational videos. Video-MMMU features a collection of 300 videos and 900 human-annotated questions across six disciplines, evaluating *knowledge acquisition* through stage-aligned question-answer pairs: Perception, Comprehension, and Adaptation. Beyond measuring final accuracy, Video-

MMMU proposes the performance gain metric that quantifies an LMM’s learning gain from video, shifting the focus of evaluation from absolute performance to learning efficiency. Our evaluation reveals a substantial gap between human learners and current LMMs, highlighting the need to improve models’ ability to learn and adapt knowledge from video content.

1 Introduction

Humans acquire knowledge through cognitive stages as outlined in Bloom’s Taxonomy (Bloom et al., 1956): perceiving information, comprehending concepts, and applying knowledge to solve problems. Video is an ideal medium for knowledge acquisition, enabling a progression from information intake to practical application (Sablić et al., 2021; Yousef et al., 2014). Consider learning neural network forward propagation through a video lecture in Fig. 1. Before watching, learners may be unable to solve a related exam question. As they

✉Corresponding author.






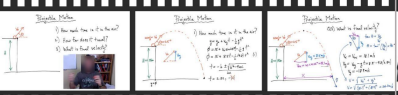

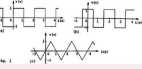
Art	Humanities	Medicine
 <p>Question: What does the speaker say when introducing Peter Paul Rubens at the end of the video? Select the option that precisely matches the speaker's statement.</p> <p>Options: (A) Peter Paul Rubens was a famous Baroque... (B) Peter Paul Rubens is regarded as a prolific artist... (I) Peter Paul Rubens was the most important... (J) Peter Paul Rubens is celebrated for his dynamic...</p> <p>Track: Perception, Video Type: Concept-introduction video, Subject: Art Theory, QA Type: Automatic Speech Recognition (ASR)</p>	 <p>Question: Based on your understanding of cultural universals from the video, determine which of the following statements are correct: Statement 1: All human cultures have some... Statement 2: The video uses the example of... Statement 3: At 3:35, the video implies that ... Statement 4: ... Statement 5: ...</p> <p>Options: (A) Statement 1 (B) Statement 2,3 (C) Statement 3,4 (D) Statement 2,4,5(J) Statement 2,4</p> <p>Track: Comprehension, Video Type: Concept-introduction video, Subject: Sociology, QA Type: Concept Comprehension (CC)</p>	 <p>Question: Can you identify the abnormality on this plain film of the pelvis? <image 1></p>  <p>Options: (A) Bone cyst (B) Acute hip fracture (C) Osteoarthritis (D) Surgical hardware (E) Resection of the pubic symphysis ... (J) Bone infection</p> <p>Track: Adaptation, Video Type: Concept-introduction video, Subject: Clinical Medicine, QA Type: Case Study Analysis (CSA)</p>
Business	Science	Engineering
 <p>Question: According to the video, a minimum price control on alcoholic drinks is intended to reduce consumption from Q_1 to _____, addressing negative externalities. The policy raises the price to _____ above the free market price of _____. Fill in the blanks based on the video content.</p> <p>Options: (A) Q^*, P_{min}, P_1 (B) Q^*, P_1, P_{min} (C) Q_1, P_{min}, P_2 (D) Q_2, P_1, P_{min} (E) Q^*, P_2, P_1 ... (F) Q_1, P_2, P_{min} (G) Q_2, P_{min}, P_1. (H).... (I).... (J) Q_1, P_1, P_{min}</p> <p>Track: Perception, Video Type: Problem-solving video, Subject: Economics, QA Type: Optical Character Recognition (OCR)</p>	 <p>Question: In the video, Example Question (1) is solved with an angle $\theta=25$ degrees. If the angle θ is adjusted to 30 degrees while all other conditions remain unchanged, what will be the updated result for Example Question (1) as explained in the video?</p> <p>Options: (A) 4.00 seconds (B) 2.82 seconds (C) 3.50 seconds (D) 2.50 seconds (E) 3.04 seconds (F) 2.00 seconds (G) 3.15 seconds (H) 1.85 seconds (I) 2.25 seconds (J) 3.85 seconds</p> <p>Track: Comprehension, Video Type: Problem-solving video, Subject: Math, QA Type: Problem-solving Strategy Comprehension (PSC)</p>	 <p>Question: Based on what you learned from the video, write the Fourier series for the three voltage waveforms in (a) of <image 1></p>  <p>Options: (A) $(4/\pi)(\sin(\pi t)+(1/2)\sin(3\pi t)+(1/4)\sin(5\pi t)+\dots)$ (B) $(4/\pi)(\sin(\pi t)+(1/3)\sin(3\pi t)+(1/5)\sin(5\pi t)+\dots)$ (C) $(4/\pi)(\sin(\pi t)+(1/2)\sin(2\pi t)+(1/4)\sin(4\pi t)+\dots)$ (J) $(4/\pi)(\sin(\pi t)+(1/4)\sin(3\pi t)+(1/6)\sin(5\pi t)+\dots)$</p> <p>Track: Adaptation, Video Type: Problem-solving video, Subject: Electronics, QA Type: Problem-solving Strategy Adaptation (PSA)</p>

Figure 2: Sampled Video-MMMU examples across the 6 disciplines and 3 cognitive tracks.

engage with the video, they identify formulas, build conceptual understanding, and eventually apply the knowledge to problem-solving. This learning gain can be quantified as normalized performance gain (Hake, 1998), which captures the difference in performance before and after watching the video.

For Large Multimodal Models (LMMs) to operate in the wild like humans (Feng et al., 2024), acquiring knowledge from videos is essential. However, existing video benchmarks primarily focus on static, general visual understanding tasks, without evaluating whether models can acquire knowledge dynamically from educational videos. To bridge this gap, we introduce **Video-MMMU**, a multi-modal, multi-discipline, multi-track video benchmark that evaluates LMMs' knowledge acquisition capability from videos. Video-MMMU provides a fundamentally new perspective on VideoQA (question answering): 1) *Video as an Educational Medium*: Video-MMMU is the first to recognize videos as educational resources. It comprises 300 college-level videos spanning 30 subjects in 6 disciplines: Art, Business, Science, Medicine, Hu-

manities, and Engineering. 2) *Cognitive Learning QA Design*. Video-MMMU shifts the focus from general visual understanding to in-context knowledge acquisition. Each video is paired with three question-answer tasks aligned with human learning stages: Perception (identifying key information), Comprehension (understanding underlying principles), and Adaptation (applying acquired knowledge to a new scenario). 3) *From Absolute Performance towards Learning Gain*: A key innovation of Video-MMMU is that it evaluates not only a model's static task-solving ability (final accuracy), but also its learning gain: how much the model improves after watching the video. Specifically, Video-MMMU measures a model's performance on the Adaptation question without video and re-evaluates it after providing the video from which a human learner could learn to solve the question. Following the educational research (Hake, 1998), we quantify the learning gain as the normalized performance gain.

Our experiments show the difficulty of Knowledge acquisition from videos, revealing a no-

table gap between humans and models in learning gain. While humans achieve substantial improvement (Gain = 36.5%) after video, even top-performing models show smaller gains (Gain = 13.0%). These findings emphasize the need to improve how LMMs acquire and apply knowledge from video, bringing them closer to human-level learning capabilities.

2 Related Work

2.1 VideoQA Benchmarks

Existing video benchmarks focus primarily on general-purpose visual understanding tasks, including action recognition (Yu et al., 2019; Patraucean et al., 2023; Mangalam et al., 2023; Xiao et al., 2021), temporal reasoning (Liu et al., 2024; Wang et al., 2025; Wu et al., 2024; Cai et al., 2024). More recently, benchmarks such as (Fu et al., 2025; Fang et al., 2024; Zhou et al., 2025; Liu et al., 2026; Yang et al., 2026) have expanded to multi-task video understanding and agentic tasks. Existing benchmarks primarily focus on visual understanding tasks in general domain videos. In contrast, Video-MMMU evaluates knowledge reasoning from educational videos.

2.2 Knowledge-Driven Benchmarks

As Artificial General Intelligence (AGI) progresses toward expert-level intelligence (Morris et al., 2024a), knowledge-driven benchmarks have emerged to evaluate models’ professional expertise. MMLU (Hendrycks et al., 2021) and MMLU-Pro (Wang et al., 2024) evaluate models using textual exam questions, while MMMU (Yue et al., 2024) and MMMU-Pro (Yue et al., 2025) expand the modality to include images. MMWorld (He et al., 2025), MMVU (Zhao et al., 2025) further introduce video modality into the evaluation. Existing benchmarks primarily evaluate static final accuracy. Video-MMMU incorporates dynamic learning gain, providing a new perspective on knowledge-driven video QA. Tab. 1 summarizes the key differences between Video-MMMU and other benchmarks.

3 Video-MMMU Dataset

Video-MMMU evaluates knowledge acquisition from college-level videos in 6 disciplines. The distribution across disciplines is shown in Fig. 3a.

Benchmarks	Video Domain	QA Len.	Video Len.	Learning Gain
Video-MME (Fu et al., 2025)	Open	35.7	1017.9	✗
MMBench-Video (Fang et al., 2024)	Open	10.9	165.4	✗
TempCompass (Liu et al., 2024)	Open	49.2	11.4	✗
MVBench (Li et al., 2024)	Open	27.3	16.0	✗
MMVU (Zhao et al., 2025)	Scientific	40.7	51.4	✗
WorldQA (Zhang et al., 2024)	Open	38.5	60.0	✗
Video-MMMU	Academic	75.7	506.2	✓

Table 1: Comparison of video benchmarks.

3.1 Curriculum-Driven Video Curation

Exam-Point-Driven Topic Selection. To ensure reasoning depth and disciplinary coverage, we adopt a curriculum-driven curation pipeline. We extract fine-grained *Exam Points* from textbooks, each corresponding to a college-level knowledge topic. This yields an initial pool of Exam Points, as detailed in Appendix D.

Educational Video Sourcing. For each Exam Point, annotators retrieve educational videos from YouTube. We prioritize lecture-style videos that provide detailed explanations or step-by-step example questions with explicit derivations, rather than fast-paced “crash course”. This ensures that each video contains clear visual cues for perception, detailed explanations for comprehension, and well-developed examples for adaptation.

Expert-Guided Quality Control. All videos undergo a rigorous quality-control process including annotator scoring, cross-checking, and final validation by human experts, as detailed in Appendix C.1. Videos are evaluated along three dimensions: visual clarity, technical depth, and alignment with the Exam Point, with each rated on a 1–3 scale. Only videos with the highest score are retained. The total video length exceeds 42.18 hours.

Video Category. Through this curriculum-guided and expert-validated process, the initial pool is systematically filtered into 300 high-quality videos. These videos fall into two categories: *Concept-Introduction Videos*, which explain foundational theory, and *Problem-Solving Videos*, which demonstrate step-by-step solution methods.

3.2 QA Taxonomy

We annotate QA pairs in three cognitive stages: Perception, Comprehension, and Adaptation, each assessing deeper levels of knowledge acquisition.

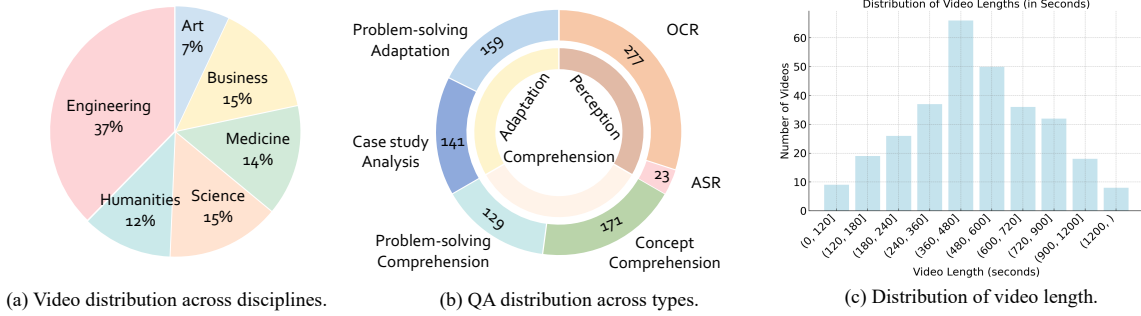


Figure 3: Taxonomy of QA types and video disciplines.

Perception Questions evaluate the ability to extract information from the video. (1) *Optical Character Recognition (OCR)*: These questions require identifying visual details such as charts and handwritten notes, as illustrated in Fig. 2 (Business). (2) *Automatic Speech Recognition (ASR)*: These questions require transcribing spoken content into text, as shown in Fig. 2 (Art).

Comprehension Questions evaluate the understanding of knowledge: (1) *Concept Comprehension (CC)*: These questions evaluate the conceptual understanding with a multiple-answer multiple-choice format (MAMC), as shown in Fig. 2 (Humanities). (2) *Problem-solving Strategy Comprehension (PSC)*: For videos with an example question, we test understanding by presenting the same question with modified input values. In Fig. 2 (Science), the video shows calculation with a 25-degree angle, and our question changes to 30 degrees. One can follow the same reasoning to demonstrate the comprehension of method.

Adaptation Questions measure the ability to transfer the knowledge to new scenarios. (1) *Case Study Analysis (CSA)*: These questions require applying concepts to realistic scenarios. In Fig. 2 (Medicine), the video explains a diagnostic concept, while the QA asks to diagnose a new patient’s radiograph. (2) *Problem-Solving Strategy Adaptation (PSA)*: These questions present a novel problem that requires adapting the method from video. As shown in Fig. 2 (Engineering), the video analyzes one waveform while the question asks a different one. The distribution of question types is shown in Fig. 3b.

3.3 Annotation and Quality Control

To ensure high-quality annotations, we adopt a time-controlled annotation protocol to reduce annotator fatigue. Annotators are compensated on an hourly basis (\$20/hr) with a strict workload limit

of 10 QAs per day. Each sample requires an average of 35.2 minutes for annotation and 8.3 minutes for validation, allowing sufficient time for annotation, and the total annotation time exceeds 217.5 hours. To further ensure annotation quality, we implement a three-stage *Rejection–Refinement Loop* to validate all questions. The details are provided in Appendix C.

Stage 1: Self-Verification. Annotators first conduct a self-check using a standardized 3-point rubric covering clarity, technical correctness, and reasoning depth. Annotators are required to refine their drafts to the best of their judgment.

Stage 2: Cross-Review. To reduce individual bias, each QA is cross-reviewed by a different annotator using the same rubric. Discrepancies or non-optimal ratings trigger a refinement, and 169/900 QAs are revised until consensus is reached.

Stage 3: Expert Validation. Finally, all QAs are reviewed by PhD domain experts to verify technical accuracy and alignment with the Exam Point. During expert review, 13.7% (123/900) of QAs are marked for refinement and iteratively revised until final consensus is reached. This rigorous multi-stage pipeline ensures reliable annotation quality.

4 Experiments

4.1 Settings

Baselines. We evaluate open-source models including LLaVA-OneVision (Li et al., 2025), LLaVA-Video (Zhang et al., 2025b), Qwen3-VL (Bai et al., 2025), MAMmoTH-VL (Guo et al., 2025), Llama-3.2-Vision (Meta, 2024). Proprietary models include GPT-4o (Hurst et al., 2024), o4-mini (OpenAI, 2025b), GPT-5 (OpenAI, 2025a), Gemini-3.0-Pro (DeepMind, 2025), Claude-3.5-Sonnet (Anthropic, 2024). The full evaluation setting is shown in Appendix A.2. Finally, we establish human base-

Model	Overall	Track			Discipline					
		Per.	Com.	Ada.	Art	Biz.	Sci.	Med.	Hum.	Eng.
Random Choice	14.00	12.00	14.00	16.00	11.11	12.88	12.12	22.48	10.48	13.57
Human (low)	74.44	84.33	78.67	60.33	80.95	78.79	74.24	70.54	84.76	69.91
Human (high)	90.45	91.67	90.67	89.00	85.71	93.94	90.15	93.02	94.29	87.91
<i>Proprietary Models</i>										
GPT-4o	59.22	67.67	55.00	55.00	71.43	69.70	50.76	64.34	72.38	50.15
Claude-3.5-Sonnet	67.67	76.67	66.00	60.33	73.02	81.82	61.36	62.79	75.24	63.13
O4-mini	77.78	83.67	77.00	72.67	77.78	84.85	71.97	78.29	83.81	75.22
GPT-5	79.22	85.33	79.00	73.33	73.02	83.33	76.52	82.17	81.90	77.88
Gemini-3.0-Pro	84.22	88.33	84.33	80.00	77.78	90.15	81.82	86.82	92.38	80.53
<i>Open-Source Models</i>										
LLaVA-OneVision-7B	33.33	39.67	30.33	30.00	49.21	28.79	34.85	31.78	45.71	28.32
Llama-3.2-11B	34.67	42.67	36.33	25.00	41.27	34.09	28.03	41.09	36.19	33.33
LLaVA-Video-7B	38.22	45.33	35.00	34.33	65.08	36.36	34.85	46.51	45.71	29.79
MAmmoTH-VL-8B	41.89	52.33	40.33	33.00	44.44	38.64	34.09	38.76	48.57	44.84
LLaVA-OneVision-72B	49.33	60.67	44.67	42.67	61.90	49.24	40.15	55.81	60.00	44.84
LLaVA-Video-72B	51.00	61.67	47.67	43.67	69.84	47.73	44.70	59.69	57.14	46.02
Qwen3-VL-30B-A3B	67.33	72.33	73.00	56.67	69.84	76.52	61.36	71.32	70.48	63.13
Qwen3-VL-235B-A22B	74.89	77.67	77.33	69.67	76.19	81.06	71.21	72.09	80.00	73.16

Table 2: Video-MMMU results across tracks (Perception, Comprehension, Adaptation) and disciplines (Art, Business, Science, Medicine, Humanities, Engineering).

lines by recruiting 60 undergraduate students from relevant majors. To ensure statistical representativeness, we maintained a minimum sample size of 4 participants per discipline. We report two human baselines: Human (high), representing top-performing students (GPA top 10% in their major) as an approximate upper bound, and Human (low), representing average undergraduate performance to reflect typical learning outcomes.

Evaluations. We use Gemini-2.5-Flash-Lite (Comanici et al., 2025) as the judge to extract the answer from model responses. The extracted answer is then compared against ground truth using exact match. We report micro-averaged accuracy. Evaluations are conducted using LMMs-Eval (Zhang et al., 2025a) for consistency and reproducibility.

4.2 Main Results

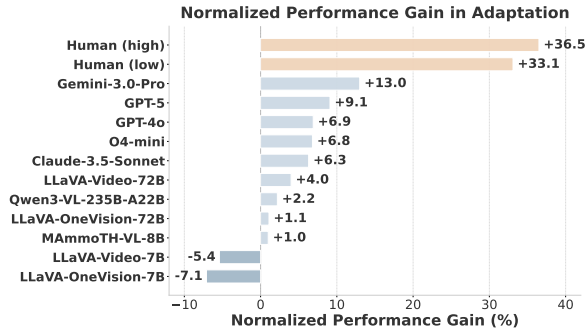
Performance by Track. As shown in Tab. 2, performance consistently declines from Perception to Comprehension and Adaptation, reflecting the increasing cognitive difficulty of these stages. The Perception track focuses on capabilities such as OCR and temporal localization, where models generally achieve higher accuracies. Performance declines in Comprehension, which requires connecting visual cues with domain knowledge. Adap-

tation yields the lowest accuracies, as it demands high-level reasoning to apply video context to novel scenarios. Proprietary models generally show a smaller performance drop across these stages compared to open-source models, highlighting superior reasoning capability and domain knowledge.

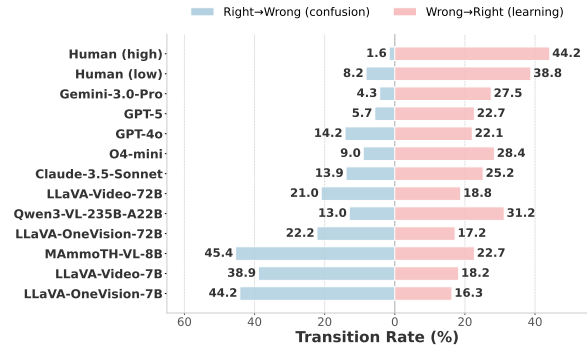
Performance by Discipline. Most models demonstrate relatively higher performance in Humanities & Social Science and Art & Design, where videos focus on conceptual presentation. In comparison, accuracy is lower in Science and Technology & Engineering that require quantitative reasoning and fine-grained visuals such as diagrams. This suggests that current models still struggle in domains that require complex reasoning and visual analysis.

4.3 Performance Gain in Adaptation

Beyond final accuracy, we ask: *how much does a model actually improve from watching the video?* To answer this, we measure the *normalized performance gain* in the Adaptation track, following Hake’s gain metric widely used in educational research (Hake, 1998). A simple analogy is that we let a model take an exam without the video lecture, and reattempt the exam after watching the video lecture, and see how it utilizes the knowledge from the video lecture to benefit its solution. For



(a) Normalized Performance Gain in Adaptation.



(b) Wrong to Right Rate and Right to Wrong Rate.

Figure 4: Pre-video and Post-video Analysis: Normalized Performance Gain and Transition Rate in Adaptation.

each model, we first collect its pre-video accuracy Acc_{pre} (answering the Adaptation question *without* the video) and its post-video accuracy Acc_{post} (answering the same question *after* watching the video). The normalized gain is then defined as:

$$\text{Gain} = \frac{Acc_{post} - Acc_{pre}}{100\% - Acc_{pre}} \times 100\%.$$

The numerator captures the raw improvement, while the denominator represents the maximum possible room for improvement given the model’s starting point. This normalization is critical because raw accuracy gain is not comparable across models with different baselines: improving from 90% to 95% achieves 50% of the remaining gap, whereas improving from 50% to 55% improves only 10%. The normalized gain thus quantifies *learning efficiency*, the fraction of learnable questions that a model actually learns from the video.

4.3.1 Human-Model Gap in Learning Gain

Fig. 4a reveals a substantial gap between humans and models in normalized performance gain after watching the video. Humans achieve a gain of 36.5%, whereas the best-performing model reaches only 13.0%. Importantly, this gap cannot be explained by differences in prior knowledge. Humans (low) start with a pre-video accuracy that is comparable to open-source models such as LLaVA-Video-72B. However, after watching the video, humans achieve a much higher post-video accuracy of 60.33%, matching proprietary models such as Claude-3.5-Sonnet. Therefore, even with normalization that accounts for different baseline performances, the human advantage arises not from stronger prior knowledge, but from a superior ability to transfer video-acquired knowledge to problem-solving.

4.3.2 Wrong to Right and Right to Wrong

A small net gain does not imply that the model leaves its answers unchanged. Performance gain in Adaptation can be decomposed into two types of answer transition after watching the video: incorrect answers that become correct (Wrong to Right) and correct answers that become incorrect (Right to Wrong). Wrong to Right transitions indicate successful learning, whereas Right to Wrong transitions reflect video-induced confusion that overturns previously correct answers. We quantify the effects using two normalized rates. The Wrong to Right rate is defined as:

$$\text{Wrong to Right Rate} = \frac{N_{\text{Wrong to Right}}}{N_{\text{Wrong-before}}} \times 100\%,$$

where $N_{\text{Wrong to Right}}$ denotes the number of questions answered incorrectly before but correctly after the video, and $N_{\text{Wrong-before}}$ denotes the number of incorrect questions before the video. The Right to Wrong rate is defined analogously, using $N_{\text{Right-before}}$ as the denominator. Note that the two rates are normalized by *different* denominators, hence a model with a high Wrong to Right rate is not guaranteed to have a net positive gain; the two rates characterize learning and confusion on different baselines and should be interpreted jointly rather than subtracted directly. As shown in Fig. 4b, models exhibit moderate Wrong to Right rates, indicating some capacity for adaptation, but they also exhibit notably higher Right to Wrong rates than humans, suggesting that video-induced confusion is a frequent failure mode. In contrast, humans achieve a higher Wrong to Right rate and a much lower Right to Wrong rate, reflecting stable learning with minimal video-induced confusion. This suggests that improving adaptation performance

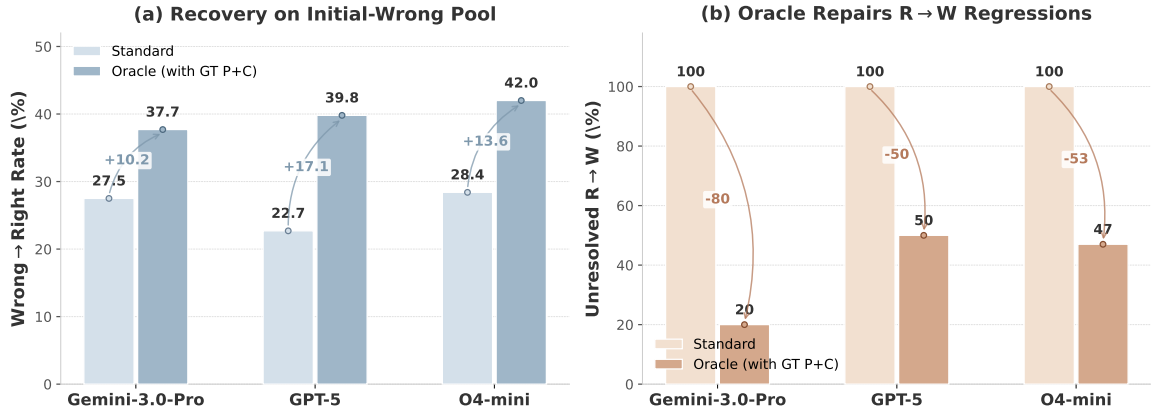


Figure 5: Exploratory oracle study on two informative subsets where standard adaptation already exhibits a clear failure mode. (a) On the wrong-before pool, standard adaptation combined with oracle (GT Perception & Comprehension) recovers 10%–17% more questions than standard adaptation alone. (b) Of the questions that standard adaptation regressed (Right→Wrong), oracle recovers roughly half or more (50–80%). The oracle setting is intentionally non-standard (it exposes ground-truth answers from other tracks) and is reported here only as a controlled probe, not as a recommended evaluation protocol.

requires not only increasing Wrong→Right transitions, but also reducing Right→Wrong transitions. **Sources of Video-Induced Confusion.** To understand why Right→Wrong transitions occur, we analyzed the human baseline, where even top-performing students exhibit a non-zero Right→Wrong rate. We identify two cognitive mechanisms. (1) *Methodological conflict*: A participant may correctly answer a question via qualitative reasoning before the video, but the video then introduces a rigorous quantitative derivation. When adopting this new method, the participant may make errors that override an initially correct judgment. For example, a student who correctly identifies the direction of a force using intuition may fail when switching to the formal vector decomposition demonstrated in the video. (2) *Information distraction*: Academic videos often provide in-depth discussion of edge cases, exceptions, or related concepts, which can create cognitive overload. A participant might then misapply a formula intended for a special case, or confuse properties of similar-looking structures. These patterns are amplified in models: without metacognitive monitoring to preserve high-confidence prior answers, models are more susceptible to both methodological conflict and information distraction, explaining their high Right→Wrong rates.

Metric Validity and Human Alignment. To validate the metric reliability, we conducted a human verification study with human experts. Experts analyzed the model responses to verify whether Wrong→Right transitions represent learning and

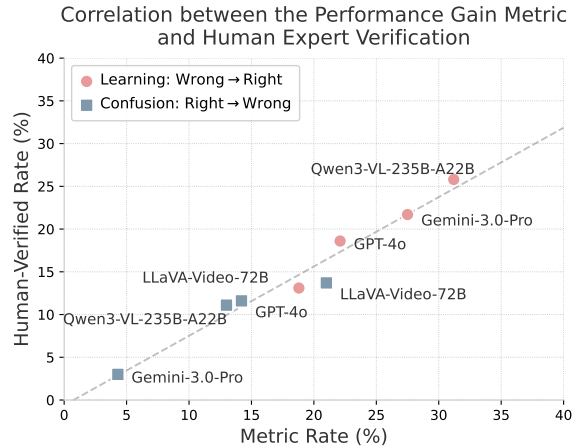


Figure 6: Validation of the performance gain metric.

Right→Wrong represent confusion, as visualized in Fig. 6. We observe that the metric closely correlates to the human-verified rates, confirming that it reflects genuine knowledge acquisition.

4.4 Oracle Study: Does Video Understanding Help Adaptation?

As a preliminary exploration into the role of video understanding in adaptation, we conducted a small-scale oracle study on three top-performing models (Gemini-3.0-Pro, GPT-5, and O4-mini). For each Adaptation question, we additionally provided the ground-truth Perception and Comprehension answers as auxiliary context in the prompt, while keeping the video and the question unchanged. The intent is to ask: *if video understanding errors were removed, what would happen to adaptation?* To

make the analysis informative rather than driven by aggregate accuracy shifts, we focus on two subsets where standard adaptation already exhibits a clear failure mode. The results are shown in Fig. 5.

Oracle expands recovery on the wrong-before pool. We start from the *wrong-before pool* (questions the model could not answer before video) and first let the model attempt the standard adaptation setting. The oracle setting then repairs questions that remained wrong after standard adaptation, yielding additional recovery of +10.2% on Gemini-3.0-Pro, +17.1% on GPT-5, and +13.6% on O4-mini, as shown in Fig. 5a. In other words, for a portion of the questions that the standard adaptation setting cannot solve from the video alone, providing reliable perception and comprehension cues may help the model arrive at the correct adaptation. Nonetheless, a substantial portion of the wrong-before pool remains unsolved even under the oracle setting, suggesting that video understanding is not the sole bottleneck.

Oracle mitigates standard adaptation’s regression. We then examine the questions that the model answered correctly before watching the video but got wrong after standard adaptation (the *Right→Wrong* subset). Oracle recovers roughly half or more of these regressions (Fig. 5b), suggesting that a portion of video-induced confusion may be attributable to incomplete or inaccurate video understanding rather than reasoning errors.

Caveats and scope. We emphasize that this is an exploratory subset-level analysis rather than a definitive claim. The oracle setting deliberately exposes ground-truth answers from other tracks, which violates the standard evaluation protocol and is therefore unsuitable as a general benchmark setting. Moreover, the analysis is restricted to two subsets and three models. We report it as a controlled probe rather than a strong conclusion: video understanding appears to be *one* contributing factor among several (e.g., reasoning depth and domain knowledge) that jointly affect adaptation performance.

4.5 Analysis by Question Type

Beyond aggregate track-level accuracy, we analyzed model performance across four question types in the Comprehension and Adaptation tracks defined in Sec. 3: Concept Comprehension (CC) and Problem-solving Comprehension (PSC), and Case Study Analysis (CSA) and Problem-solving Adaptation (PSA). We omit the Perception track as

Model	Comprehension		Adaptation	
	CC	PSC	CSA	PSA
Gemini-3.0-Pro	84.8	83.7	81.6	78.6
GPT-5	76.6	82.2	73.8	73.0
O4-mini	73.7	81.4	79.4	66.7
Claude-3.5-Sonnet	63.2	69.8	65.2	56.0
GPT-4o	55.0	55.0	68.8	42.8
Qwen3-VL-235B-A22B	74.9	80.6	73.8	66.0
Qwen3-VL-30B-A3B	70.2	76.7	64.5	49.7
LLaVA-Video-72B	59.1	32.6	53.2	35.2
LLaVA-OneVision-72B	57.9	27.1	52.5	34.0
MAmmoTH-8B	46.8	31.8	33.3	32.7
LLaVA-Video-7B	46.8	19.4	41.1	28.3
Llama-3.2-11B	44.4	25.6	28.4	22.0
LLaVA-OneVision-7B	39.8	17.8	36.2	24.5

Table 3: Accuracy (%) by question type for Comprehension and Adaptation.

ASR is a long-tail task. Tab. 3 reports the results.

Comprehension: procedural comprehension as a capability threshold. The most discriminative question type is PSC (Problem-solving Comprehension), which requires following the step-by-step method demonstrated in the video. Proprietary models generally score higher on PSC than CC. In contrast, open-source models score substantially higher on CC than PSC regardless of scale. This suggests that the ability to follow procedural or methodological demonstrations is a threshold capability that separates model tiers.

Scaling effects differ across architectures. Within the Qwen3-VL family, scaling from smaller to larger variants reverses the CC–PSC relationship: Qwen3-VL-30B-A3B scores 6.5% higher on PSC than CC, suggesting an emergent procedural comprehension capability at larger scale. In contrast, the LLaVA-OneVision family shows no such reversal: scaling from 7B (22% gap) to 72B (31% gap) retains the CC advantage, indicating architecture or training-specific limitations in procedural or methodological learning.

Conceptual knowledge transfers more reliably than procedural knowledge. In Adaptation, PSA is consistently harder than CSA across all models, but the gap narrows for stronger models (e.g., Gemini-3.0-Pro: 3.0%) compared to weaker ones (e.g., LLaVA-OneVision-7B: 11.6%). Comparing the Comprehension-to-Adaptation drop within each question type, CC-to-CSA retention rates range from 85% to 100% for most models, indicating that conceptual knowledge acquired in Com-

prehension transfers reliably to Adaptation. In contrast, PSC-to-PSA retention rates are much lower for models, showing that procedural knowledge is substantially harder to transfer to novel scenarios.

4.6 Error Analysis

We manually analyze adaptation errors of Gemini-3.0-Pro on the incorrectly answered Adaptation questions to provide a systematic overview of failure modes. Fig. 7 summarizes the distribution of errors. More case studies for each error type are provided in Appendix 8–12.

Input-level: Visual Extraction Errors (19%).

The model misreads fine-grained visual information from the video or the adaptation question, such as small-font axis labels on a titration curve or numerical values on a circuit diagram (Fig. 11). The reasoning chain is otherwise correct, but a single misread input invalidates downstream computation.

Knowledge-level: Method Understanding Errors (18%). The model misunderstands the knowledge presented in the video and selects an inappropriate method (Fig. 8). In such cases, the subsequent reasoning chain may remain internally consistent, but it starts from a wrong foundation, suggesting that the model does not fully comprehend which method is suitable for the problem.

Reasoning-level: Reasoning and Adaptation Errors (63%). The majority of errors occur at the reasoning level, where the model fails during reasoning. We identify two sub-categories of reasoning errors. (1) *Method Adaptation Errors* (~41%). The model identifies the right method but produces errors when applying it to the new scenario, such as arithmetic mistakes in multi-step calculations or incorrect variable substitution (Fig. 9, 10). (2) *Context Confusion Errors* (~22%). The model conflates the video’s worked example with the new adaptation scenario (Fig. 12). A common pattern is *value leakage*, where the model reuses numerical inputs from the video instead of the values in the Adaptation question. Another pattern is *condition carryover*, where the model applies assumptions from the video that no longer hold in the Adaptation question. These patterns suggest that processing long, multimodal contexts remains challenging for current models: it is difficult for them to separate instructional content (the method or concept to learn) from instance-specific content (the particular example in the video), and locating the correct input values and relevant details from the video to answer the adaptation question can be unreliable. This is

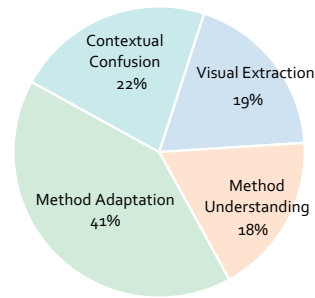


Figure 7: Distribution of Gemini-3.0-Pro errors.

an important capability for video-based learning: a model needs to identify what is essential in the video and what is distracting or less relevant, as irrelevant content introduces cognitive load that interferes with correct reasoning.

4.7 Guide for Future Models

Required capabilities for video learning. Effective video-based learning requires both reliable video understanding and adaptation reasoning. As shown in Sec. 4.4 and 4.6, Perception and Comprehension are necessary for acquiring correct knowledge, while effective reasoning is also critical.

Future directions. *Understanding-wise:* Models should strengthen video perception and comprehension to reduce error propagation from unreliable understanding. More training on lecture-style, information dense videos can be a promising direction for improvement. *Reasoning-wise:* Models should improve adaptation reasoning under video contexts to increase the Wrong→Right transitions. Future models should better control how video information is integrated into reasoning, enabling more stable knowledge adaptation.

5 Conclusion

Video-MMMU systematically evaluates how large multimodal models acquire knowledge from instructional videos across three stages: Perception, Comprehension, and Adaptation. Our results reveal an existing gap between humans and models, not only in final performance but also in learning efficiency from video. Through quantitative analysis, we show that effective video-based learning depends on both video understanding and adaptation reasoning, providing diagnostic insights to guide future research toward more effective video-based learning in multimodal models.

Limitations

Video-MMMU represents an initial effort toward evaluating knowledge acquisition from college-level academic videos and introducing the learning gain as a new perspective. However, Video-MMMU currently focuses exclusively on college-level difficulty, leaving other educational stages and specialized domains (e.g., law) unexplored. In addition, the video curation and QA annotation are manual, resulting in substantial expert labor costs and limited scalability. It is resource-heavy and difficult to scale up. We acknowledge that this manual approach restricts the current size of the benchmark. Additionally, our use of normalized performance gain follows classic educational research (Hake’s Gain). We recognize that this is one possible approach and encourage the community to investigate alternative metrics.

Ethical Considerations

Video-MMMU is developed with a focus on responsible research use and transparent evaluation. All questions are manually constructed to assess domain knowledge, avoiding any personal, sensitive, or biased content. Videos are selected to align with standard academic fair-use practices. The dataset is under CC-BY-NC-SA-4.0 license and is strictly for academic research and non-commercial use. The AI tools are used for grammar checking. For data protection, all human baseline test-taker data were collected without personal, offensive or sensitive content. Participation was anonymous, informed consent was obtained, and the data are not shared publicly to protect participant privacy.

Acknowledgements

This research is supported by cash and in-kind funding from NTU S-Lab and industry partner(s). This study is also supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20223-0002).

References

Richard C Anderson and P David Pearson. 1984. A schema-theoretic view of basic processes in reading comprehension. *Handbook of reading research*, 1:255–291.

Anthropic. 2024. Claude Team. Introducing Claude 3.5 Sonnet. <https://www.anthropic.com/claude/sonnet>.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Benjamin S. Bloom, Max D. Engelhart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. 1956. *Taxonomy of educational objectives: The classification of educational goals, handbook i: Cognitive domain*.

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. 2024. Temporal-bench: Towards fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsen, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*. *Preprint*, arXiv:2507.06261.

Google DeepMind. 2025. Gemini 3 Pro model card. <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf>. Accessed 2025.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. *Mmbench-video: A long-form multi-shot benchmark for holistic video understanding*. In *Advances in Neural Information Processing Systems*, volume 37, pages 89098–89124. Curran Associates, Inc.

Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. 2024. *How far are we from AGI: Are LLMs all we need?* *Transactions on Machine Learning Research*. Survey Certification.

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, Caifeng Shan, and 2 others. 2025. *Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis*. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 24108–24118. Computer Vision Foundation / IEEE.

Jiawei Guo, Tianyu Zheng, Yizhi Li, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Graham Neubig, Wenhui

- Chen, and Xiang Yue. 2025. **MAMmoTH-VL: Eliciting multimodal reasoning with instruction tuning at scale**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13869–13920, Vienna, Austria. Association for Computational Linguistics.
- Richard R. Hake. 1998. Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1):64–74.
- Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, Kevin Lin, William Wang, Lijuan Wang, and Xin Wang. 2025. **Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos**. In *International Conference on Representation Learning*, volume 2025, pages 26931–26964.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, and 397 others. 2024. **Gpt-4o system card**. *ArXiv*, abs/2410.21276.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. **LLaVA-onevision: Easy visual task transfer**. *Transactions on Machine Learning Research*.
- 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. 2024. **Mvbench: A comprehensive multi-modal video understanding benchmark**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.
- Shuai Liu, Shulin Tian, Kairui Hu, Yuhao Dong, Zhe Yang, Bo Li, Jingkan Yang, Chen Change Loy, and Ziwei Liu. 2026. **Filegram: Grounding agent personalization in file-system behavioral traces**. *Preprint*, arXiv:2604.04901.
- Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. **TempCompass: Do video LLMs really understand videos?** In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8731–8772, Bangkok, Thailand. Association for Computational Linguistics.
- Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. **Egoschema: A diagnostic benchmark for very long-form video language understanding**. In *Advances in Neural Information Processing Systems*, volume 36, pages 46212–46244. Curran Associates, Inc.
- Meta. 2024. **Llama 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models**.
- Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2024a. **Position: Levels of AGI for operationalizing progress on the path to AGI**. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36308–36321. PMLR.
- Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2024b. **Position: levels of agi for operationalizing progress on the path to agi**. In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Thomas O. Nelson. 1990. **Metamemory: A theoretical framework and new findings**. volume 26 of *Psychology of Learning and Motivation*, pages 125–173. Academic Press.
- OpenAI. 2025a. **GPT-5 system card**. <https://cdn.openai.com/gpt-5-system-card.pdf>. Accessed 2025.
- OpenAI. 2025b. **Openai o3 and o4-mini system card**. <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>. Accessed 2025.
- Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchet, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, and 5 others. 2023. **Perception test: A diagnostic benchmark for multimodal video models**. In *Advances in Neural Information Processing Systems*, volume 36, pages 42748–42761. Curran Associates, Inc.
- Marija Sablić, Ana Miroslavljević, and Alma Škugor. 2021. **Video-based learning (vbl)—past, present and future: An overview of the research published from 2008 to 2019**. *Technology, Knowledge and Learning*, 26(4):1061–1077.
- Wei Han Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. 2025. **Lvbench: An extreme long video understanding benchmark**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22958–22967.

- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhua Chen. 2024. [Mmlu-pro: A more robust and challenging multi-task language understanding benchmark](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290. Curran Associates, Inc.
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. [Longvideobench: A benchmark for long-context interleaved video-language understanding](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 28828–28857. Curran Associates, Inc.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. 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.
- Zhe Yang, Shulin Tian, Kairui Hu, Shuai Liu, Hoang-Nhat Nguyen, Yichi Zhang, Zujin Guo, Mengying Yu, Zinan Zhang, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2026. [Hippocamp: Benchmarking contextual agents on personal computers](#). *Preprint*, arXiv:2604.01221.
- Ahmed Mohamed Fahmy Yousef, Mohamed Amine Chatti, and Ulrik Schroeder. 2014. The state of video-based learning: A review and future perspectives. *International Journal on Advances in Life Sciences*, 6(3):122–135.
- Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yuet-ing Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134.
- 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, and 3 others. 2024. 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.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhua Chen, and Graham Neubig. 2025. [MMMU-pro: A more robust multi-discipline multimodal understanding benchmark](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, Vienna, Austria. Association for Computational Linguistics.
- Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025a. [LMMS-eval: Reality check on the evaluation of large multimodal models](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. 2025b. [LLaVA-video: Video instruction tuning with synthetic data](#). *Transactions on Machine Learning Research*.
- Yuanhan Zhang, Kaichen Zhang, Bo Li, Fanyi Pu, Christopher Arif Setiadharm, Jingkang Yang, and Ziwei Liu. 2024. [Worldqa: Multimodal world knowledge in videos through long-chain reasoning](#). *arXiv preprint arXiv:2405.03272*.
- Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shang-guan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8475–8489.
- Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13691–13701.

A Dataset Overview

A.1 Overview of Subjects

Discipline	Subjects
Art	Art History
	Art Theory
	Design
	Music
Business	Accounting
	Economics
	Finance
	Management
	Marketing
Science	Biology
	Chemistry
	Geography
	Math
	Physics
Medicine	Basic Medical Science
	Clinical Medicine
	Diagnostics and Laboratory Medicine
	Pharmacy
	Public Health
Humanities	History
	Literature
	Psychology
	Sociology
Engineering	Agriculture
	Architecture and Engineering
	Computer Science
	Electronics
	Energy and Power
	Materials
	Mechanical Engineering

Table 4: Subjects grouped under six disciplines.

A.2 Experimental Settings

Evaluation Framework. To ensure reproducibility and fair comparison, all LMM evaluations are conducted using the LMMs-Eval framework (Zhang et al., 2025a).

Model Configuration and Sampling. For proprietary models (e.g., GPT series (OpenAI, 2025a,b), Gemini-3.0-Pro (DeepMind, 2025)), we utilize the Azure APIs. We sample GPT series models at 50 frames, and process full videos for Gemini-3.0-Pro. When facing API errors, we retry with fallback to 32 frames. For open-source models, we adhere to the recommended sampling strategies. For the Qwen-VL series (Bai et al., 2025), the frame sampling is capped at 768 frames. For other models, we apply uniform sampling with the following frame counts: 64 frames for LLaVA-Video (Zhang et al.,

2025b); 32 frames for LLaVA-OneVision (Li et al., 2025) and MAMmoTH-VL (Guo et al., 2025).

Answer Extraction via LLM Judge. We use Gemini-2.5-Flash-Lite (Comanici et al., 2025) as the LLM judge to extract the final answer from each model’s free-form response. The judge is called with greedy decoding (temperature = 0) and a maximum output length of 50 tokens. The extraction prompt varies by question type:

Multiple-choice questions:

```
You are extracting the final answer from a model's response to a multiple choice question.
Question: {question}
Model's Response: {prediction}
Extract ONLY the option letter (A, B, C, or D) that the model chose as its final answer.
Output ONLY the single letter, nothing else.
If the model response is 'No Answer', API Error or empty string, return exactly: "No Answer".
```

Open-ended questions:

```
You are extracting the final answer from a model's response to an open-ended question.
Question: {question}
Model's Response: {prediction}
Extract ONLY the final answer from the response.
Rules:
- Strip currency symbols, units, and commas from numbers
- Convert number words (e.g., "one", "two") to digits
- For fractions, keep as fraction form
- Output ONLY the raw answer value, nothing else.
```

The extracted answer is then compared against the ground truth using exact match, followed by a rule-based post-processing step to normalize currency symbols and units. For model inputs, we follow a standard protocol that concatenates sampled video frames, the question, and the answer options into the input. The evaluations are validated with multiple runs, and we report the average value.

A.3 Dataset Scale and Comparative Advantages

We have verified the reliability of results. In this section, we discuss the scale of Video-MMMU and highlight three distinct advantages that differentiate it from existing video benchmarks.

Dataset Scale and Annotation Effort. Video-MMMU distinguishes itself through its broad disciplinary coverage and substantial temporal scale.

Spanning 6 disciplines and 30 subjects, the benchmark comprises a total video duration exceeding **42.18 hours**. Unlike general video captioning, the annotation of academic video content demands high-level domain expertise. Consequently, the annotation process required over 217.5 hours of expert labor. We believe this combination of broad coverage, long-form duration, and rigorous expert annotation constitutes a sufficiently large and diverse data source for evaluating LMMs.

Unique Advantages. Beyond scale, Video-MMMU offers three critical advantages designed to push the boundaries of current video LMM research.

Distinct Task Domain. Video-MMMU is the first benchmark specifically targeting academic video learning. Unlike existing datasets that focus on everyday activities or entertainment content, our benchmark evaluates the model’s ability to process dense, professional knowledge. This capability is critical for deploying LMMs in real-world educational applications, such as intelligent tutoring systems.

Acquisition-based QA Design. The QAs are structured into three distinct tracks that mirror human cognitive learning processes. Furthermore, the significantly longer QA length in Video-MMMU reflects the higher complexity inherent in academic evaluation, requiring models to engage in more sustained reasoning rather than simple pattern matching.

Evaluation of Knowledge Acquisition. Video-MMMU is among the first to explore Learning Gain from video, quantifying it through performance gain metrics (e.g., Net Gain). This approach highlights that videos serve not merely as visual signals to be described, but as rich sources of knowledge. This marks a pivotal step from mere video understanding towards acquiring knowledge from videos.

B Metric Validity and Human Alignment

B.1 Human Verification Protocols

To rigorously validate the proposed Normalized Performance Gain metric, we conducted a “Reasoning Verification” study with a group of PhD experts. Unlike simple output scoring, experts analyzed the Chain-of-Thought (CoT) reasoning traces to distinguish genuine knowledge acquisition from statistical noise. Model behaviors were classified into four categories:

Verified Learning (Genuine): The model correctly changes from Wrong (Pre) to Right (Post), *and* the reasoning explicitly utilizes correct information from the video.

Lucky Guess (Spurious): The model shifts to the correct answer post-video, but the supporting reasoning is hallucinated, logically invalid, or merely matches keywords without genuine comprehension. Our metric records this as a gain, whereas human verification rejects it due to lack of grounding.

Verified Confusion (Genuine): The model changes from Right to Wrong because it was actively misled by specific video content.

Spurious Refusal / Response Collapse: The model shifts from a correct answer to a less informative state (e.g., “I cannot answer,” repetition, or empty output). This regression stems from over-sensitive alignment or long-context instability, rather than genuine video-induced confusion.

B.2 Quantitative Results

We sampled responses from four representative models. Tab. 5 and Tab. 6 present the comparison between Automated Metric results and Human-Verified results. The human-verified rate is calculated as the average value of the pool of human experts. The automated metric closely tracks the human-verified rates at the system level, confirming that it is a reliable measure aligned with human judgment.

Model	Metric (%)	Human Verified (%)
Gemini-3.0-Pro	27.5	21.7
Qwen3-VL-235B	31.2	25.8
GPT-4o	22.1	18.6
LLaVA-Video-72B	18.8	13.1

Table 5: **Learning Assessment** (Wrong → Right). Comparison of Automated Metric vs. Human-Verified rates.

Model	Metric (%)	Human Verified (%)
Gemini-3.0-Pro	4.3	3.0
Qwen3-VL-235B	13.0	11.1
GPT-4o	14.2	11.6
LLaVA-Video-72B	21.0	13.7

Table 6: **Confusion Assessment** (Right → Wrong). Comparison of Calculated Metric vs. Human-Verified rates.

B.3 Analysis of Discrepancies

Based on Tab. 5 and 6 (visualized in Fig. 6), we analyze two key patterns regarding the divergence between automated metrics and human verification:

Learning Alignment (Blue Points). SOTA models such as Gemini-3.0-Pro exhibit minimal discrepancies between automated and human outcomes, indicating that their performance gains indeed stem from genuine, video-grounded reasoning. This substantial gap confirms that weaker models frequently rely on spurious lucky guesses, where the model hits the correct answer via hallucinations or superficial keyword matching rather than actual comprehension. Overall, the metric is validated by human experts and aligns with human preference.

Confusion Alignment (Red Points). The automated metric accurately reflects genuine video-induced confusion. Across the evaluated models, the automated and human-verified rates track closely, with the largest discrepancies concentrated on weaker models where spurious refusals and response-collapse behaviors inflate the automated R→W count relative to genuine confusion.

C Annotation Pipeline

C.1 Video Curation

We will go through the standardized protocol used by annotators to curate instructional videos for Video-MMMU. The guideline specifies (i) how videos are sourced and filtered, and (ii) how candidate videos are evaluated using a concrete, multi-dimensional scoring rubric. The objective is to ensure that all selected videos reliably support knowledge acquisition across the Perception, Comprehension, and Adaptation stages.

Annotators curate videos following a curriculum-driven pipeline. Each candidate video must correspond to a predefined Exam Point in Appendix D and undergo multi-stage filtering before scoring.

C.1.1 Step 1: Exam-Point Matching

For each Exam Point, annotators retrieve candidate educational videos based on the following criteria.

- 1) Each video must be mapped to exactly one Exam Point.
- 2) The Exam Point must be explicitly covered and substantively explained or demonstrated in the video via definition, derivation, or worked example questions.

Dimension	Score	Operational Decision Criteria
Visual Clarity	1	Critical visual content (e.g., formulas, subscripts, axes labels, handwritten text) is unreadable at $\geq 720p$, even when the video is paused.
	2	Most visual content is readable at $\geq 720p$, but some critical details are briefly blurred, occluded, or require guessing.
	3	All critical visual content is consistently readable at $\geq 720p$ without ambiguity or guesswork.
Technical Depth	1	Intuition-only or descriptive explanation; no equations, derivations, or explicit problem-solving steps.
	2	Correct but introductory-level explanation.
	3	College-level lecture or tutorial with explicit derivations, equations, or step-by-step solution procedures.
Exam-Point Alignment	1	Video mentions a related topic but does not explain or demonstrate the specified Exam Point.
	2	Exam Point is mentioned or partially illustrated, but not developed in sufficient detail.
	3	Exam Point is explicitly explained or demonstrated with sufficient depth.
Retention Rule	–	Retain the video if and only if the dimensions receive a score of 3.

Table 7: Summary of the video curation scoring rubric.

- 3) Videos that only mention the Exam Point without explanation must be rejected.

C.1.2 Step 2: Educational Video Sourcing

Preferred videos include university instructor-led tutorials that provide detailed explanations or step-by-step example questions with derivations.

Discouraged videos include crash courses, high-level summaries, or entertainment-oriented content without formal structure.

C.1.3 Video Scoring Rubric

All candidate videos passing the sourcing stage are evaluated using the three-dimensional rubric summarized in Tab. 7. Videos are independently scored along visual clarity, technical depth, and alignment with the exam point on a 1–3 scale.

C.1.4 Retention Policy and Quality Control

We enforce a strict *Perfection-or-Rejection* policy.

- 1) A video is retained if and only if it receives a score of 3 on all three rubric dimensions.
- 2) Any score below 3 in any dimension results in rejection.

C.2 Initial Annotation Guideline

This guideline specifies how annotators construct QA pairs for each instructional video. For every video, annotators must create exactly one QA pair for each cognitive track: Perception, Comprehension, and Adaptation.

1) Perception QA. Construct a question that targets information directly observable from the video, including on-screen text, numerical values, symbols, diagrams, or spoken content (OCR or ASR). The correct answer must rely only on visual or auditory cues and must not require conceptual reasoning. Reject questions that can be answered without watching the video.

2) Comprehension QA. Construct a question that tests understanding of concepts or procedures explicitly explained in the video. The question should require interpretation or evaluation of the presented knowledge rather than surface recall. Concept comprehension (e.g., MAMC) and problem-solving strategy comprehension with modified inputs are both acceptable.

3) Adaptation QA. Construct an exam-style question that applies the same concept or problem-solving strategy taught in the video to a new scenario. Do not reuse the original example, inputs, or numerical values shown in the video.

Options and Answer. Each multiple-choice question may include up to 10 options with exactly one correct answer. All options should be plausible, mutually exclusive, and closely related to the target concept. Clearly mark the correct answer and ensure it is unambiguous based on the video.

C.2.1 Annotator Compensation and Workload

All annotations in Video-MMMU were conducted by graduate or PhD students in the corresponding academic disciplines. Annotation was performed on a paid, time-controlled basis with explicit workload limits to ensure sustainable working conditions.

Annotators were compensated using a hybrid scheme that combines a base hourly rate with per-task bonuses for annotation and validation. Each

Aspect	Details
Annotator background	Graduate and PhD students in relevant disciplines
Base compensation	20 USD per hour
Per-video bonus	+2 USD per completed annotation (1 video = 3 QAs)
Per-validation bonus	+1 USD per validated video (3 QAs)
Avg. annotation time	35 min 10 s per video (3 QAs)
Avg. validation time	8 min 20 s per video
Daily workload cap	Max 10 annotations or 20 validations
Total annotation effort	>217.5 total annotation hours

Table 8: Summary of annotator compensation and workload.

annotation corresponds to one video paired with three QA items. Daily workloads were explicitly capped to reduce fatigue and maintain annotation quality.

Due to the academic complexity of the dataset and the requirement for multi-stage quality control, the construction of Video-MMMU required more than 217.5 total annotation hours. Detailed annotation protocols and quality-control procedures are provided in Tab. 8.

C.3 Review and Cross-Review Guideline for QA Annotation

This section provides the operational guideline for the three-stage *Rejection–Refinement Loop* described in the main paper: self-verification, inter-annotator cross-review, and expert validation. Reviewers must follow the format and scoring rules below.

Review Format

Track: [Perception / Comprehension / Adaptation]

Exam Point: [Exam Point Name]

Video: [Video Link]

Question: [Question Text]

Options: [Option List]

Answer: [Correct Option]

Stage 1: Self-Verification (Annotator)

Annotators must self-check each QA before submission using three rubric dimensions.

1) Clarity (1–3).

3 = question is grammatically correct and unambiguous; a reader can understand it on first read.

2 = understandable but requires minor edits (wording, phrasing, or option format).

1 = unclear, ambiguous, or underspecified (e.g.,

missing condition, unclear referent, vague wording).

2) Technical Correctness (1–3).

3 = correct answer is fully supported by the video; no factual or conceptual errors.

2 = mostly correct but with minor imprecision (e.g., notation mismatch, missing assumption) that may confuse readers.

1 = incorrect, not supported by the video, or relies on external knowledge not taught in the video.

3) Reasoning Depth (1–3).

3 = the QA matches the intended cognitive track and requires the expected level of processing (Perception: extraction; Comprehension: understanding; Adaptation: transfer).

2 = partially matches the intended depth.

1 = violates the track definition (e.g., Perception requires inference; Adaptation is identical to the video example).

Stage 2: Inter-Annotator Cross-Review (Independent Annotator)

A different annotator performs a blind cross-review using the same rubric and evidence requirement.

Stage 3: Expert Validation (PhD Domain Experts)

PhD domain experts perform the final validation focusing on technical rigor and Exam-Point alignment.

Expert Checks.

1) Technical validity. Confirm that the correct answer is technically correct under standard domain conventions (notation, definitions, units, and assumptions) and consistent with the video explanation.

2) Exam-Point alignment. Confirm that the QA targets the intended Exam Point and that solving the QA relies on the knowledge unit defined by that Exam Point.

3) Track alignment. Confirm that the QA matches the intended track definition (Perception/Comprehension/Adaptation) and does not leak information across tracks (e.g., Perception requiring reasoning; Adaptation becoming a copy of the example).

Expert Decision. Experts may mark a QA as Accept or Refine. Any QA marked Refine must be revised and re-validated until consensus is reached.

C.4 Instructions for Video-MMMU Human Participants

Thank you for participating in the Video-MMMU evaluation. This task simulates how students learn by watching educational videos. Please follow the steps below carefully.

Step 1: Pre-Video Question (Adaptation)

You will first be shown one question without watching the video.

This question tests your existing knowledge of the topic.

Please answer it based on what you already know.

Step 2: Watch the Video

Next, you will watch a lecture video related to the topic.

Watch the entire video attentively, as if you are attending a real class.

Take mental notes if needed, as this will help with the next questions.

Step 3: Answer Three Post-Video Questions

You will then answer three questions, each targeting a different aspect of learning:

1) Perception Question

Tests your ability to recall specific details directly shown in the video.

2) Comprehension Question

Tests your understanding of key concepts and ideas presented in the video.

3) Adaptation Question (Reattempt)

You will see the same question from Step 1.

Answer it again using the information you learned from the video.

Final Notes

You may refer to your course notes and other personal study materials. However, you are prohibited from searching for answers online.

All responses will remain anonymous. We will not share or distribute your raw answers. Your response will be used solely to compute aggregate results for academic research purposes. By proceeding with this task, you acknowledge that you have read and agree to the above instructions. Thank you for your participation.

C.5 Recruitment

Human participants were recruited to establish both average and upper-bound human performance baselines. We first recruited undergraduate students from relevant universities as part-time participants,

each drawn from the corresponding disciplines listed in Appendix A. These participants represent average human learners. All human participants are English speakers, ensuring they do not have a language barrier. They were compensated at an hourly rate of 20 USD.

We observe that the average human performance before watching videos does not exceed that of top-performing models. This indicates that, prior to video exposure, human participants and models start from a comparable performance level, mitigating concerns that the observed human performance gains are driven by higher prior knowledge rather than learning from video.

To estimate an upper bound on human performance, we additionally recruited top-performing undergraduate students from each subject area. These participants serve as a high-competence reference group. As a result, each QA item is evaluated with at least two human samples, covering both average and top-performing participants.

Discipline	Per Group	# Groups	Total
Art	4	2	8
Business	5	2	10
Medicine	5	2	10
Science	5	2	10
Humanities	4	2	8
Engineering	7	2	14
All	30	2	60

Table 9: Human participant distribution across disciplines.

D Exam Points Summary

Art

Art: Elements of Art

Art Theory: Art Historical Period Identification, Post-Impressionism, Baroque Art, Baroque Period, Stylistic Characteristics of Baroque Art, Stylistic Characteristics of Baroque Painting

Design: Principles of Design, Color Schemes, Classification of Sleeves

Music: Time Signature and Measures, Musical Interval Identification, Technical Names of Scale Degrees, Functional Harmony, Types of Seventh Chords, Inversions of Seventh Chords

Business

Accounting: Cost Behavior Analysis, Statement of Retained Earnings, Net Present Value (NPV), Equivalent Units under the Weighted-Average

Method, Analysis of Cash Transactions, Prime Cost and Conversion Cost, Cost Estimation using Regression Analysis, Arbitrage Pricing Theory (APT), Information Ratio, Arbitrage Pricing Theory, Break-Even Analysis, Debt→Equity Ratio

Economics: Money Multiplier, Price Ceiling, National Saving and Investment Identity, Inflationary Gap and Long-Run Self-Adjustment, Arc Elasticity of Demand, GDP Calculation using the Expenditure Approach, Natural Monopoly

Finance: Arbitrage Pricing Theory (APT), Financial Ratio Analysis, Calculating Additional Funds Needed (AFN) with Excess Capacity, Internal Rate of Return (IRR) Calculation, Fisher Effect, Bond Valuation, Arbitrage Pricing Theory, Discriminatory Auction, Free Cash Flow to Equity (FCFE) Calculation, Discounted Payback Period

Management: Cash Receipts Budget, Trend Projection, Joint Cost Allocation using the Net Realizable Value (NRV) Method, Belbin's Team Roles, Calculating Brokerage Commission

Marketing: Analysis of Variance (ANOVA), Paired Samples T-Test, Chi-Squared Test of Independence, Degrees of Freedom in One-Way ANOVA, Conditional Probability for a Uniform Distribution

Science

Chemistry: Phase Diagram, Markovnikov and Anti-Markovnikov Rules, Acid-Base Titration Curves, Molarity Calculation, Polyprotic Acid-Base Titration, VSEPR Theory, Newman Projections of Meso Compounds, Identification of Chirality Centers

Geography: Internal Torque in Torsion, Influence Lines, Shear Force and Bending Moment Diagrams, Area Moment of Inertia of Composite Shapes

Math: Tangent-Secant Theorem, Permutation Group, Hamiltonian Cycle, Convexity and Concavity of Functions, Dynamic Programming, Related Rates, Optimization Problems, Breadth-First Spanning Tree, Prim's Algorithm, Projectile Motion

Physics: Kepler's Laws of Planetary Motion, Definition of Torque, Thin-Film Interference, Charging by Induction, Displacement from Velocity-Time Graphs, Snell's Law and Refractive Index, Thin Film Interference, Photoelectric Effect: Kinetic Energy vs. Frequency, Graphical Analysis of the Photoelectric, Variation of Strong Nuclear Force with Nucleon Separation

Medicine

Basic Medical Science: Neural Mechanisms of Sound Localization, Osteon (Haversian System), Histological Identification of Tissues, Differences between Prokaryotic and Eukaryotic Cells, Central Vestibular Dysfunction, Histology of Arteries, Pathophysiology of Asthma, Phases of an Action Potential, Hydrogen Bonding, Cardiac Muscle Histology, Innervation of Extraocular Muscles, Internal capsule and corona radiata

Clinical Medicine: Smith's fracture, Pelvic X-ray Interpretation, Ataxia-Telangiectasia, Erythrocyte Sedimentation Rate (ESR), Cardiomegaly on Chest Radiograph, Sensory Innervation of the Tongue, Approach to a Thyroid Nodule

Diagnostics and Laboratory Medicine: Histopathology of Oligodendroglioma, Grading and Staging of Cancer, 2021 WHO Classification of Central Nervous System Tumors, Subependymoma, Normal Radiographic Anatomy of the Canine Thorax, Negri bodies, Risk Factors for Neonatal Intraventricular

Pharmacy: IUPAC Nomenclature of Halocarbons, Classification of Organisms by Carbon and Energy Source, Classification of Chemical Reactions, Muscarinic Antagonists for COPD, Law of Definite Proportions, Acid-Base Titration Curves, Drug-Receptor Binding Interactions, Thermal Decomposition of Ammonium Carbonate, Predicting Products of Electrolysis

Public Health: Chi-squared test, Vaccine Efficacy Calculation, Sensitivity and Specificity, Attributable Risk, Measures of Association: Odds Ratio and Attributable Risk, Odds Ratio Calculation in Case-Control Studies, Calculation of Odds

Humanities

History: Historical Source Analysis, The 6 Cs of Historical Source Analysis, Dynastic Styles of Hanfu, River Valley Civilizations, Identification of Historical Architectural Styles, Historical Architectural Styles, Divisions of the Stone Age, Paleolithic Era, Forms of Government, Core Tenets of World Religions

Literature: Simile and Metaphor, Freytag's Pyramid, Elements of Plot, Schools of Literary Criticism, Literary Genres and Subgenres, Poetic Devices

Sociology: Cultural Universals, Marxist Theory of

Social Class, Sociological Theories of Social Class, Sociological Theories of Culture, Sociological Perspectives on Culture, Theoretical Views of Trade, Theories of Global Development, Theoretical Approaches to Globalisation, Modernization Theory's view on urbanization

Psychology: Emotion-Focused Coping, Structure and Function of Neurons, Schedules of Reinforcement, Mediational Processes of Social Learning Theory, Types of Reinforcers, Psychological Perspectives, Types of Research Methods in Psychology

Engineering

Agriculture: Plant Virus Symptoms, Early Blight vs. Late Blight, Plant Pathogens, Disease Cycle of Grapevine Downy Mildew

Architecture and Engineering: Stadia Tacheometry: Inclined Sight with Vertical Staff, Analysis of Three-Hinged Trussed Arches, Influence Lines for Trusses, Flight Planning for Aerial Photography, Bulk Specific Gravity of Combined Aggregates, Earthwork Volume Calculation by Prismoidal Method, Area Calculation by Double Meridian Distance (DMD) Method, Maximum Dry Unit Weight and Optimum Moisture Content, Reduction to Center, Finite Summation Method for Beam Deflection, Reservoir Capacity Calculation using the Prismoidal Formula, Cross-Sectional Area of Earthworks on Sloping Ground, Three-Point Resection Problem

Computer Science: Forward Propagation in Neural Networks, Bias-Variance Tradeoff, Ford-Fulkerson Algorithm, Minimization of Deterministic Finite Automata (DFA), Leave-One-Out Cross-Validation for K-Nearest Neighbors, Insertion into a 2-3 Tree, Deletion from a Leftist Heap, Function of the Network Layer, DFA Minimization, Maximum Flow Problem, Prim's Algorithm, Discovery and Finishing Times in Depth-First Search, Cost Function for Linear Regression, Conflict Serializability, Kruskal's Algorithm for Minimum Spanning Tree, Serializability

Electronics: Step Response of an RC Circuit, Transient Response of First-Order RC Circuits, Fourier Series of a Half-Wave Rectifier, Equivalent Impedance, DC Steady-State Analysis of First-Order RL Circuits, Transistor Saturation Region, Root Mean Square (RMS) Value of Periodic Waveforms, Ramp Response of a First-Order RL Circuit, BJT Voltage-Divider Bias Circuit Analysis, Output

Resistance of Common-Collector Amplifier, AC Voltage Divider, AC Steady-State Analysis, Fourier Series, Fourier Series of a Sawtooth Waveform, Mesh Analysis for AC Circuits, Nodal Analysis, DC Analysis of BJT Fixed-Bias Circuit, Laplace Transform of Periodic Functions, Signal Representation using Unit Step Functions

Energy and Power: Heat Pump Coefficient of Performance (COP), Hydrostatic Forces on Curved Surfaces, Hydrostatic Force on Submerged Inclined Surfaces, Absolute and Gage Pressure, First and Second Law Efficiencies, Differential Manometer, Adiabatic Process, Manometry, Three-Reservoir Problem, Stagnation Pressure, Bernoulli's Equation, Flow Rate and Average Velocity from Velocity Profile

Materials: Average Normal Stress in Truss Members, Average Normal Stress, Transformed Section Method, Normal Strain, Miller Indices for Directions, Shear Stress and Strain, Equilibrium of Frames and Machines, Shear Modulus, Average Normal Stress under Axial Loading, Ductile→Brittle Transition Temperature, Axial Deformation, Average Normal Strain

Mechanical Engineering: Selection of the Main View, Transfer Function of a Two-Degree-of-Freedom Rotational System, Darcy-Weisbach Equation and Moody Diagram, Rectilinear Kinematics: Acceleration as a Function of Velocity, Orthographic Projection, Routh-Hurwitz Stability Criterion, Kinematics of a Particle: Acceleration as a Function of Velocity, Newton's Law of Universal Gravitation, Axial Stiffness of Composite Bars, Stress Concentrations, Orthographic Projection from an Isometric View, Orthographic Projection from Isometric View, Strength of Parallel-Loaded Fillet Welds, Equivalent Torsional Spring Constant

E Additional Discussions

E.1 A Metacognitive Perspective on Human's High Learning Efficiency.

To interpret why humans exhibit higher learning efficiency from videos compared to current models, we draw on Metacognitive Monitoring Theories. Our analysis offers an exploratory explanation for why humans outperform current models in video-based learning, aiming to provide insights that may inform future model development.

Schema-based knowledge integration limits Right→Wrong reversals. Cognitive schema the-

ory (Anderson and Pearson, 1984) posits that new information is not memorized word by word, but rather integrated into existing mental frameworks, or schemata. When watching lecture-style videos, human learners engage in active self-explanation, interpreting each graphic or worked example in relation to their prior knowledge. They selectively update their schemata only when inconsistencies arise. This gradual integration-not-replacement process enhances performance gain, as the updated schema becomes more complete, while also minimizing Right→Wrong reversals by preserving correct prior knowledge. In contrast, current LMMs lack explicit mechanisms for schema updating. They tend to encode new video content independently of existing representations, which can overwrite correct associations and lead to the high Right→Wrong rates as observed in Tab. 2.

Metacognitive monitoring calibrates when and how knowledge is revised. According to metacognition theory (Nelson, 1990), humans continually evaluate the confidence and accuracy of their own answers while learning. When a silent video introduces unfamiliar material, learners detect low-confidence regions, allocate extra attention, or mentally “rewind” by re-inspecting key frames; conversely, they leave high-confidence knowledge untouched. This strategic self-regulation yields a large Wrong→Right gain while simultaneously shielding high-confidence correct answers from being reversed. In contrast, most LMMs lack real-time confidence calibration. They process visual inputs in a single forward pass without self-monitoring, allowing new evidence to overwrite earlier answers indiscriminately, which undermines net learning gain. Our findings suggest that incorporating such metacognitive dynamics could be a promising direction for future model development.

E.2 Broader Implications: Cognitive Evaluation from an AGI Perspective

While Video-MMMU is primarily a benchmark for evaluating multimodal learning from video, its emphasis on knowledge acquisition and learning efficiency also connects with broader trends in AGI research. As general-purpose AI systems move beyond static, task-specific capabilities, the ability to learn, adapt, and evolve from experience is increasingly emphasized in recent AGI roadmaps. Video-MMMU's cognitive structure and learning-oriented evaluation naturally align with several emerging directions in the field.

Cognitive capability as a foundation of AGI. "How Far Are We From AGI?" (Feng et al., 2024) outlines a structured three-layer framework for AGI development: internal (cognitive mechanisms), interface (modality-level integration), and system (coordination and learning over time). At the internal level, they highlight perception, reasoning, memory, and cognition as essential components, each of which maps closely to our benchmark’s structure. Video-MMMU explicitly follows the cognitive learning process of human, as outlined in Bloom’s taxonomy (Bloom et al., 1956): Perception, Comprehension, and Adaptation tracks, which collectively assess these core faculties in a stepwise manner. This supports Video-MMMU’s alignment for studying cognitive capability in models.

Adaptivity through real-world instructional video understanding. Additionally, the authors (Feng et al., 2024) further stress that future AGI systems must exhibit adaptivity by integrating multimodal input, retaining memory, and generalizing across contexts. Video-MMMU embodies this principle by evaluating a model’s ability to learn from instructional videos and apply knowledge to novel exam-style questions. The task design mimics how humans attend to lectures, monitor comprehension, and transfer knowledge to problem-solving scenarios, encouraging models to process information more actively and flexibly, rather than relying on static memorization.

Learning efficiency as an evaluation to self-evolution. At the highest level of their AGI roadmap (Feng et al., 2024), they propose self-evolution as the defining trait of “Ultimate AGI.” In this context, Video-MMMU offers a stepping stone by simulating in-context knowledge acquisition from video, where models must update internal representations based on observed content. Our performance gain metric directly measures this learning efficiency, highlighting how much a model improves after watching a video. This reflects an early form of model-driven adaptation and aligns with AGI’s broader goal of enabling systems to evolve cognitively from interaction with the world. Videos serve as the model’s “eyes” and evaluating how well it learns from them provides insight into its self-improving potential. Hence Video-MMMU provides valuable insights for how models can learn as it operates, and provides a new perspective to measure such capabilities through a VideoQA setting.

Human-like learning and metacognition. A com-

plementary view is presented in the Levels of AGI position paper (Morris et al., 2024b), which offers concrete case studies for evaluating AGI capabilities. Case Study 4 argues that AGI must solve cognitive tasks similar to those performed by humans, particularly abstract reasoning and conceptual problem solving. Video-MMMU reflects this aim by assessing models across three cognitive stages of human learning, making the evaluation setting intuitive and naturally aligned with human intellectual behavior. Case Study 5, drawing from Shanahan’s definition, highlights metacognition as a critical attribute of AGI. Our findings in Video-MMMU show that humans exhibit high Wrong→Right gains and low Right→Wrong reversals, likely due to metacognitive strategies such as monitoring uncertainty and preserving schema. These results provide insights into the abilities that current LMMs lack. This underscores the benchmark’s value in exposing gaps between human and model learning processes.

In summary, Video-MMMU supports several goals shared by the AGI community, including the evaluation of cognitive capability, adaptive learning, and self-improvement. By grounding these evaluations in real-world, information-dense videos, the benchmark offers a concrete and cognitively motivated framework for measuring progress beyond static accuracy. We believe Video-MMMU contributes meaningfully to the broader effort to design general-purpose, self-improving AI systems grounded in human-like learning efficiency.

F Error Case Studies

This section presents detailed case studies of representative errors across all three cognitive tracks, as well as a successful learning example. We first examine the four categories of Adaptation errors identified in Sec. 4.6, then present error cases from the Perception and Comprehension tracks. Finally, we analyze a Wrong→Right example demonstrating successful knowledge acquisition.

F.1 Adaptation Track: Method Understanding Error

Fig. 8 presents a case where the model misunderstands the problem-solving method demonstrated in the video. Although the model correctly perceives the visual content and identifies the relevant topic, it confuses the specific method with a superficially similar alternative. The resulting error is

systematic: the model applies a coherent but incorrect procedure, producing a plausible answer that differs from the ground truth. This type of error is particularly challenging to detect because the model’s reasoning chain appears logical and well-structured, yet is built on an incorrect methodological foundation extracted from the video.

F.2 Adaptation Track: Method Adaptation Error

Fig. 9 illustrates a case where the model correctly identifies the method from the video but fails during the adaptation process. The model successfully extracts the general approach and begins applying it to the new scenario, but introduces errors when substituting new parameter values or executing intermediate computation steps. In this example, the model’s chain-of-thought shows the correct formula and setup, but a computational mistake in an intermediate step propagates through the remaining derivation, leading to an incorrect final answer. This highlights that even when comprehension is successful, the transfer of procedural knowledge to novel inputs remains error-prone.

Adaptation Track: Method Adaptation Error.

Fig. 10 provides a side-by-side comparison of a model and a human learner on the same adaptation question, before and after watching the same video. The question asks about valid DFS forests when discovery/finishing time stamps are modified in a graph with cycles. Before watching the video, both the model and the human answer incorrectly—both rely on surface-level connectivity patterns rather than DFS principles. After watching the video, both *verbalize* the correct principle (“a root node’s interval must contain its descendants’ intervals”), suggesting that the core concept was acquired from the video. However, only the human successfully adapts this principle to the new graph, identifying that c ’s interval $[1/10]$ contains the intervals of g , f , h , and d , and arriving at the correct answer. The model, despite articulating the same principle, fails to systematically apply interval containment to the complex graph and instead defaults to local connectivity heuristics, producing an incorrect grouping. This contrast highlights that articulating a principle is not equivalent to correctly adapting it under the structural complexity of a new scenario, and suggests that method adaptation is a distinct capability from method comprehension.

F.3 Adaptation Track: Visual Extraction Error

Fig. 11 shows a case where the model misreads fine-grained visual information from either the video frames or the adaptation question’s accompanying image. The model’s reasoning logic is otherwise sound, but it operates on incorrect input values extracted from the visual content. Common triggers include small font sizes in axis labels, overlapping annotations in diagrams, and ambiguous visual elements that require precise spatial reasoning. In this case, the misreading of a single value leads to an entirely different computation path, demonstrating how input-level errors can cascade into fundamentally wrong answers despite correct reasoning.

F.4 Adaptation Track: Context Confusion Error

Fig. 12 demonstrates a context confusion error where the model conflates specific details from the video’s worked example with the new adaptation scenario. Rather than abstracting the general method from the video and applying it with the new question’s parameters, the model carries over instance-specific values or conditions from the video. This *value leakage* pattern—where the model substitutes the video example’s numerical inputs instead of the adaptation question’s values—suggests that current models have difficulty separating the instructional content (the method to be learned) from the instance-specific content (the particular example used to demonstrate it) within the same video context.

F.5 Perception Track Error

Fig. 13 presents a Perception error where the model fails to accurately extract information directly presented in the video. Unlike Adaptation errors, which involve higher-order reasoning, Perception errors reflect fundamental limitations in the model’s ability to process video content at the information extraction level. In this case, the model either misidentifies a visual element, misreads text rendered in the video frames, or fails to localize the relevant temporal segment. These errors indicate that even basic video understanding—the first stage in the knowledge acquisition pipeline—remains imperfect, and errors at this stage can prevent successful comprehension and adaptation downstream.

F.6 Comprehension Track Error

Fig. 14 shows a Comprehension error where the model fails to correctly understand the underlying concept or principle presented in the video. The model may correctly perceive the visual content (passing the Perception stage) but draws an incorrect conclusion about the knowledge being conveyed. In this example, the model misinterprets the relationship between concepts demonstrated in the video, leading to an incorrect selection in the multiple-answer multiple-choice format. This type of error is especially common in Concept Comprehension questions, where the model must synthesize information across multiple video segments to form a coherent understanding.

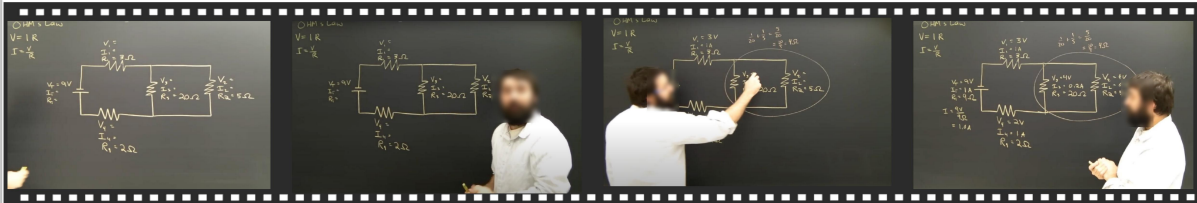
F.7 Wrong→Right Case Analysis

Fig. 15 presents a positive example where the model initially answers the Adaptation question incorrectly without the video (pre-video), but answers correctly after watching the video (post-video). The model's pre-video response reveals reliance on parametric knowledge that is insufficient or incorrect for this specific problem. After watching the video, the model's chain-of-thought explicitly references the method demonstrated in the video and correctly applies it to the new adaptation scenario, providing evidence of genuine knowledge acquisition. This case validates that the Wrong→Right transitions captured by our metric can reflect authentic learning from video content, consistent with the human verification results reported in Sec. 4.3.2.

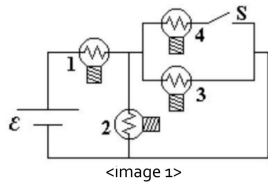


Method Understanding Error

Video Lecture --- "Physics Help: Series and Parallel Circuits Electricity Diagrams Part 4"



Adaptation Question



Question: A circuit is connected as shown. All light bulbs are identical. When the switch in the circuit is closed illuminating bulb #4, which other bulb(s) also become brighter?

Option:

- (A) Bulb #1 only
- (B) Bulb #2 only
- (C) Bulbs #2 and #3 only
- (D) Bulbs #1, #2, and #3
- (E) Bulbs #1 and #2 only
- (F) Bulbs #1 and #3 only
- (G) Bulb #4
- (H) None
- (I) All
- (J) Not sure

Before Video

Model:

When the switch is closed, bulb #4 is added in parallel with bulb #3. This changes the total resistance in the circuit and affects the brightness of the other bulbs.

1. When the switch is closed, bulbs #3 and #4 are in parallel, and this parallel combination is in series with bulbs #1 and #2.
2. Closing the switch reduces the total resistance of the circuit because the parallel combination of bulbs #3 and #4 has a lower resistance than bulb #3 alone. This increases the total current from the battery.
3. Since bulbs #1 and #2 are in series with the parallel combination of bulbs #3 and #4, the increased total current flows through bulbs #1 and #2, making them brighter. Therefore, the correct answer is: **F. Bulbs #1 and #2 only.**

Human:

When the switch is closed, Bulb #4 is added in parallel with Bulb #3. This means the current through the circuit increases overall. Since all the bulbs are identical, the increased current is equally shared among all the bulbs in the circuit. Therefore, all the bulbs including Bulbs #1, #2, and #3 will become brighter because the total power delivered to the circuit has increased. Hence the answer is **D. Bulbs #1, #2, and #3.**

Comment:

- The video introduces a **problem-solving approach: use Ohm's law to calculate current through each bulb to determine brightness.** The adaptation question presents a new scenario with an added switch, but we can use the identical approach to determine the current through each bulb.
- Although seeing the same circuit problem, the model **failed to adopt the video's quantitative approach of calculating currents.** Instead, it **still used qualitative analysis.** This might be insufficient to determine current change for each bulb, leading to wrong answer.
- In contrast, human effectively adopts the video's approach to the new scenario. Human calculate the actual currents (ϵ/R) through each bulb before/after switch closes, which clearly shows only bulb #1 increases.
- In essence, while the same problem-solving approach (calculate currents) works in both scenarios, the model still uses its own qualitative analysis instead of learning from the video, leading to incorrect conclusions.

After Video

Model:

To determine which bulbs become brighter when the switch is closed, we need to analyze the circuit configuration and the effect of closing the switch on the current distribution. When the switch (S) is closed:

1. It creates a parallel path with bulb #3
2. The total resistance in that branch decreases (parallel resistors have less total resistance)
3. This means more current will flow through the main circuit. Following the circuit: Bulb #1 will get brighter because more total current flows through the main circuit. Bulb #2 will get brighter for the same reason as Bulb #1. Bulb #3 will actually get dimmer because some of its current is now diverted through Bulb #4
4. Therefore, when switch S is closed and Bulb #4 lights up, Bulbs #1 and #2 will become brighter, while Bulb #3 becomes dimmer. The correct answer is **F. Bulbs #1 and #2 only.**

Human:

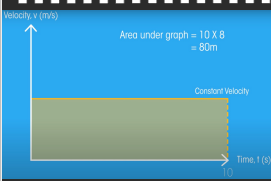
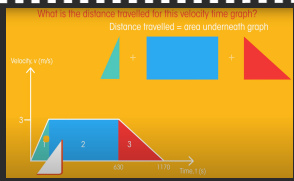
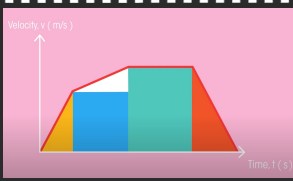
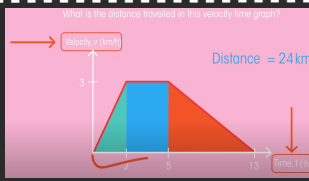
As explained in the video, we can calculate the current through each bulb to determine the brightness.

Before switch closes: Total $R = R + (R \times R)/(R+R) = 3R/2$
 Initial current $I = \epsilon/(3R/2) = 2\epsilon/3R$, $I_1 = 2\epsilon/3R$, $I_2 = I_3 = \epsilon/3R$
 After switch closes: Total $R = R + (R/2 \times R)/(R/2 + R) = 4R/3$
 New current $I = \epsilon/(4R/3) = 3\epsilon/4R$, $I_1 = 3\epsilon/4R$ (increased), $I_2 = I_3 = \epsilon/4R$ (decreased)
 $I_4 = \epsilon/4R$ (turns on)
 Only I_1 increases, hence we choose **A. Bulb #1 only.**

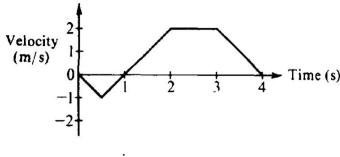
Figure 8: **Method Understanding Error** in the Adaptation track. The model misidentifies the problem-solving method demonstrated in the video, applying a related but incorrect procedure to the adaptation question.

Method Adaptation Error

Video Lecture --- "Area Under Velocity Time Graphs | Forces & Motion"

Adaptation Question



<image 1>

Question: <image 1>The graph above shows the velocity versus time for an object moving in a straight line. At what time after $t = 0$ does the object again pass through its initial position?

Option:

(A) 1 s	(B) Between 1 and 2 s ✓	(C) 2 s
(D) Between 2 and 3 s	(E) 3 s	(F) Between 3 and 4 s
(G) 4 s	(H) Between 4 and 5 s	(I) 5 s
(J) Between 5 and 6 s		

Before Video	After Video
<p>Model:</p> <p>From $t = 0$ to $t = 1$ second, the area under the curve is a triangle with a base of 1 second and a height of -1 m/s. The area (displacement) is -0.5.</p> <p>From $t = 1$ to $t = 2$ seconds, the area under the curve is a triangle with a base of 1 second and a height of 2 m/s. The area (displacement) is: 1</p> <p>From $t = 2$ to $t = 3$ seconds, the velocity is constant at 2 m/s, and the area under the curve is a rectangle with a base of 1 second and a height of 2 m/s. The area (displacement) is: 2.</p> <p>.....</p> <p>The object does not pass through its initial position within the given time frame of 4 seconds. However, it is possible that it passes through the initial position after 4 seconds. Hence, the closest answer would be (H) Between 4 and 5 s. But since the graph does not show this, we cannot definitively determine the exact time from the provided graph. ✗</p>	<p>Model:</p> <p>To determine when the object passes through its initial position again, we need to consider the area under the velocity-time graph, which represents the displacement of the object. The object will pass through its initial position when the net displacement is zero. ✓</p> <p>From 0 to 1 second: The object has a negative velocity, so the area is below the time axis. Area = $0.5 * \text{base} * \text{height} = 0.5 * 1 * (-1) = -0.5$ m.</p> <p>From 1 to 2 seconds: The area is a triangle with base 1 second and height 2 m/s. Area = $0.5 * \text{base} * \text{height} = 0.5 * 1 * 2 = 1$ m.</p> <p>.....</p> <p>As all the area is above zero, we cannot derive a position when the net displacement is zero. The object does not return to its initial position within the 4 seconds shown in the graph. Based on the given options, the closest answer would be: J. Between 5 and 6 seconds ✗</p>
<p>Human:</p> <p>As shown in the graph, the position should be the point where all the area under the curve become zero. Hence the answer is (B) Between 1 and 2 s. ✓</p>	<p>Human:</p> <p>As explained in the video, the area under the curve is the displacement of the object. Hence, we can find the position where the area under the curve is zero. From 0 to 1 seconds, the area is -0.5. From 1 to 2 seconds, the area is 1. Hence the position is between 1 and 2 seconds. The answer should be (B) Between 1 and 2 s. ✓</p>
<p>Comment:</p> <ul style="list-style-type: none"> The video introduces the concept of calculating displacement using the area under a velocity-time graph. The adaptation question asks for the point where the object returns to its initial position, which means the net displacement must equal zero. The examples in the video only show examples with <u>curves above zero</u>, while <u>this question involves a scenario with both positive and negative areas</u>. The model demonstrates an understanding of the method by calculating the area under each interval but fails to correctly apply it during reasoning, mistakenly concluding that the object cannot return to its initial position. In contrast, the human accurately adapts the knowledge from the video to this new scenario, correctly summing up the positive and negative areas to determine that the object returns to its initial position between 1 and 2 seconds, showcasing a concise and correct reasoning process. 	

Figure 9: **Method Adaptation Error** in the Adaptation track. The model correctly identifies the method but introduces computational errors when adapting the reasoning steps to the new scenario.

Method Adaptation Error

Adaptation Track Question
"Engineering – Computer Science"

Question: In Depth First Search (DFS), discovery/finishing time stamps are x/y , where x is discovery time and y is finishing time. If the time stamps of g and h in the graph are changed to $2/7$ and $5/6$, respectively, which of the following traversals are valid depth first forests in <image 1>?

Options:

(A) {a, b, c, d, e} {f, g, h} (B) {a, b, e} {f, g} {c, d} {h}

(C) {a, b, e, h} {c, d, f, g} (D) {a, b, e} {c, d, h} {f, g}

(E) {a, b, e} {c, d, h} {f, g} (F) {a, b, e} {c, d, f, g, h} ✓

(G) {a, b, e, f, g} {c, d, h} (H) {a, b, e} {c, g, f, h} {d}

(I) {a, b, c, e} {d, f, g, h} (J) {a, b, f, e} {c, g, h, d}

Video Lecture
"Pre and Post visited Times in DFS"

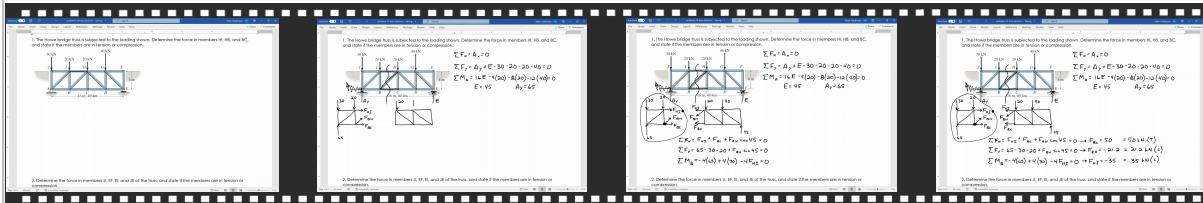
Before Video	After Video
<p>Model:</p> <p>(1) Given the new timestamps and analyzing the connectivity and timing constraints: h must be in a separate tree from g due to their new timestamps and the cycles, c and d form a strongly connected component, a, b, and e form another connected component, f and g are connected.</p> <p>(2) The answer is D: {a, b, e} {c, d, h} {f, g}. This is correct because... ✗</p>	<p>Model:</p> <p>Let me help solve this step by step using what was taught in the video...</p> <p>(1) First, let's recall the key principles: Vertices in the same DFS tree have overlapping intervals. Parent's interval must fully contain child's interval...</p> <p>(2) Analyzing the graph: - Node a is discovered first in the graph - Nodes a, b, e form one tree component - Nodes c, d form another component... ✗</p> <p>(3) Therefore, the answer is B: {a, b, e} {f, g} {c, d} {h}. ✗</p>
<p>Human:</p> <p>(1) Looking at the timestamps, c (1/10) is visited first, followed by g, f, h.</p> <p>(2) Node h has a cycle and does not have any outgoing edges, therefore it must be the end of the first DFS tree. Then d (8/9) was visited and will form a second DFS tree. Then b was discovered followed by e and a.</p> <p>(3) Hence the answer should be H: {a, b, e} {c, g, f, h} {d}. ✗</p>	<p>Human:</p> <p>(1) The video helped me understand the principle: the DFS tree's root node's interval (discovery/finish timestamps) should contain all its descendants' intervals, as the root is discovered first and finished in the end.</p> <p>(2) Applying this principle, I could see that c's interval (1/10) contains the intervals of nodes g, f, h, and d, making them all part of c's DFS tree. ✓</p> <p>(3) Hence the answer is F: {a, b, e} {c, d, f, g, h}. ✓</p>
<p>Comment:</p> <ol style="list-style-type: none"> The video introduces basic DFS principles with timestamps, while the adaptation question tests these concepts in a more complex graph with cycles. Before the video, both Model and Human incorrectly focused on cycles and connectivity patterns. After learning the video, both started correctly by understanding the core principle: a root node's interval must contain its descendants' intervals. However, Model demonstrated a Method Adaptation Error - despite understanding the principle, it failed to correctly adapt it to the complex graph (as shown in blue). In contrast, Human successfully applied the method of interval containment in this new scenario (as highlighted in the box). This reveals the challenges of adapting the method from the video to novel, real-world scenarios. 	

Figure 10: **Method Adaptation Error: Human–Model comparison on a DFS question.** Before watching the video, both the model and the human fail by relying on surface-level connectivity patterns. After watching the video, both verbalize the correct principle (“a root’s interval must contain its descendants’ intervals”), but only the human adapts it correctly to the complex graph; the model reverts to local heuristics and produces an incorrect grouping. This illustrates that articulating a principle does not guarantee correct adaptation under novel structural complexity.

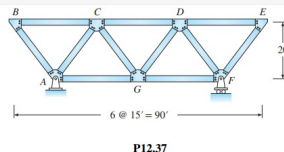


Visual Extraction Error

Video Lecture – “Photoelectric Effect Graphs”



Adaptation Question



Question: If the truss is to be designed for a uniform live load of 0.32 kip/ft that can be placed anywhere on the span in addition to a concentrated live load of 24 kips that can be positioned where it will produce the largest force in bar CG, determine the maximum value of live load force (tension, compression, or both) created in bar CG. <image 1>

- Option:**
- (A) 9.75 kips compression ✓ (B) 5.75 kips compression (C) 7.75 kips compression
 (D) 10.25 kips compression (E) 6.25 kips compression (F) 8.25 kips compression
 (G) 11.50 kips compression (H) 4.50 kips compression (I) 9.00 kips compression (J) 7.00 kips compression

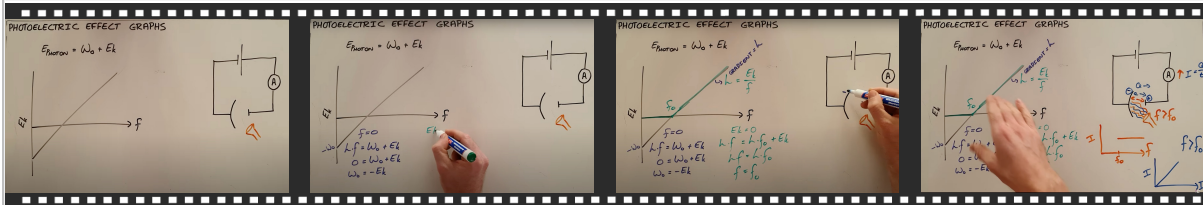
Before Video	After Video
<p>Model: Let me help solve this truss problem step by step. 1) First, let's identify what we're looking for... 2) Key information: - Span = 90 feet Height = 20 feet - Bar CG is a diagonal member ✓ 3) Analysis: - For bar CG, the worst case scenario will be when: The uniform load covers the entire span. The 24-kip concentrated load is placed at joint C. 4) Calculations: Total uniform load = 0.32 kip/ft × 90 ft = 28.8 kips. Load per panel point = 28.8/6 = 4.8 kips. The 24-kip concentrated load at C will create the maximum compression 5) Using method of sections or influence lines, the force in CG can be determined. The answer is A: 9.75 kips compression. ✓</p>	<p>Model: Let me help analyze this step by step based on the video content and the truss problem shown. For bar CG in the truss: 1) We need to consider two types of live loads: - Uniform live load of 0.32 kip/ft(n) - Concentrated live load of 24 kips 2) Key points to consider: - Bar CG is a vertical member as shown in the image ✗ - The truss span is 90 feet and the height is 20 feet 3) For maximum force in CG: Need to consider influence lines for the force... 4) Looking at the given options and considering: The method of sections... Based on the video content and analysis, the maximum compressive force in bar CG would be: C: 7.75 kips compression ✗</p>
<p>Human: We can first calculate the Total Uniform Load: 28.8 kips Then we calculate the Reactions Due to Uniform Load: 1. Since the load is uniformly distributed, the reactions at A and F will be equal. 2. $R_A = R_F = 14.4$ kips To maximize the force in bar CG, place the 24 kip load at point C (mid-span). We can calculate the Reactions Due to Concentrated Load: The concentrated load of 24 kips at point C will be equally distributed to supports A and F... The maximum force in bar CG is G. 11.50 kips in compression. ✗</p>	<p>Human: For maximum force in CG: • Concentrated load should be placed at point C • Uniform load should be placed from A to F (full span) • This creates maximum compression in CG ✓ Calculations: Total uniform load = 0.32 kip/ft × 90 ft = 28.8 kips Total load = 28.8 kips + 24 kips = 52.8 kips Due to truss geometry and load position, bar CG will experience compression Using method of sections and equilibrium equations, max compression in CG = 9.75 kips. The answer should be A: 9.75 kips compression. ✓</p>
<p>Comment: The video explains an example problem about truss analysis and load distribution. The Adaptation track question presented is a similar question to the video example, focusing on finding the maximum force in bar CG under specified loading conditions.</p> <p>For the model: • Before Video: model correctly perceived CG as a diagonal member, it showed proper initial perception without video. • After Video: model adopted correct method from the video, BUT misidentified CG as a vertical member. This perceptual error led to incorrect calculations.</p> <p>In comparison, human maintained correct perception of CG as diagonal. And before the video, human did not have a correct reasoning. After learning from the video, human effectively integrated the video knowledge and arrived at accurate force calculations.</p> <p>This case reveals an interesting challenge: a visual detail that was correctly perceived in a static image became misinterpreted when presented in the context of video content. The complexity of processing the long video context may pose additional challenge to the accurate perception of information.</p>	

Figure 11: **Visual Extraction Error** in the Adaptation track. The model misreads fine-grained visual information, causing downstream reasoning errors despite correct methodology.

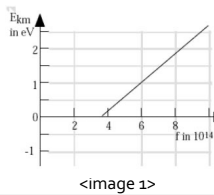


Context Confusion Error

Video Lecture – "Photoelectric Effect Graphs"



Adaptation Question



Question: A student performs the photoelectric effect experiment and obtains the data depicted in the accompanying graph of E_{km} (maximum kinetic energy) of photoelectrons v. the frequency of the photons. What is the approximate work function of this material?

Option:

- (A) 1.5 eV ✓
- (D) 3.5 eV
- (G) 3.0 eV

- (B) 2.0 eV
- (E) 1.8 eV
- (H) 3.8 eV

- (C) 2.7 eV
- (F) 2.2 eV
- (I) 2.5 eV

- (J) 4.0 eV

Before Video

Model:

The equation governing the photoelectric effect is:

$$E_k = hf - \phi \quad \checkmark$$

where:

- E_k is the maximum kinetic energy of the photoelectrons,
- h is Planck's constant, $h = 4.14 \times 10^{-15} \text{eV}$,
- f is the frequency of the photons,
- ϕ is the work function of the material.

The work function ϕ corresponds to the threshold frequency f_0 where the kinetic energy $E_k = 0$.

From the graph: The threshold frequency $f_0 = 4 \times 10^{14} \text{Hz}$. ✗

The work function is calculated as: $\phi = h \cdot f_0$

Substitute $f_0 \approx 4 \times 10^{14} \text{Hz}$ and $h = 4.14 \times 10^{-15} \text{eV}$: $\phi \approx 1.8 \text{eV}$ ✗

Human:

We can derive from the photoelectric effect equation:

$$E_k = hf - \phi$$

To calculate the work function ϕ , the E_k must be $E_k = 0$.

Hence, we can calculate ϕ by substituting $f_0 = 4 \times 10^{14} \text{Hz}$ and $h = 4.14 \times 10^{-15} \text{eV}$: $\phi \approx 1.8 \text{eV}$ ✗

Comment:

The video introduces a similar example related to Photoelectric Effect Graphs, demonstrating that the work function (ϕ) can be simply calculated by finding the y-intercept of the line. This question can be solved without using formulas, as we only need to recognize that **the y-intercept provides the answer directly**.

Before the video:

- Both humans and model approached the problem by using formulas and substituting values
- This indirect method led to incorrect answers and was not the most efficient approach

After the video:

- The model correctly understood that it should look at the y-intercept of the graph.
- However, **model recalls details from the video example, and assume the input from the video example to be the adaptation question condition.**
- Humans, on the other hand, demonstrated concise and accurate reasoning. They correctly identified the y-intercept to find the answer of 1.5.

This is a context confusion error that model relies overly on video details and affect its reasoning.

After Video

Model:

To determine the work function ϕ of the material from the graph of maximum kinetic energy E_k of photoelectrons versus the frequency f of the photons, we can use the photoelectric equation:

$$E_k = hf - \phi \quad \checkmark$$

This equation represents a straight line $y = mx + c$, where:

- y is E_k
- x is f
- m is h
- c is $-\phi$

From the video, the y-intercept is -2.0 eV. Therefore, the work function $\phi = -c = 2.0 \text{eV}$ ✗

Human:

According to the video lecture, we can directly solve this work function ϕ by directly find the y-intercept of the line. ✓

The y-intercept of the line in this graph is approximately -1.5 eV. Hence, the work function $\phi = 1.5 \text{eV}$. ✓

Figure 12: **Context Confusion Error** in the Adaptation track. The model reuses instance-specific values from the video's worked example instead of the adaptation question's parameters.

Error in Perception Track

Video Lecture – "Identifying Muscle | Review and Practice"

Perception Question

Question:
At the beginning of the video, what are the muscles in the lower left corner, upper left corner, and lower right corner, respectively?

Option:

A. Cardiac muscle, Smooth muscle, Skeletal muscle;	B. Skeletal muscle, Cardiac muscle, Smooth muscle
C. Skeletal muscle, Smooth muscle, Cardiac muscle;	D. Smooth muscle, Cardiac muscle, Skeletal muscle
E. Smooth muscle, Skeletal muscle, Cardiac muscle;	F. Smooth muscle, Cardiac muscle, Cardiac muscle
G. Skeletal muscle, Skeletal muscle, Smooth muscle;	H. Cardiac muscle, Smooth muscle, Smooth muscle
I. Skeletal muscle, Smooth muscle, Smooth muscle;	J. Cardiac muscle, Skeletal muscle, Smooth muscle ✓

Model response

Model:
Looking at the initial frame of the video, the arrangement of the muscle types is: **B. Skeletal muscle (upper left), Cardiac muscle (lower left), Smooth muscle (lower right).** ✗

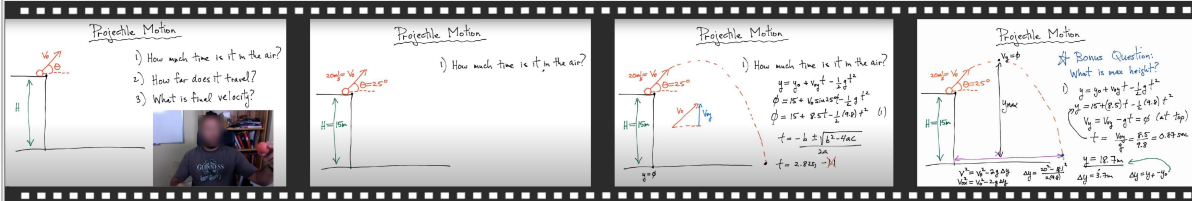
Comment:

- This example is an error in the perception track.
- The user specifically arranged the **question** to test the model's ability to recognize the **positions of the muscles** when the **sequence is deliberately randomized (lower left, upper left, lower right)**.
- However, the model interpreted the **positions** in a **habitual reading order (left to right, top to bottom)** and provided an **incorrect sequence** of muscle identifications. This highlights a gap in the model's capacity to accurately process and respond to spatial arrangements in visual content.
- Future development should focus on enhancing the model's spatial perception capabilities to prevent such misjudgements.

Figure 13: **Perception Track Error.** The model fails to accurately extract information directly presented in the video content.

✘ Error in Comprehension Track

Video Lecture – "Projectile launched off a cliff at an angle"



Comprehension Question

Question:

In the video, Example Question (1) is solved with an angle $\theta=25$ degrees. If the angle θ is adjusted to 30 degrees while all other conditions remain unchanged, what will be the updated result for Example Question (1) as explained in the video?

Options:

- (A) 4.00 seconds (B) 2.82 seconds (C) 3.50 seconds (D) 2.50 seconds (E) 3.04 seconds ✔
 (F) 2.00 seconds (G) 3.15 seconds (H) 1.85 seconds (I) 2.25 seconds (J) 3.85 seconds

Model response

Model:

Let me help you solve this step by step for $\theta = 30^\circ$.

- 1) To find the time in the air, we can use the same approach shown in the video, but with $\theta = 30^\circ$: Initial velocity (V_0) = 20 m/s, $\theta = 30^\circ$
- 2) First, we need V_{0y} (initial vertical velocity): $V_{0y} = V_0 \sin(30^\circ) = 20 \times 0.5 = 10$ m/s
- 3) For time in the air, we can use:
 - The object returns to its initial height ✘
 - Vertical velocity at that point equals $-V_{0y}$
- 4) Using $V_y = V_{0y} - gt$ ✘
 - $10 = V_{0y} - gt$
 - $t = 2V_{0y}/g$
 - $t = 2(10)/9.8$
 - $t = 20/9.8$
 - $t = 2.04$ seconds
- 5) Looking at the options provided, the closest answer is: H. 1.85 seconds. Therefore, H. 1.85 seconds is the answer. ✘

Comment:

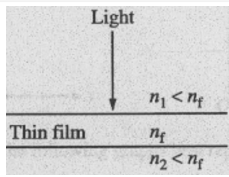
- The case is a comprehension error. This situation clearly demonstrates how the model fails to understand **problem-solving strategies** shown in the video.
- For comprehension questions based on video examples, the test typically uses **the same problem but changes some input values**. To demonstrate true comprehension, the model simply needs to follow the **exact same reasoning process** shown in the video, just with different numbers.
- In this case, the video explains how to calculate a projectile's air time when launched from a cliff at an angle, with the example shown in the second and third frames. The comprehension question uses the identical scenario but only changes one input - the launch angle from 25 to 30 degrees.
- However, the model solved the comprehension question with the **incorrect method and formula**: $t = (2 * V_{0y}) / g$. This formula only works for projectiles that return to their starting height, like when launching from ground level. Since this problem involves a 15-meter elevated platform with the projectile landing below its launch point, this shortcut formula does not apply.
- **The correct solution** requires using the quadratic equation that accounts for different launch and landing heights: $y(t) = y_0 + V_{0y} * t - (1/2) * g * t^2$. This is the same method demonstrated in the video - a process that humans can easily follow but seems challenging for models.
- This example highlights how models struggle to fully comprehend problem-solving strategies shown in educational videos. While a human learner would simply follow the same reasoning steps with new numbers, the model failed to replicate this basic aspect of learning.

Figure 14: **Comprehension Track Error.** The model misinterprets the underlying concept demonstrated in the video, leading to incorrect answers in the multiple-answer format.

Wrong-to-Right Example

Video Lecture – “Thin Film Interference Summary”

Adaptation Question



Question: A thin film with index of refraction n_f separates two materials, each of which has an index of refraction less than n_f . A monochromatic beam of light is incident normally on the film, as shown above. If the light has wavelength λ within the film, maximum constructive interference between the incident beam and the reflected beam occurs for which of the following film thicknesses?

- Option:**
- (A) 2λ (B) λ (C) $\lambda/2$ (D) $\lambda/4$ (E) 3λ
 (F) $5\lambda/2$ (G) $3\lambda/2$ (H) $\lambda/8$ (I) $\lambda/16$ (J) 4λ

Before Video

Model:

Let me help solve this thin film interference problem.

For constructive interference in a thin film:

- 1) First, we need to consider phase shifts at the boundaries
- 2) Then, consider the path difference for constructive interference

Key points:

- Light is incident normally (perpendicular to surface) $n_1 < n_f$ and $n_2 < n_f$ (both surrounding media have lower refractive index than film)
- We want maximum constructive interference
- Phase shifts: At top boundary: 180° (π) phase shift because $n_1 < n_f$. At bottom boundary: 180° (π) phase shift because $n_2 < n_f$. Total phase shift from reflections = 360° (2π) = **no effective phase shift.**

For constructive interference:

- Total path = $2t$ (down and up through film)
- Condition: $2t = m\lambda$ (where m is an integer)

Therefore, $t = m\lambda/2$. The smallest non-zero thickness that satisfies this is: $t = \lambda/2$. Looking at the options, **C ($\lambda/2$)** is the correct answer.

After Video

Model:

Let me help you solve this step by step using the concepts shown in the video:

1) From the video, we learned that for thin film interference:

- When there's a phase shift at both or neither boundary: $2t = (m + 1/2)\lambda$ for constructive interference
- When there's a phase shift at only one boundary: $2t = m\lambda$ for constructive interference

2) In this problem:

- The film has higher index of refraction (n_f) than both surrounding materials. Light is incident normally (vertically)

3) Key points:

- When light reflects from a medium with higher n , there's a phase shift of $\lambda/2$
- When light reflects from a medium with lower n , there's no phase shift

4) Analysis:

Since n_f is higher than both surrounding materials. This means one phase shift total.

5) Therefore:

- We use the formula $2t = (m + 1/2)\lambda$ for constructive interference
- For $m = 0$, this gives us $t = \lambda/4$. Therefore, the answer is **D: $\lambda/4$.**

Comment:

This example illustrates how the model **successfully learned from a video lecture on thin film interference** to correct its initial misunderstanding of phase shifts at boundaries.

The video demonstrates when phase shifts occur at interfaces between materials with different refractive indices. The adaptation question tests this knowledge by asking about the film thickness required for maximum constructive interference.

Before video, the model:

- Incorrectly assumed 180° phase shifts at both boundaries
- Failed to understand that phase shift depends on relative refractive indices
- Reached wrong conclusion that $t = \lambda/2$ for constructive interference

After watching the video, the model:

- Correctly identified phase shift only at top boundary (low to high index)
- Applied proper interference conditions to determine $t = \lambda/4$ for constructive interference

In conclusion, this case demonstrates successful knowledge acquisition from video. The model **corrects its understanding of phase shifts at interfaces** and thin film interference through learning from the video, and applies the demonstrated principles to solve the adaptation problem correctly.

Figure 15: **Wrong→Right Example.** The model answers incorrectly without the video but correctly after watching it, with the chain-of-thought explicitly referencing the video-demonstrated method.