PedagogyBench: A Cognitive-Driven Benchmark for Multimodal Instructional Video Understanding

Xiaokang Jin[1,2*] **Jia Zhu**[1*†] **Jingjiang Liu**[1,2] **Yabing Shi**[1]
Jueqi Guan[1] **Hao Chen**[3†] **Pasquale De Meo**[4]
[1]Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University
[2]School of Computer Science and Technology, Zhejiang Normal University
[3]Tencent Financial Technology [4]Department of Computer Science, University of Messina
jiazhu@zjnu.edu.cn herrickchen@tencent.com

Abstract

Existing video understanding benchmarks mainly emphasize general visual recognition and reasoning, but do not adequately capture the pedagogical logic embedded in instructional videos. To address this gap, we present PedagogyBench, a multimodal benchmark for instructional video understanding grounded in pedagogical cognition. We introduce a pedagogy-driven segmentation strategy and a dual-stream semantic injection pipeline that combines machine pre-annotation with expert refinement, enabling the construction of a dataset organized around a cognitive pyramid with four levels and 20 fine-grained tasks. We further propose the Cognitive Fidelity Score (CFS) to measure the balance of model performance across pedagogical cognitive dimensions. Experiments on 12 multimodal large language models reveal a clear generative gap, where models perform relatively well on discriminative tasks but degrade on higher-order pedagogical diagnosis, often relying on parametric memory rather than grounded visual perception. Project resources are available at https://github.com/Shallcom/PedagogyBench.

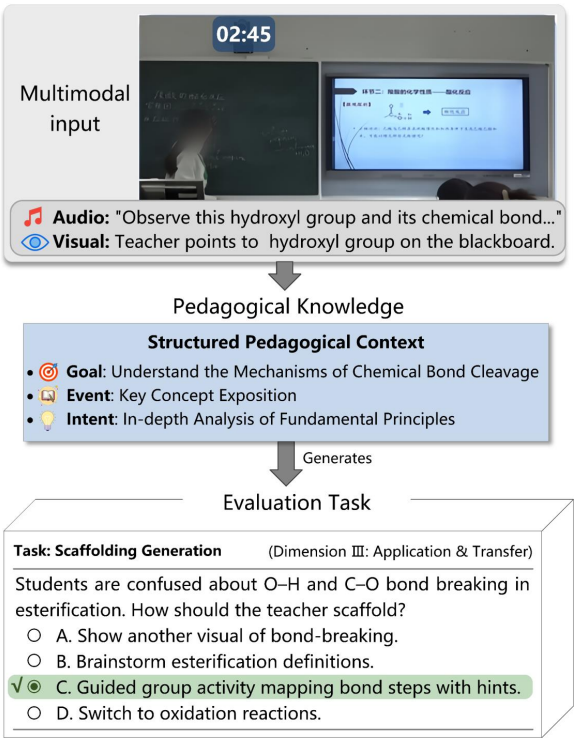


Figure 1: PedagogyBench transforms multimodal inputs into structured pedagogical knowledge, which then supports the generation of evaluation tasks. This example challenges MLLMs to infer the most appropriate scaffolding strategy for a chemistry concept, requiring reasoning beyond surface-level perception.

1 Introduction

In recent years, Multimodal Large Language Models (MLLMs) have advanced rapidly in cross-modal perception and reasoning (Chen et al., 2024; Dai et al., 2023; Qin et al., 2025). Representative models such as GPT-4o (Team et al., 2024), Gemini (Team, 2025), and Video-LLaMA (Zhang and Li, 2023) have shown strong performance on image understanding, video comprehension, and visual question answering (Li et al., 2025; Pang et al., 2024). These advances have broadened the applicability of MLLMs to more complex real-world scenarios (Xi et al., 2025).

Instructional videos have become an important medium for knowledge delivery and skill development in digital education. Unlike general videos, they are organized around pedagogical goals and structured teaching phases (Mayer, 2020). A classroom video is not merely a sequence of visual and auditory signals; it also reflects the teacher's instructional design, including topic introduction, concept explanation, example demonstration, interaction, and lesson summary (Merrill, 2002). Understanding such videos requires not only multimodal perception, but also the ability to model pedagogical organization, teaching intent, and the evolving

*These authors contributed equally to this work.
†Corresponding authors: Jia Zhu and Hao Chen.

Findings of the Association for Computational Linguistics: ACL 2026, pages 12621–12647
July 2-7, 2026

Benchmark	Domain	Video Source	Segmentation Logic	#Units	#Q&A	Len. (s)	Format
Video-Bench (Ning et al., 2026)	General	Combined	Task-oriented	5,917	17,036	56.0	MC
Video-MME (Fu et al., 2024)	General	Open-domain	Multi-level temporal	900	2,700	1,017.9	MC
MVBench (Li et al., 2024)	General	Combined	Task-oriented	3,641	4,000	16.0	MC
Video-MMMU (Hu et al., 2025)	Education	Lecture videos	Stage-aligned	300	900	506.2	MAMC
InstructionBench (Wei et al., 2025)	Education	Instructional datasets	Step-aligned	713	5,000	282.9	MC
PedagogyBench	Education	Teaching competitions	Pedagogy-driven	1,852	11,112	184.2	MC/OE

Table 1: Comparison of PedagogyBench with representative video understanding benchmarks. #Units and #Q&A denote the numbers of evaluation units and QA pairs; Len. denotes average input length in seconds; MC = multiple choice; OE = open-ended; MAMC = multiple-answer multiple-choice.

logic of knowledge delivery.

However, existing multimodal benchmarks, such as MME (Fu et al., 2025), MVBench (Li et al., 2024), and Video-MME (Fu et al., 2024), primarily focus on visual recognition, common-sense reasoning, and factual question answering in general scenarios. These benchmarks usually treat videos as temporal sequences of events, while underemphasizing the bidirectional teaching–learning process in classroom settings. At the same time, the lack of a benchmark tailored to educational scenarios makes it difficult to systematically evaluate whether MLLMs can capture the deeper pedagogical logic of instructional videos.

To address this gap, we present PedagogyBench, a benchmark for multimodal instructional video understanding in educational settings. Figure 1 shows a representative example from PedagogyBench. Our main contributions are as follows:

1. **A Benchmark Grounded in Instructional Design**: We construct PedagogyBench using a pedagogy-driven segmentation strategy that aligns video granularity with teaching logic. This design evaluates whether a model can follow pedagogical flow rather than only recognize isolated events.

2. **A Dual-Stream Annotation Pipeline with Expert Refinement**: We develop a dual-stream semantic injection pipeline that combines machine pre-annotation and expert refinement. This process captures the interactions among spoken explanations, blackboard writing, and visual demonstrations in classroom videos.

3. **A Cognitive Evaluation Framework for Instructional Video Understanding**: We establish a cognitive framework with four levels spanning 20 fine-grained tasks and introduce the CFS to measure balanced performance across dimensions. Experiments on 12 MLLMs reveal a clear generative gap from option-guided discrimination to higher-order pedagogical reasoning.

2 Related Work

2.1 MLLMs for Dynamic Visual Reasoning

MLLMs have advanced rapidly in recent years. Early models like LLaVA (Liu et al., 2023) connected visual encoders with LLMs, validating the effectiveness of visual instruction tuning. To enhance fine-grained perception, approaches such as LLaVA-NeXT(Zhang et al., 2024), Qwen2.5-VL(Bai et al., 2025b), and InternVL 2.5 (Chen et al., 2025b) have advanced general visual understanding by integrating dynamic high-resolution mechanisms and rigorous data filtering. Addressing the challenges of video temporality and long sequences, Video-LLaMA(Zhang and Li, 2023) utilizes Q-Formers to process temporal data, while mPLUG-Owl3(Ye et al., 2024) optimizes attention computation via a hyper attention mechanism. Furthermore, Qwen3-VL(Bai et al., 2025a) and InternVL 3.0 (Zhu et al., 2025) have overcome context length limitations, supporting coherent reasoning for ultra-long video and multi-image inputs. GLM-4.1V-Thinking (Team et al., 2025b) has significantly enhanced multimodal chain-of-thought capabilities. While proprietary models such as GPT-4o (Team et al., 2024) and Gemini 2.5 Pro (Team, 2025) show stronger long-context performance, open-source MLLMs remain prone to hal-

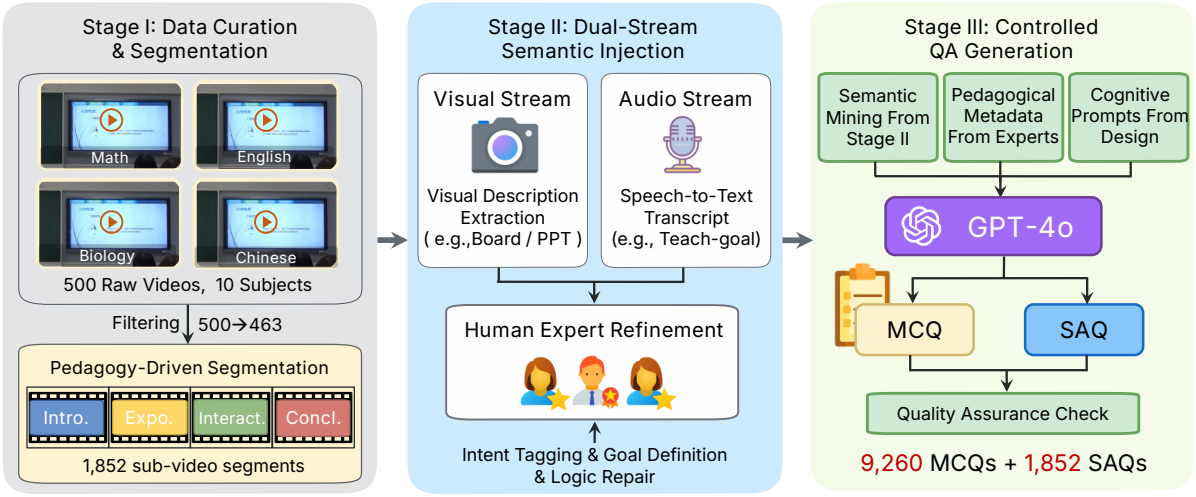


Figure 2: The data construction pipeline of PedagogyBench. It comprises three stages: (I) Data Curation and Pedagogy-Driven Segmentation; (II) Dual-Stream Semantic Injection with Rigorous Human Expert Refinement; and (III) Context-Aware QA Generation synthesizing MCQs and SAQs.

lucinations (Li et al., 2023) and temporal inconsistencies (Yang et al., 2025) in complex scenarios.

2.2 Benchmarks for Video Understanding

Video understanding benchmarks have shifted from basic perception to complex reasoning. Early benchmarks like MSVD-QA (Xu et al., 2017) and MSRVTT-QA (Xu et al., 2016) targeted short video descriptions, while ActivityNet-QA (Yu et al., 2019) and TGIF-QA (Jang et al., 2017) introduced foundational temporal localization. With the rise of MLLMs, MME (Fu et al., 2025) established a unified standard for perception and cognition. MVBench (Li et al., 2024) and Video-MME (Fu et al., 2024) emphasized fine-grained temporal perception and cross-duration logical reasoning. To explore deeper cognitive abilities, EgoSchema (Mangalam et al., 2023) and NExT-QA (Xiao et al., 2021) focus on causal inference and long-term dependencies. Meanwhile, VQAGuider (Chen et al., 2025a) and MoReVQA (Min et al., 2024) investigate modular, step-by-step reasoning. However, these benchmarks largely evaluate what happens or why it occurs, neglecting the pedagogical logic and cognitive scaffolding essential to instruction. This lack of pedagogically grounded evaluation limits how accurately current metrics can assess model capability in educational settings.

2.3 Instructional Video Analysis

Instructional video analysis focuses on equipping models with procedural knowledge. HowTo100M (Miech et al., 2019) laid the foundation for weak supervision using large-scale narrated videos, while COIN (Tang et al., 2019) and CrossTask (Zhukov et al., 2019) improved fine-grained step recognition through hierarchical annotations. Subsequently, StepFormer (Dvornik et al., 2023) improved step localization accuracy via self-supervised temporal modeling, and InstructionBench (Wei et al., 2025) utilized multi-granularity QA to assess step sequencing and reasoning. Additionally, EgoProceL (Bansal et al., 2022) and Ego4D (Grauman et al., 2022) addressed key state changes and egocentric perspectives. However, existing works mainly emphasize procedural execution rather than pedagogical metacognition and instructional rationale. PedagogyBench is designed to address this gap.

3 PedagogyBench

PedagogyBench is a benchmark for pedagogical cognition in instructional videos, covering 10 K–12 subjects, 1,852 video segments, and 11,112 QA pairs. Unlike existing benchmarks that mainly emphasize general event understanding or procedural steps, it focuses on the teaching–learning logic of real classroom instruction. Table 1 compares PedagogyBench with representative general and instructional benchmarks.

3.1 Dataset Construction Pipeline

The construction of PedagogyBench follows a three-stage pipeline, and the overall workflow is illustrated in Figure 2.

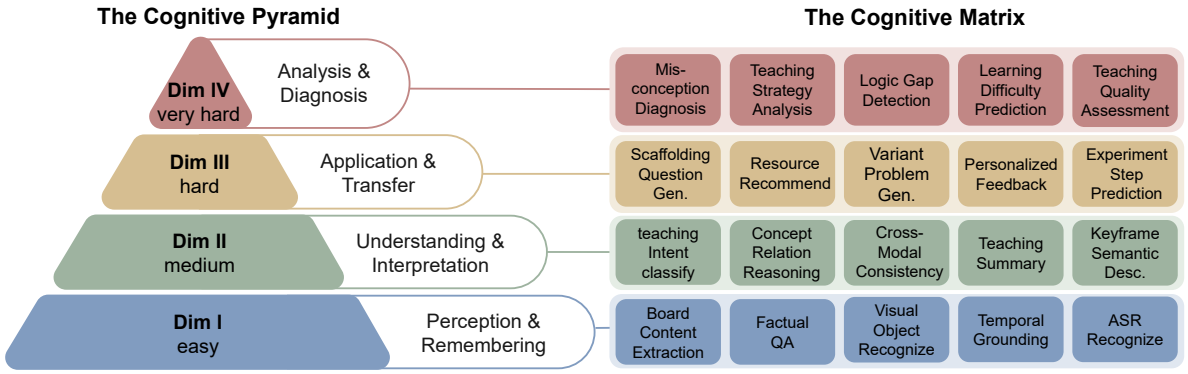


Figure 3: The left side organizes the evaluation into four cognitive levels, forming a pyramid from easy to very hard. The right side expands each level into concrete task types, yielding a cognitive matrix that covers perception, understanding, application, and diagnostic reasoning in instructional videos.

3.1.1 Data Curation and Pedagogy-Driven Segmentation

Stage I: The raw videos in PedagogyBench come from teaching skills competitions, rather than online videos or public datasets. Unlike typical classroom recordings, these videos exhibit significantly higher quality in both instructional design and recording conditions. They typically contain a complete "Introduction–Exposition–Interaction–Conclusion" structure, together with clear blackboard writing and clean audio, making them suitable for benchmark construction. All source videos are drawn from Chinese K–12 teaching competitions, with classroom speech and on-screen text primarily in Chinese. The benchmark uses English QA pairs by design to evaluate cross-lingual pedagogical reasoning rather than surface-level language matching.

We collected 500 raw long videos across 10 subjects. To ensure pedagogical integrity, we retained only videos lasting 10–15 minutes that included all four teaching segments, resulting in a curated set of 463 high-quality videos.

We then performed pedagogy-driven segmentation based on instructional event boundaries, moving beyond traditional shot segmentation toward cognitively meaningful semantic units. Specifically, we jointly exploited multimodal cues such as speech rate variations, slide transitions, blackboard layout changes, and interaction signals to automatically detect instructional event boundaries. This process divided each video into four distinct segments: Introduction, Exposition, Interaction, and Conclusion, yielding 1,852 coherent sub-video segments (463×4) and providing the structured temporal basis for subsequent annotation.

3.1.2 Dual-Stream Semantic Injection

Stage II converts visually observable yet hard-to-verbalize classroom details into structured semantic representations. To this end, we develop a dual-stream semantic injection mechanism that combines automated extraction with expert refinement.

For the audio stream, we use the speech recognition model FunASR (Gao et al., 2023) to transcribe all sub-videos, capturing teaching explanations, questions, and feedback. Beyond raw transcripts, this process also produces phase-level summaries and pedagogical objectives, which provide useful temporal cues about instructional pacing. **For the visual stream**, we employ a phase-aware prompting system based on a vision LLM (Zhu et al., 2025). These prompts guide the model to adopt observation roles aligned with different pedagogical phases. Detailed prompts and annotation examples are provided in Appendix B.

To improve data quality, especially for STEM subjects (Physics, Chemistry, Biology, Mathematics, and Technology), we incorporate scientific notation constraints into the prompts, requiring formulas to be transcribed in standardized form. To further reduce hallucination, we adopt a "**Facts-First, Narrative-Second**" strategy, which first enumerates chronological visual facts and then synthesizes them into a narrative.

To improve annotation quality, we incorporate a human-in-the-loop verification process. As shown in Figure 2, the Human Expert Refinement module in Stage II involves three doctoral candidates who revise the machine-generated drafts at the instance level, including intent tagging, logic repair, and goal definition. This process improves the faithfulness of the annotations to subject knowledge and

Subject	# Seg.	MCQs	SAQs	Tokens (k)		Avg. Duration (min)				Total Duration	
				Audio	Visual	Intro	Expo	Inter.	Concl.	(min)	Avg. (min)
Biology	180	900	180	154.0	122.1	2.59	4.06	3.32	2.11	543.0	3.02
Chemistry	188	940	188	155.9	151.6	2.51	4.05	3.24	2.41	573.5	3.05
Chinese	196	980	196	172.8	146.0	2.60	3.57	3.79	2.82	626.1	3.19
English	196	980	196	96.5	141.8	2.62	3.95	3.49	2.50	614.9	3.14
Geography	144	720	144	116.4	116.2	2.36	4.00	3.58	2.23	437.8	3.04
History	180	900	180	168.8	141.2	2.39	3.80	3.44	2.45	543.3	3.02
Mathematics	200	1000	200	163.6	142.2	2.52	4.03	3.53	2.19	612.3	3.06
Physics	192	960	192	163.4	133.8	2.50	4.53	3.64	1.92	577.1	3.00
Politics	180	900	180	160.3	138.4	2.43	4.32	3.32	2.20	561.2	3.12
Technology	196	980	196	157.7	152.7	2.70	3.76	3.37	2.48	603.3	3.08
Total / Avg.	**1,852**	**9,260**	**1,852**	**1,509.4**	**1,386.0**	**2.52**	**4.01**	**3.47**	**2.33**	**5,692.5**	**3.07**

Table 2: Detailed statistics of PedagogyBench across 10 subjects. The table reports the scale of video segments and QA pairs, multimodal token counts, and the average duration for each of the four pedagogical phases.

pedagogical intent. Details of the refinement procedure, together with a before-and-after example, are provided in Appendix B.3.

3.1.3 Controlled QA Generation Engine

Stage III converts the structured annotations into evaluation tasks. We developed a controlled QA generation engine powered by GPT-4o (Team et al., 2024) to ensure question quality while maintaining comparability and fairness across models. Details of the generation engine are provided in Appendix C.

For each video segment, the engine generates five multiple-choice questions (MCQs) and one short-answer question (SAQ) based on the multimodal annotations. We explicitly specify the mapping between questions and the cognitive pyramid, enforcing a difficulty-increasing constraint. This ensures a balanced distribution of complexity, progressively advancing from basic factual perception to higher-order pedagogical diagnosis.

To mitigate shortcut reasoning, we remove text snippets that may directly reveal answers and apply post-validation checks for formatting compliance and option completeness. Importantly, all QA generation is grounded in the expert-refined annotations from Stage II, including teaching goals, visual facts, and pedagogical summaries, which reduces hallucinations at the source.

We further perform sampled human validation on the generated QA pairs. Three PhD students reviewed 300 segments (1,800 QA items), of which only 23 cases required minor revisions, mainly in option formatting, while no answer keys or factual content needed correction. For the most challenging Dim IV tasks, two K–12 in-service teachers additionally reviewed 60 diagnostic items, confirming that 57 of them (95%) reflected realistic classroom challenges. Only QA pairs retained after this multi-round verification process are included in the final benchmark.

3.2 The Cognitive Evaluation Framework

To systematically assess the pedagogical cognitive capabilities of MLLMs, we draw upon Bloom's Taxonomy (Anderson, 2001) and the Pedagogical Content Knowledge (Shulman, 1986) framework to organize the evaluation tasks into a hierarchical pyramid of four progressive dimensions, with each dimension encompassing five specific subtasks, as detailed in Figure 3 and Appendix C.1.

Dim I: Perception and Remembering. Grounded in Mayer's learning theory (Mayer, 2020), this dimension evaluates the ability to perceive audio-visual signals by grounding visual and auditory details. We design tasks such as blackboard OCR and speech transcription check to verify whether the inputs are free from hallucination, thereby establishing the foundation for subsequent reasoning.

Dim II: Understanding and Interpretation. This dimension leverages dual-coding theory (Paivio, 1991) to assess multimodal meaning construction. We design tasks such as multimodal alignment and pedagogical phase recognition, requiring the model to determine whether the teacher's gestures align with spoken explanations and thereby form a coherent understanding of instructional messages.

Dim III: Application and Transfer. This level simulates the process of learning transfer (Haskell, 2000), assessing the model's ability to apply internalized logic to novel contexts. We require the model to synthesize logically consistent examples or predict subsequent experimental steps based on its understanding of the current pedagogical logic.

Model	Subjects										Avg	CFS
	Math	Phys	Chem	Bio	Geog	Tech	Hist	Eng	Chi	Pol		
Gemini-2.5	**88.38**	**88.91**	85.11	87.29	88.00	**77.94**	**87.36**	88.07	**87.75**	88.12	86.69	**83.07** ↑
GPT-5.1	87.95	87.83	**87.43**	**87.99**	**89.39**	75.51	86.36	**88.47**	87.25	**88.82**	**86.70**	82.78 ↓
Qwen3-VL:8B	84.68	85.29	80.14	83.76	88.45	74.09	86.46	84.38	85.66	85.16	83.81	80.03 –
Qwen2.5-VL:7B	83.44	82.39	80.79	81.53	88.11	72.45	85.07	82.91	82.20	81.53	82.04	78.48 –
InternVL3:8B	83.38	84.88	79.40	80.08	86.99	73.50	82.58	84.94	83.93	79.85	81.95	76.80 –
Gemma3:12B	81.13	81.58	76.53	79.03	80.99	71.56	76.66	81.46	81.89	71.25	78.21	72.11 –
MiniCPM-V:8B	75.56	76.11	71.21	69.93	76.82	64.66	76.39	75.77	78.83	73.19	73.85	70.33 ↑
InternVL2.5:8B	80.81	77.41	74.93	76.32	78.65	71.55	74.52	77.74	79.78	76.11	76.78	69.85 ↓
GLM-4.1V-Thinking:9B	75.19	74.87	68.62	67.92	73.70	64.48	73.68	71.69	71.24	70.83	71.22	67.77 ↑
InternVL2:8B	78.94	72.46	69.22	73.47	75.69	69.77	73.47	73.02	74.30	72.57	73.29	65.40 ↑
LLaVA-NeXT:7B	80.00	73.57	71.61	75.21	79.34	72.58	74.72	81.31	80.87	72.01	76.12	64.87 ↓
mPLUG-Owl3:7B	75.75	71.03	67.82	72.09	79.17	70.28	72.50	79.86	81.06	70.07	73.96	60.67 ↓

Table 3: Main Results on PedagogyBench. Reported are accuracy scores (%) across 10 subjects, Average Accuracy (Avg), and Cognitive Fidelity Score (CFS). Models are ranked by CFS. Bold denotes best performance. Arrows (↑ / ↓ / –) indicate rank improvement, decline, or stability in CFS relative to Avg.

Dim IV: Analysis and Diagnosis. This dimension evaluates diagnostic competence (Klug et al., 2013) and requires models to emulate the metacognitive processes of an expert teacher. We design tasks such as error diagnosis and teaching strategy analysis, requiring the model to identify the conceptual roots of student errors or assess the effectiveness of specific instructional strategies.

3.3 Data Analysis and Statistics

PedagogyBench comprises 1,852 video segments and 11,112 QA pairs across 10 subjects. A detailed statistical breakdown is provided in Table 2; additional analyses are included in Appendix A.

In total, the benchmark contains 5,692.5 minutes of instructional video. It includes 1,509.4k audio tokens and 1,386.0k visual tokens, providing dense multimodal evidence for pedagogical understanding. The average segment duration is 3.07 minutes, which is consistent with the pacing of real classroom instruction.

Temporally, instructional content is concentrated in the Exposition (4.01 min) and Interaction (3.47 min) phases, while Introduction (2.52 min) and Conclusion (2.33 min) are relatively concise. This non-uniform distribution reflects authentic teaching practice and requires models to capture both core knowledge delivery and its surrounding pedagogical context.

PedagogyBench also maintains broad disciplinary coverage across both STEM and humanities subjects, including Mathematics, Physics, Chemistry, Biology, Geography, History, Chinese, English, Politics, and Technology. This relatively balanced subject distribution helps reduce domain bias and supports the evaluation of cross-disciplinary generalization.

4 Experiments and Analysis

We conduct a systematic evaluation of 12 MLLMs on PedagogyBench. Our analysis addresses four questions: the general proficiency of current models across pedagogical subjects, the extent to which PedagogyBench depends on genuine visual grounding rather than language priors, the balance of model competence across cognitive dimensions, and the failure modes revealed by this benchmark.

4.1 Experimental Setup

Evaluated Models: To ensure a comprehensive evaluation, we assess 12 representative MLLMs, including two proprietary frontier models and ten open-source models mostly in the 7B–12B range. Specifically, the evaluated models include: Gemini-2.5(Team, 2025), GPT-5.1 (OpenAI, 2025), Qwen2.5-VL:7B (Bai et al., 2025b), Qwen3-VL:8B (Bai et al., 2025a), InternVL2:8B (Wang et al., 2025), InternVL2.5:8B (Chen et al., 2025b), InternVL3:8B (Zhu et al., 2025), MiniCPM-V:8B (Yao et al., 2024), mPLUG-Owl3:7B (Ye et al., 2024), LLaVA-NeXT:7B (Zhang et al., 2024), Gemma3:12B (Team et al., 2025a), and GLM-4.1V-Thinking:9B (Team et al., 2025b). All models are evaluated in a zero-shot setting to assess their generalization ability under a unified protocol.

Evaluated Metrics: We evaluate model performance using three types of metrics: overall accuracy, SAQ accuracy under deterministic matching,

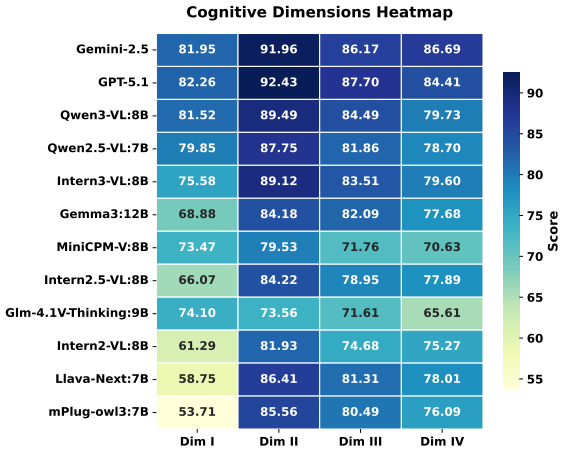


Figure 4: Performance heatmap across four cognitive dimensions. Darker colors indicate higher scores.

and inference time. For MCQs, we report standard accuracy. For SAQs, we do not use an MLLM-as-a-judge protocol. Although the SAQ format is open-ended, the task is essentially subject classification, with answers restricted to 10 predefined subject categories. We therefore adopt a deterministic rule-based matching strategy: a prediction is counted as correct if the generated text contains the correct subject name or its common abbreviation. This design avoids additional judgment bias and ensures objectivity and reproducibility.

Beyond raw accuracy, we introduce the Cognitive Fidelity Score (CFS) to measure whether a model exhibits balanced pedagogical competence across the four cognitive dimensions defined in Sec. 3.2, namely Dim I (D_1), Dim II (D_2), Dim III (D_3), and Dim IV (D_4). A reliable pedagogical assistant should not only perform well on average, but also maintain consistency across dimensions rather than excel in only a subset while failing in others. To capture this property, we define CFS as the geometric mean of the four dimensional scores penalized by their dispersion. Formally, let $\mathbf{S} = [S_{D_1}, S_{D_2}, S_{D_3}, S_{D_4}]$ denote the score vector over the four dimensions; then CFS is defined as:

$$\mathrm{CFS} = \left(\prod_{i=1}^{4} S_{D_i}\right)^{\frac{1}{4}} - \sqrt{\frac{1}{4}\sum_{i=1}^{4}\left(S_{D_i} - \bar{S}\right)^2} \quad (1)$$

where $\bar{S}$ is the arithmetic mean of the dimensional scores. In this way, CFS rewards strong overall performance while penalizing lopsided cognitive profiles. In addition, we record inference time as a supplementary efficiency metric and report the detailed comparison in the appendix.

Implementation Details: All experiments were conducted on a server equipped with 8 NVIDIA A100 GPUs. To ensure fairness and efficiency, we uniformly employed BF16 precision for inference, sampling 16 frames evenly from the video as visual context input. For decoding, we used the official default chat template for each model to guarantee result reproducibility.

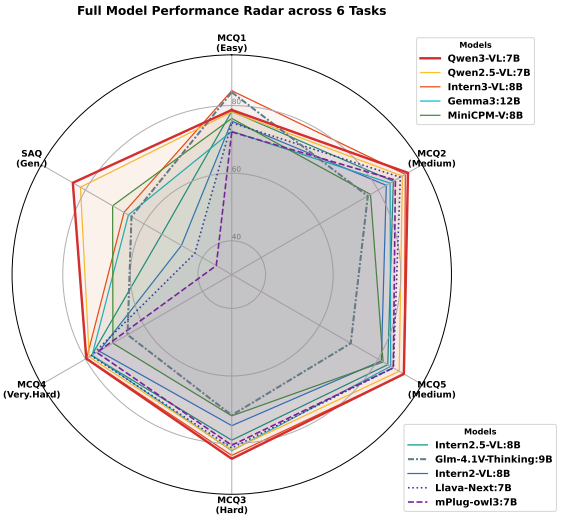


Figure 5: Task-level performance radar. The chart displays model accuracy across six tasks arranged by difficulty.

4.2 Overall Performance

We first present the overall results. Table 3 summarizes the main benchmark results, and the full breakdown is provided in Appendix D. Three observations emerge:

First, the two proprietary frontier models achieve the strongest overall performance. GPT-5.1 obtains the highest Average Accuracy (86.70), while Gemini-2.5 achieves the best CFS (83.07), indicating a stronger balance across cognitive dimensions. Among open-source models, Qwen3-VL:8B ranks highest in both Avg and CFS, demonstrating the strongest all-around capability within the open-source group.

Second, the comparison between Avg and CFS reveals cognitive-profile differences that conventional accuracy alone cannot capture. MiniCPM-V improves its ranking under CFS, suggesting that although its absolute accuracy is lower, its performance is more evenly distributed across cognitive dimensions. In contrast, models such as LLaVA-

Setting	Math	Phys	Chem	Bio	Geog	Tech	Hist	Eng	Chi	Pol	Avg	CFS
Text-Only (Blind)	69.25	62.50	58.91	68.61	73.78	66.52	68.89	66.64	66.52	70.42	67.20	44.44
Audio-Only	74.19	77.47	73.94	75.97	77.34	63.65	77.29	67.86	75.51	76.94	74.02	64.67
Qwen3-VL:8B (Full)	84.68	85.29	80.14	83.76	88.45	74.09	86.46	84.38	85.66	85.16	83.81	80.03

Table 4: Modality ablation on Qwen3-VL:8B. Removing visual input leads to clear drops in both Avg and CFS, confirming the importance of visual grounding on PedagogyBench.

Model	Math	Phys	Chem	Bio	Geog	Tech	Hist	Eng	Chi	Pol	Avg	CFS
GPT-5.1 (Chinese QA)	87.00	86.39	86.17	87.16	88.54	77.82	86.47	88.94	86.10	86.46	86.10	82.44
GPT-5.1 (English QA)	87.95	87.83	87.43	87.99	89.39	75.51	86.36	88.47	87.25	88.82	86.70	82.78

Table 5: Comparison of GPT-5.1 under English-QA and Chinese-QA settings on PedagogyBench. The two settings yield highly similar subject-wise profiles, Avg, and CFS.

NeXT and mPLUG-Owl3 drop noticeably in the CFS ranking, indicating greater imbalance across tasks. These shifts confirm that CFS captures not only overall performance, but also consistency across pedagogical competencies.

Technology remains the most challenging subject for nearly all models, especially in the open-ended SAQ setting. This gap does not indicate flawed QA design. Rather, because General Technology is inherently cross-disciplinary and practice-oriented, models often default to more intuitive subject labels such as Chemistry or Physics instead of the standardized curriculum category *Technology*. This pattern further highlights the gap between option-guided discrimination and precise open-ended generation.

4.3 Modality Contribution Analysis

To examine whether PedagogyBench can be solved mainly through language priors, we conduct a modality ablation on Qwen3-VL:8B under three settings: Text-Only (Blind), Audio-Only, and Full (Audio+Visual). As shown in Table 4, removing visual input causes substantial drops in both Avg and CFS. In particular, the text-only setting drops from 83.81 Avg / 80.03 CFS to 67.20 Avg / 44.44 CFS. These results indicate that visual grounding plays an important role in pedagogical reasoning on PedagogyBench and that the benchmark is not reducible to language priors alone.

We further test the effect of the cross-lingual setting by re-evaluating GPT-5.1 on a fully Chinese version of the QA pairs. The results remain highly consistent with the original English-QA setting, suggesting that the main challenge of PedagogyBench lies in multimodal pedagogical reasoning rather than superficial language mismatch.

4.4 Cognitive Dimensions Analysis

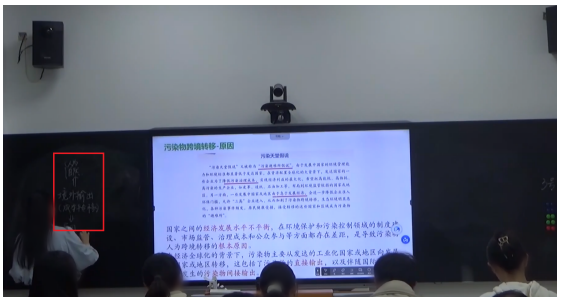
Figure 4 shows model performance across the four levels of the cognitive pyramid. A consistent pattern is that Dim II scores are generally higher than Dim I scores. This suggests that current MLLMs can often rely on the linguistic priors of their LLM backbones for interpretation, while still facing bottlenecks in fine-grained visual perception. As a result, some models can produce plausible reasoning while remaining weak in grounding their answers in visual evidence.

We observe two broad capability profiles. Models such as Qwen3-VL and MiniCPM-V exhibit relatively balanced scores across dimensions, indicating more stable vision-language integration. By contrast, models such as LLaVA-NeXT show larger variance across dimensions, with weaker performance on perception-related tasks. This pattern suggests a stronger dependence on textual priors and less reliable visual grounding.

These dimension-level differences help explain the rank shifts between Avg and CFS. MiniCPM-V benefits from its more even distribution of performance, whereas models with sharper dimension gaps receive larger CFS penalties. Overall, the results suggest that robust pedagogical reasoning depends not only on strong average accuracy, but also on balanced competence across perception, interpretation, application, and diagnosis.

4.5 Generative Gap and Failure Mode Analysis

To examine fine-grained model behavior, Figure 5 visualizes performance across six tasks arranged by cognitive complexity. A clear generative gap emerges: models perform relatively well on discriminative tasks, especially MCQs, but deteriorate



Dimension	Mixed (Dim I–Dim IV)
Phase	Interaction
Model	GLM-4.1V-Thinking:9B
Performance	**Discriminative (MCQ): 100% Accuracy** *The model correctly answers all five MCQs related to pollution transfer and teaching strategy.*
SAQ Question	Based on the visual content and teaching context, identify the specific subject being taught.
Ground Truth	Geography
Prediction	Chemistry
Analysis	**Textual Bias Over Visual Evidence.** The model is misled by the keyword "Pollutant" in the slide title and retrieves a chemistry prior, while failing to ground its answer in the visible blackboard notes and classroom context.

Figure 6: A representative failure case from PedagogyBench. Despite perfect MCQ performance, GLM-4.1V-Thinking:9B fails on the open-ended SAQ because language cues override visual grounding.

sharply on the open-ended SAQ setting. This suggests that many MLLMs are considerably stronger at option-guided discrimination than at precise free-form generation grounded in multimodal evidence.

Performance also declines steadily as tasks move from factual recall to higher-order reasoning. This radial retraction supports the difficulty design of PedagogyBench: as models progress from perception and interpretation to application and diagnosis, their performance weakens correspondingly. Only models with stronger vision-language alignment remain relatively robust under higher cognitive load.

We further observe a recurring speech/text interference effect. In several failure cases, models rely excessively on salient spoken or textual cues while underweighting decisive visual evidence, leading to plausible but incorrect answers driven by language priors rather than grounded classroom context.

Figure 6 shows a representative example from the Interaction phase of a Geography lesson. Although GLM-4.1V-Thinking:9B answers all MCQs correctly, it fails on the SAQ subject-classification task, predicting "Chemistry" instead of "Geography". The keyword "Pollutant" in the slide title biases the model toward a chemistry prior, while the visible blackboard cues and broader classroom context clearly indicate a human geography lesson. This case highlights two recurring patterns in PedagogyBench: strong discriminative performance does not guarantee faithful open-ended generation, and language cues can override visual grounding in pedagogical reasoning.

We report additional qualitative case studies in Appendix E and inference-efficiency comparisons in Appendix F.

5 Conclusion

We present PedagogyBench, a benchmark for multimodal instructional video understanding grounded in pedagogical cognition. The benchmark combines pedagogy-driven segmentation, dual-stream semantic injection, and a cognitive evaluation framework with four levels covering 20 fine-grained tasks. We also introduce the CFS to measure the balance of model performance across cognitive dimensions.

Experiments on 12 MLLMs reveal a clear generative gap: although current models perform relatively well on discriminative tasks, their performance drops on open-ended pedagogical reasoning and diagnosis, where they often rely on linguistic priors or parametric memory instead of grounded visual evidence. We hope PedagogyBench will provide a useful testbed for studying multimodal pedagogical reasoning and support future work on educationally grounded MLLMs.

Limitations

While PedagogyBench provides a useful benchmark for instructional video understanding, we note several limitations.

Subject and Cultural Scope: Although the benchmark covers 10 K–12 subjects, the current data is primarily drawn from standard Chinese classroom settings. Since pedagogical practice is shaped by cultural and curricular contexts, future versions will incorporate more cross-cultural instructional data to improve generalizability.

Interactivity Setting: PedagogyBench currently focuses on single-turn QA in order to isolate and measure discrete pedagogical cognitive abilities.

This static setting provides a necessary starting point, but it does not yet capture the multi-turn nature of real instructional scaffolding.

Granularity of Modal Analysis: Our current evaluation emphasizes multimodal benchmark performance rather than fine-grained modality attribution. Although we include modality ablations, more systematic analyses of the respective contributions of visual, audio, and textual signals remain for future work.

Sample Size: High-level pedagogical annotation requires scarce interdisciplinary expertise and is therefore costly to scale. While the current benchmark is sufficient to reveal major performance trends and model failure patterns, larger-scale expansion would further strengthen the robustness of the conclusions.

Ethical Considerations

The dataset used in this study is derived from archived videos of a teaching skills competition held by the authors' institution. The data collection process received formal approval from the university's administrative authorities. Because the videos were recorded in a public competition setting, participants were informed in advance that the recordings might be used for pedagogical assessment and academic research.

To maximize the protection of personal privacy, we implemented strict screening and processing protocols. The selected video segments primarily capture the back or side views of the teachers. We strictly excluded any segments containing clear frontal faces. Furthermore, All experts involved in data annotation signed confidentiality agreements and received compensation above the local minimum wage standard applicable to their location.

Although this study focuses on evaluation rather than model training, we acknowledge that the pedagogical norms may reflect specific regional styles. Users should be aware of this potential bias when using PedagogyBench. We recommend broader fairness testing before deploying models in educational scenarios. Finally, AI tools were used only for language polishing.

Acknowledgments

This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2026C02A1236), the National Natural Science Foundation of China (No. 62577050), and the Jinhua Major Science and Technology Project (No. 2024-1-005).

References

Lorin W. Anderson. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, et al. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, et al. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Siddhant Bansal, Chetan Arora, and C. V. Jawahar. 2022. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision*, pages 657–675, Cham. Springer Nature Switzerland.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, et al. 2024. On Scaling Up a Multilingual Vision and Language Model . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14432–14444.

Yuyan Chen, Jiyuan Jia, Jiaxin Lu, Siyue Li, Yu Guan, Ming Yang, and Qingpei Guo. 2025a. VQAGuider: Guiding multimodal large language models to answer complex video questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 7821–7834, Vienna, Austria. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, et al. 2025b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Nikita Dvornik, Isma Hadji, Ran Zhang, Konstantinos G. Derpanis, Richard P. Wildes, and Allan D. Jepson. 2023. Stepformer: Self-supervised step discovery and localization in instructional videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18952–18961.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, et al. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, et al. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. In *INTERSPEECH*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18973–18990.

Robert E Haskell. 2000. *Transfer of learning: Cognition, instruction, and reasoning*. Academic Press.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *Preprint*, arXiv:2501.13826.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1359–1367, Los Alamitos, CA, USA. IEEE Computer Society.

Julia Klug, Simone Bruder, Augustin Kelava, Christiane Spiel, and Bernhard Schmitz. 2013. Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior and instructional adjustments. *Teaching and Teacher Education*, 30:38–46.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2025. Videochat: chat-centric video understanding. *Science China Information Sciences*, 68(10):200102.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, et al. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Karttikeya Mangalam, Raiymbek Akshkulakov, and Jitendra Malik. 2023. Egoschema: a diagnostic benchmark for very long-form video language understanding. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23.

Richard E. Mayer. 2020. *Multimedia Learning*. Cambridge University Press.

Marriner Merrill. 2002. First principles of instruction. *Educational Technology Research and Development*, 50:43—-59.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. 2024. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2026. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *Computational Visual Media*, 12(1):71–84.

OpenAI. 2025. GPT-5.1: A smarter, more conversational ChatGPT. OpenAI Blog.

Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3):255.

Jinhui Pang, Xinyun Yang, Xiaoyao Qiu, Zixuan Wang, and Taisheng Huang. 2024. Mmaf: Masked multimodal attention fusion to reduce bias of visual features for named entity recognition. *DATA INTELLIGENCE*, 6(4):1114–1133.

Yang Qin, Huiming Xie, Yujie Li, Benying Tan, and Shuxue Ding. 2025. Enhancing intermodal interaction for unified vision-language understanding and generation. *DATA INTELLIGENCE*, 7(2):358–380.

Lee S Shulman. 1986. Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2):4–14.

Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1207–1216.

Gemini 2.5 Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, et al. 2025a. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

OpenAI Team, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, et al. 2025b. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, et al. 2025. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *Preprint*, arXiv:2411.10442.

Haiwan Wei, Yitian Yuan, Xiaohan Lan, Wei Ke, and Lin Ma. 2025. Instructionbench: An instructional video understanding benchmark. *Preprint*, arXiv:2504.05040.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, et al. 2025. The rise and potential of large language model based agents: a survey. *SCIENCE CHINA Information Sciences*, 68(2):121101–.

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 Conference on Computer Vision and Pattern Recognition*, pages 9777–9786.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, page 1645–1653, New York, NY, USA. Association for Computing Machinery.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

Garry Yang, Zizhe Chen, Man Hon Wong, Haoyu Lei, Yongqiang Chen, Zhenguo Li, Kaiwen Zhou, and James Cheng. 2025. Mesh - understanding videos like human: Measuring hallucinations in large video models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 4827–4836, New York, NY, USA. Association for Computing Machinery.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: a dataset for understanding complex web videos via question answering. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*.

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

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: A strong zero-shot video understanding model.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3532–3540.

A Subject-Specific Details and Data Distribution

To ensure a rigorous and fair evaluation of MLLMs across diverse academic domains, PedagogyBench includes 1,852 instructional video segments, spanning 5,692.5 minutes and 11,112 QA pairs covering 10 K–12 subjects. We provide a detailed visualization of the dataset's characteristics, specifically focusing on subject distribution, temporal composition, and cognitive dimensions.

A.1 Subject Distribution

Figure 7 illustrates the distribution of video segments across the 10 subjects: Biology, Chemistry,

Chinese, English, Geography, History, Mathematics, Physics, Politics, and Technology.

As shown in Figure 7, PedagogyBench maintains a high degree of categorical balance. The sample size for most subjects is controlled at approximately 10%, with Mathematics contributing 200 segments, Chinese and English each contribute 196 segments, and even Geography, the subject with the fewest samples, contains a substantial 144 segments. This balanced distribution helps mitigate domain bias, prevents the evaluation from skewing toward high-resource subjects, and better reflects cross-disciplinary generalization.

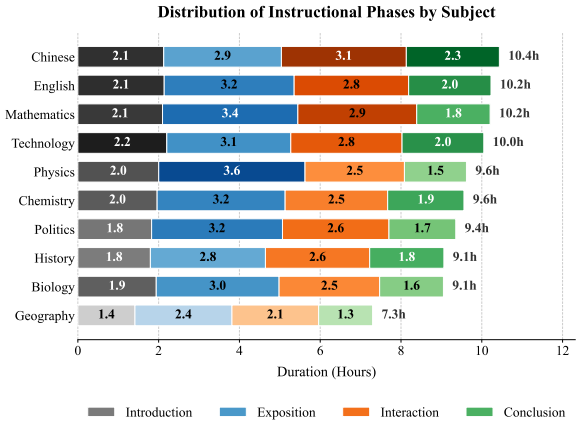


Figure 7: Subject-wise distribution of video segments. The dataset maintains a balanced composition across 10 core K–12 disciplines to ensure unbiased evaluation.

A.2 Video Duration and Pedagogical Phases

Figure 8 reports the total video duration for each subject and illustrates the content into four pedagogical phases. The statistical results indicate that the Exposition and Interaction phases occupy the vast majority of the duration, with shorter durations for the Introduction and Conclusion phases. This structure is consistent with the logic of knowledge transmission and interactive learning in real classrooms, further supporting the realism of the dataset.

In terms of subject duration, Chinese exhibits the longest total duration at 10.4 hours, followed closely by Mathematics and English, both at 10.2 hours. This reflects the central role of these core subjects in the curriculum and their inherent knowledge richness. In contrast, Geography has a relatively shorter duration, totaling only 7.3 hours.

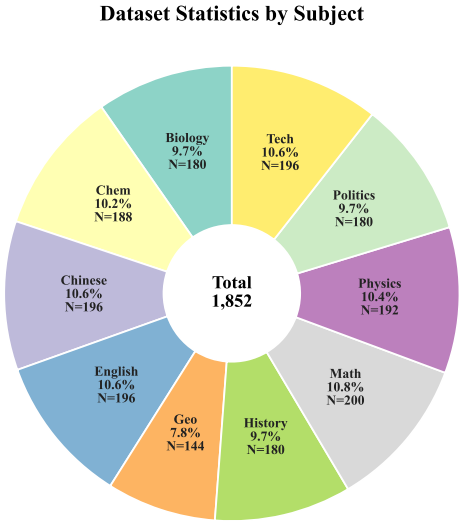


Figure 8: Total video duration per subject, segmented by pedagogical phases. The dominance of Exposition and Interaction phases reflects the authentic structure of classroom instruction.

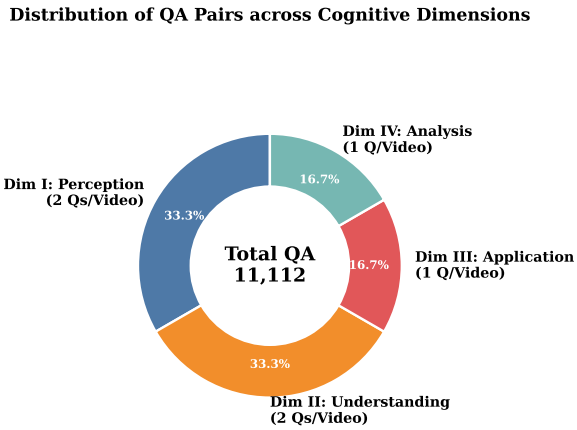


Figure 9: Distribution of QA pairs across cognitive dimensions. The structured "2:2:1:1" allocation forms a Cognitive Pyramid, using a solid perception foundation to support higher-order analysis.

A.3 Cognitive Dimension Distribution

Figure 9 displays the proportional distribution of QA pairs across the four cognitive dimensions (Dim I–Dim IV). To comprehensively assess the progressive cognitive chain from perception to analysis, we employed a structured QA allocation strategy during data construction. For each video segment, six evaluation questions are generated with a fixed distribution: Dim I and Dim II together account for four questions (66.7%), forming the

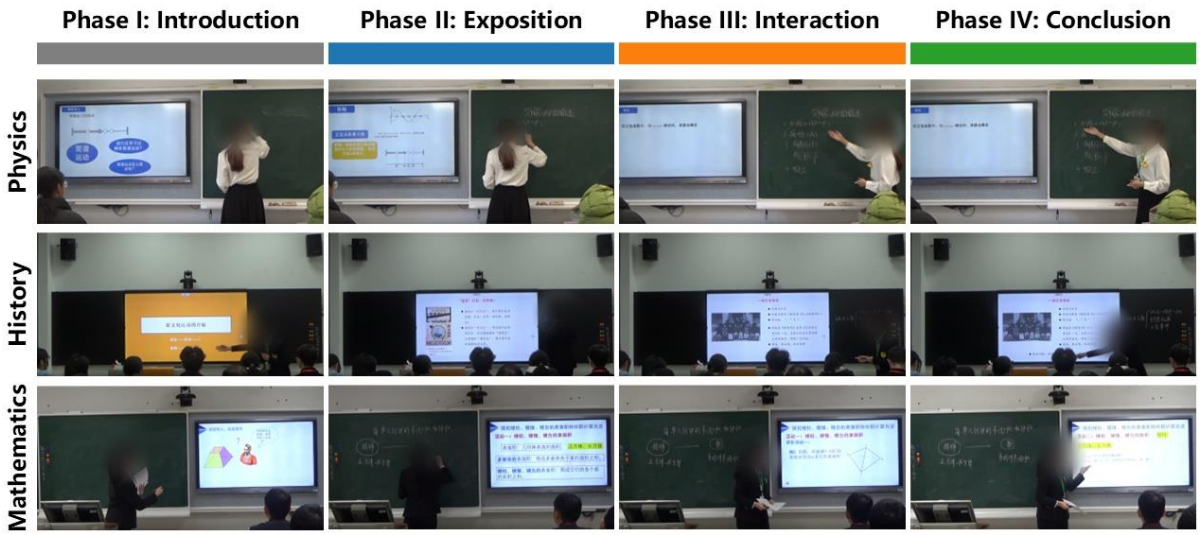


Figure 10: Visual examples of pedagogical phases across Physics, History, and Mathematics. The frames illustrate domain-specific visual patterns.

perceptual foundation of the evaluation by verifying whether the model can accurately extract basic audiovisual facts. Dim III and Dim IV together account for two questions (33.3%), with one question assigned to each dimension. These questions assess the model's capabilities for application transfer and diagnostic analysis in complex contexts.

This 2:2:1:1 distribution forms a robust cognitive pyramid structure. Although higher-order reasoning questions are fewer in quantity than foundational ones, their computational reasoning cost is significantly higher. This widened-base design ensures that a model's performance on high-level reasoning is credible only if supported by solid perception capabilities, thereby mitigating the possibility of models achieving high scores through mere guessing.

A.4 Visual Evidence

To provide visual evidence of our pedagogy-driven segmentation strategy, Figure 10 displays representative video frames from three distinct subjects across the four pedagogical phases. For Physics, exemplified by "Simple Harmonic Motion," the frames illustrate the diagrammatic representation of physical concepts and the process of theoretical derivation. For History, focusing on the "New Culture Movement," the visuals display courseware rich in text and imagery, highlighting the reliance of humanities subjects on multimodal documentary materials. For Mathematics, centered on the "Surface Area of Geometric Solids," the scenes highlight a typical logical teaching environment in which the Exposition phase is characterized by the sketching of geometric figures and formula calculations.

These examples illustrate cross-subject visual diversity within a unified pedagogical structure, covering both abstract concepts and documentary materials.

B Phase-Aware Prompting System

This appendix describes the phase-aware prompting system used in PedagogyBench, which simulates the observational perspective of human experts and adapts visual extraction strategies to different pedagogical phases.

B.1 Hierarchical Prompting Strategy

Table 6 presents the base instruction, a common prefix module for prompts across all pedagogical phases. This instruction explicitly defines the role of the vision-language model as a multimodal pedagogical analysis expert, enforcing strict protocols for action-content binding and scientific notation. In particular, the scientific notation protocol mandates that the model convert all mathematical formulas and chemical equations extracted from blackboard or courseware into standard LaTeX format. This design is critical for ensuring the accuracy of downstream reasoning in STEM subjects.

Below we present phase-specific prompts for Introduction, Exposition, Interaction, and Conclusion.

Table 7 presents the prompts for the Introduction phase, where the model acts as a visual foren-

Base Instruction Prompts (Shared across all phases)

Role: Act as an expert Multimodal Pedagogical Analyst specializing in classroom observation.

Task: Perform a fine-grained visual-temporal analysis of this video segment.

Core Requirement 1 [Action-Content Binding]: You must explicitly link the teacher's physical actions to specific visual content. Do not just say 'The teacher pointed at the board'. Instead, specify EXACTLY what is being indicated (e.g., 'The teacher's finger traced the trajectory of the sine wave', 'The teacher circled the variable x in the equation').

Core Requirement 2 [Visual Necessity]: Focus on capturing visual details that cannot be inferred from audio alone (e.g., exact formulas written, color coding used, spatial layout of the diagram).

Output Structure: Synthesize a coherent narrative describing the evolution of visual elements (slides, blackboard, gestures) from Beginning to Middle to End. Ensure every described action is grounded to a specific visual object.

Core Requirement 3 [Scientific Notation Protocol]: When describing formulas, chemical equations, or physical units, you MUST strictly follow standard LaTeX format.
- Math: Use x^2 for superscripts, a_1 for subscripts. Do not use text descriptions like 'x squared'.
- Chemistry: Use H_2O for molecules.
- Physics: Ensure units are separated, e.g., 10 m/s.
Transcription Rule: If the handwriting is messy, infer the correct standard symbol based on context. Do NOT hallucinate non-existent indices.

Table 6: The foundational System Prompt applied across all pedagogical phases. It enforces strict constraints on action-grounding and scientific notation formatting.

Introduction Phase: Prompts

[Round 1: Visual Fact Extraction]
Please analyze this video clip and perform visual fact extraction. This is the 'Introduction' phase of the lesson. Before summarizing, act as a 'Visual Forensic Expert' to list the raw facts. Focus on:
1. **The Hook Object**: What exact object/image initiated the lesson? (Transcribe text or describe the image in detail).
2. **The Title Event**: What is the exact Title written on the board/screen? Describe the teacher's action when writing/showing it (e.g., 'Underlined it twice').
3. **The Objectives**: List the bullet points of Learning Goals if visible.
4. **Key Gestures**: Record any crucial hand movements that linked the teacher to the board content (e.g., 'Pointing to the date', 'Holding a beaker').
Output Requirement: A structured chronological log of visual events, strictly linking teacher actions to board content changes.

[Round 2: Visual Summary Generation]
Task: Generate the Final Visual Summary (summary_vis)
Based on your observation, synthesize a cohesive paragraph that describes the visual narrative of this 'Introduction' phase. This text will be combined with the audio transcript, so DO NOT repeat what is being said. Instead, focus on the VISUAL context.
Drafting Guidelines:
1. **Start with the Scene**: Describe the teacher's initial state and the opening visual stimulus (The Hook).
2. **Describe the 'Doing'**: Replace generic verbs with specific actions. (e.g., instead of 'He explained', use 'He held up a model of a molecule' or 'He pointed to the equation on the left').
3. **Capture the Transition**: Explicitly describe how the visual aid shifted from the 'Hook' to the 'Main Topic' (e.g., 'The slide switched from a photo of a bridge to the title: Forces in Physics').
4. **Grounding References**: If the teacher gestured to something, specify it. (e.g., 'While speaking, the teacher circled the keyword "Velocity" on the board').
5. **Analysis of the teaching process**: explain how the introduction link connects the new knowledge to the student's existing experience or future learning, and how the knowledge point is introduced.
Output Format: A structured paragraph following the 'Hook -> Bridge -> Topic' logic.

Table 7: Specific prompts for the **Introduction** phase, focusing on capturing hook strategies and topic initiation.

sic expert to capture opening stimuli, lesson titles, teaching objectives, and other cues that mark how the lesson begins.

Table 8 presents the prompts for the Exposition phase, where the model acts as a cognitive process auditor to track the dynamic derivation of knowledge and resolve deictic gestures into specific visual referents.

Table 9 presents the prompts for the Interaction phase, where the model acts as a dialogue audi-

Exposition Phase: Prompts

[Round 1: Visual Fact Extraction]

Phase 2: Exposition - Knowledge Flow Tracking
In this phase, the teacher transforms abstract concepts into concrete knowledge. Act as a 'Cognitive Process Auditor' to extract the step-by-step visual evolution. Focus on:
1. **The Core Artifact (Static)**: What is the MAIN visual anchor? (e.g., A geometric diagram? A chemical formula? A historical map?). Describe its initial state in high detail (labels, colors, structure).
2. **The Derivation Process (Dynamic)**: Track the teacher's 'Pen-and-Voice' synchronization.
- **Additions**: What did the teacher add to the board while speaking? (e.g., 'Drew an auxiliary line connecting A to C', 'Wrote the coefficient 2').
- **Modifications**: Did the teacher erase or modify any part to show change?
- **Deictic Gestures**: When the teacher says 'this' or 'here', what specific part of the diagram is being pointed at?
3. **The Highlighting Strategy (Emphasis)**: Record visual cues used to mark importance. (e.g., 'Boxed the final result', 'Used red chalk for the variable', 'Zoomed in on the map').
Output Requirement: A structured chronological log of visual events, strictly linking teacher actions to board content changes.

[Round 2: Visual Summary Generation]

Task: Generate the Final Visual Summary (summary_vis)
Synthesize a high-density visual narrative that reconstructs the knowledge delivery process. This text will be fused with the audio transcript, so it must provide the MISSING visual context.
Drafting Guidelines:
1. **Establish the Visual Anchor**: Start by describing the primary visual model on the screen/board.
2. **Narrate the Evolution**: Describe how the concept was built step-by-step. Use dynamic verbs like 'Constructed', 'Derived', 'Annotated'. (e.g., 'The teacher constructs a triangle ABC and draws an altitude from A to BC').
3. **Resolve Deixis (Crucial)**: Resolve all pointing gestures into specific entities. (e.g., Instead of 'He pointed at the graph', say 'He traced the rising slope of the curve to illustrate the increasing trend').
4. **Capture Emphasis**: Mention how visual highlighting aligned with the conclusion. (e.g., 'He circled the final formula $E=mc^2$ in red').
5. **Cognitive load**: Describe how visual information in videos (e.g., PPT, board books, animations) works with these strategies to reduce cognitive load.
Output Format: A dynamic knowledge point teaching construction paragraph. Describe evolution: capture the process. 'Initially, the bulletin board showed... Then, the teacher...'

Table 8: Specific prompts for the **Exposition** phase, designed to capture the dynamic visual derivation of knowledge points.

tor to capture pedagogical scaffolding and visual feedback during teacher–student exchanges.

Table 10 presents the prompts for the Conclusion phase, where the model acts as a knowledge consolidator to extract summary artifacts, review gestures, and visual cues related to homework or subsequent learning tasks.

B.2 Data Annotation Examples

To intuitively demonstrate the actual output of a phase-aware prompting system, this section presents a complete annotation record for the "Exposition" phase of a high school mathematics course video on "Exponential Functions." As shown in the example below, the instructional process is decomposed into multiple structured components: the **Audio Summary** distills the core knowledge trajectory of the course; the **Visual Facts** precisely document the teacher's blackboard writing and screen deictic gestures within different time slices; and the **Visual Narrative** reassembles these fragmented visual facts into a coherent pedagogical narrative.

This multi-layered, structured data provides a solid grounding for the subsequent generation of questions requiring deep reasoning.

[Sample: Math_Exposition]

1. Teaching Goal

Recall and explain the definition and graph characteristics of exponential functions; describe the pattern of graph changes for exponential functions with different bases; plot and analyze the symmetry of exponential functions with various bases.

2. Audio Summary

This lesson's content focuses on the definition of exponential functions, their graph characteristics, and symmetry. It begins with a review of the previous class, recalling the definition of exponential functions and the steps for graphing them. Using concrete examples, students learn how to plot exponential function graphs by the point-plotting method and observe the graph features and symmetry of exponential functions with

Interaction Phase: Prompts

[Round 1: Visual Fact Extraction]

Phase 3: Interaction - Teacher-Orchestrated Knowledge Co-construction
In this phase, the teacher probes student understanding and co-constructs knowledge. Your task is to audit the teacher's visual strategies for managing this dialogue. Focus on:
1. **The Invitation Gesture (Elicitation)**: How did the teacher visually signal the start of the interaction? (e.g., 'Made a sweeping gesture to the class', 'Pointed directly to a student', 'Tapped on a blank area of the board inviting a solution'). This is distinct from the content itself.
2. **The Visual Focus Point**: What specific visual artifact is the subject of the interaction? (e.g., 'The practice problem on slide 5', 'The chemical model on the desk', 'The historical timeline on the board').
3. **Real-time Scaffolding (Teacher's Action-While-Listening)**: While a student is responding (even verbally), what did the teacher do VISUALLY to guide them? (e.g., 'Nodded and smiled to encourage', 'Simultaneously wrote the student's spoken words on the board', 'Pointed to a relevant formula from the previous section to hint at the answer').
4. **Visual Confirmation & Elaboration (Feedback)**: How did the teacher visually validate or correct the student's input? (e.g., 'Circled the student's correct answer on the board', 'Drew a diagram to visually explain the student's error', 'Gave an emphatic thumbs-up').
Output Requirement: A structured chronological log of visual events, strictly linking teacher actions to board content changes.

[Round 2: Visual Summary Generation]

Task: Generate the Final Visual Summary (summary_vis)
Synthesize a narrative that describes how the teacher visually facilitated the application and verification of knowledge. The teacher is the main actor. This text must supply the visual orchestration missing from the audio.
Drafting Guidelines:
1. **Describe the Prompt & Invitation**: Start by describing the problem on the board and how the teacher gesturally invited students to participate.
2. **Narrate the Teacher's Facilitation**: Detail the teacher's movements and supportive gestures during the student's response. This is the core of the visual narrative. (e.g., 'While the student explained their reasoning, the teacher walked over to the board and jotted down their key terms...').
3. **Detail the Visual Feedback**: Clearly describe the teacher's final visual action that confirmed or corrected the knowledge. This is the climax of the interaction. (e.g., 'To validate the answer, the teacher gave a distinct nod and circled the final number on the screen.').
4. **Connect to the Knowledge**: Describe how the teacher's gestures linked the student's response back to the core lesson content. (e.g., '...he then drew an arrow connecting the student's correct answer back to the main theorem discussed earlier.').
5. **Constructing their own understanding**: Describe how to help students move from passively receiving knowledge to actively constructing their own understanding.
Output Format: A dynamic knowledge point teaching interactive construction paragraph. Describe evolution: capture the process.

Table 9: Specific prompts for the **Interaction** phase, focusing on visual scaffolding and feedback during teacher-student dialogue.

different bases (e.g., $y = 2^x$ and $y = (1/2)^x$). The teacher guides students to analyze and discover that these function graphs are symmetric with respect to the y-axis. By comparing the graphs, the importance of understanding function properties is emphasized, and through questioning and interaction, students deepen their understanding of the characteristics of exponential functions, helping them learn how to derive the symmetric graph from a known function graph.

3. Visual Facts

Visual-Temporal Analysis

- **Beginning:**
  - The teacher writes on the blackboard with white chalk, forming the title "§4.2 exponential function" (Section 4.2 Exponential Function).
  - The screen displays a slide with the title "introduction to new lessons" (Introduction to New Lesson) and text explaining exponential functions, including the definition $y = a^x$ where $a > 0$ and $a \neq 1$.
- **Middle:**
  - The teacher turns to face the class, gesturing towards the screen, which now shows a blank slide.
  - The teacher writes on the blackboard, but the content is not visible in the frames provided.
  - The screen updates to show a new slide with a graph of an exponential function, including axes and a curve.
- **End:**
  - The teacher points to the graph on the screen, likely explaining its features.

Conclusion Phase: Prompts

[Round 1: Visual Fact Extraction]
Phase 4: Conclusion - Visual Knowledge Consolidation
This is the final phase of the lesson, where the teacher visually summarizes and closes the session. Your task is to extract fine-grained visual facts about how key knowledge points are reorganized and highlighted. Focus on:
1. **Summary Artifact (Static View)**: What is the MAIN visual carrier of the summary? (e.g., a bullet list of key points, a completed mind map, a final version of a formula or diagram, a comparison table). Describe its overall layout and structure.
2. **Key Knowledge Items (Content View)**: List the main keywords, formulas, or headings that appear in this summary artifact. Transcribe short phrases or labels if they are clearly visible on the board/screen.
3. **Review Gestures (Dynamic View)**: How does the teacher move to visually review these items? (e.g., 'Pointed to each bullet point in order', 'Traced along the arrows in the concept map', 'Underlined the final conclusion statement'). Link each gesture to the specific text or region it refers to.
4. **Transition to Next Step**: Is there any visual element indicating homework, further reading, or next lesson? If a homework/assignment slide or board section appears, briefly transcribe what is visible (e.g., page numbers, exercise ranges).
Output Requirement: A structured list of visual facts about the final summary board/slide and the teacher's gestures over it.

[Round 2: Visual Summary Generation]
Task: Generate the Final Visual Summary (summary_vis)
Based on the visual facts above, synthesize a cohesive paragraph that describes how the teacher visually wrapped up the lesson in this 'Conclusion' phase. This text will be fused with the audio-based summary, so DO NOT repeat verbal explanations. Focus only on what can be seen.
Drafting Guidelines:
1. **Start from the Summary Artifact**: Begin by describing the main summary board/slide or diagram that is visible at the end of the lesson (e.g., 'The class ends with a slide showing three bullet points summarizing...').
2. **Describe the Visual Review of Knowledge Points**: Narrate how the teacher moves through this artifact with gestures (e.g., 'The teacher points to each bullet in sequence, briefly pausing on the final formula...').
3. **Highlight Final Emphasis**: Mention any strong emphasis actions that visually mark the core takeaway (e.g., 'He draws a box around the key concept at the center of the mind map' or 'He underlines the conclusion line twice with red chalk').
4. **Mention Visual Handover to Homework/Next Task if Present**: If a homework or next-step instruction appears visually, describe how it is presented (e.g., 'The last slide shows the homework: Page 45, problems 3 to 5...').
Output Format: A structural description of the knowledge map.

Table 10: Specific prompts for the **Conclusion** phase, capturing the visual synthesis of knowledge maps and final assignment instructions.

  - A student stands up, possibly to answer a question or participate in the lesson.

Knowledge Flow Tracking

- **The Core Artifact (Static):**
  - The main visual anchor is the graph of the exponential function displayed on the screen. It includes:
  - X and Y axes labeled.
  - A curve representing the exponential function.
  - No specific labels or equations are visible on the graph itself.
- **The Derivation Process (Dynamic):**
  - *Additions:* The teacher writes the section title on the blackboard at the beginning. The screen transitions from text-based slides to a graphical representation of the exponential function.
  - *Modifications:* No visible erasures or modifications to the blackboard or screen content during the frames provided.
  - *Deictic Gestures:* The teacher gestures towards the screen, likely indicating the graph and its components while explaining. Points to the blackboard title, emphasizing the lesson's focus on exponential functions.
- **The Highlighting Strategy (Emphasis):**
  - The teacher uses hand gestures to direct attention to the screen, particularly the graph.
  - The use of the screen to transition from text to a visual graph emphasizes the conceptual shift from definition to graphical representation.

Note: This structured visual evolution helps students connect the abstract definition of exponential functions to their graphical interpretation, fa-

cilitated by the teacher's deliberate gestures and screen transitions.

4. Visual Narrative

> Initially, the blackboard displayed the title "§4.2 exponential function" alongside a slide introducing exponential functions with the definition $y = a^x$ where $a > 0$ and $a \neq 1$. The teacher constructed the foundational concept by writing the section title on the blackboard, establishing the lesson's focus. As the lecture progressed, the screen transitioned to a blank slide, prompting the teacher to annotate verbally, likely elaborating on the properties of exponential functions. The teacher then derived the graphical representation by displaying a slide with an exponential function graph, featuring labeled axes and a characteristic curve. Throughout, the teacher resolved deixis by tracing the rising slope of the curve, illustrating the increasing trend of the function. To emphasize key points, the teacher gestured towards the graph, aligning visual cues with explanations to highlight the function's behavior. Finally, the teacher engaged a student, reinforcing the interactive learning process. This strategic use of visual aids and gestures reduced cognitive load by linking abstract definitions to concrete graphical interpretations, facilitating a comprehensive understanding of exponential functions.

B.3 Expert Refinement Procedure and Before–After Example

To further clarify the rigor of Stage II human expert refinement, we provide a concrete before-and-after example here. Building on machine pre-annotations, three PhD students with relevant subject expertise refine each segment at the instance level along three aspects: (1) teaching objective alignment, (2) key information and terminology correction, and (3) pedagogical logic and structural consistency.

More specifically, the experts revise model-generated descriptions into curriculum-aligned teaching objectives, correct common ASR errors and inaccurate visual references, and reorganize the narrative structure so that the chain among teaching objectives, visual anchors, blackboard evolution, and gesture emphasis becomes clearer and more faithful to real classroom practice. This expert refinement process is applied to all 1,852 video segments.

Table 11 shows a representative example from a Biology lesson on the applications of genetic engineering in agriculture. Compared with the machine-generated draft, the expert-refined version is more specific in instructional goals, more precise in terminology, and more coherent in pedagogical structure.

C Controlled QA Generation Engine Details

This appendix supplements Section 3.1.3 by detailing the implementation and quality-control design of the QA Generation Engine. It includes three parts: the cognitive-driven prompt template, the difficulty ladder mechanism, and a generative case study from a Physics lesson.

C.1 The Cognitive-Driven Prompt Template

To ensure that the generated assessment questions exhibit both multimodal dependency and cognitive coverage, we designed a structured system prompt. Table 12 presents the complete prompt template used to drive GPT-4o. The core feature of this template is its explicit sub-task mapping mechanism. We categorize the 20 fine-grained sub-tasks defined in Section 3.2 of the main text, such as "misconception diagnosis" and "blackboard extraction," into four cognitive dimensions. During generation, the model is instructed to construct questions based on these specific sub-task requirements, thereby preventing the generation of generic, non-pedagogical content. Furthermore, the prompt mandates that the model perform a lightweight thought process prior to question generation to plan the logic for distractor design and difficulty progression.

C.2 Difficulty Ladder

To prevent the generated questions from remaining at a superficial level of visual perception, we incorporated a mandatory difficulty ladder mechanism into the generation engine. As illustrated in the difficulty assignment module of Table 12, we eschewed random question generation in favor of a deterministic mapping between question indices and cognitive dimensions.

This design compels the model to follow a cognitive trajectory from foundational fact retrieval to higher-order pedagogical diagnosis. Specifically, Q1 and SAQ are fixed to Dimension I, focusing on fundamental perception and memory, Q2 and Q5 are mapped to Dimension II, assessing the model's comprehension capabilities, while Q3 and Q4 are mandated to target Dimensions III and IV, requiring the model to simulate the reasoning process of a teacher. This controlled difficulty design ensures that PedagogyBench evaluates the hierarchical cognitive depth of MLLMs, rather than merely testing simple pattern recognition abilities.

Dimension	Before Refinement	After Expert Refinement
Teaching objectives	Describe the application domains of genetic engineering; analyze the characteristics of transgenic pest-resistant rice.	Describe the application domains of genetic engineering **and their impact on agriculture**; analyze the characteristics of transgenic pest-resistant rice and **its role in agricultural production.**
Audio summary	Introduce the application domains of genetic engineering, present the case of transgenic pest-resistant rice, and use yellow and green rice to guide students in discussing its practical effects.	Introduce, in sequence, the four application domains of genetic engineering, then place particular emphasis on the case of transgenic pest-resistant rice, and use a **contrast between** yellow and green rice to guide students in discussing its practical effects **and significance**.
Visual facts / narrative	First display a PPT slide with the title and plant diagrams as visual anchors, then derive the concepts on the blackboard, using chalk circles and underlining to provide in-depth explanations of the content.	First display a PPT slide with the title and plant diagrams as visual anchors, then derive the **key** concepts on the blackboard, and use **colored** chalk circles and underlining to **emphasize the critical steps.**

Table 11: Before-and-after comparison of machine-generated and expert-refined annotations in Stage II. The expert-refined version is more specific in teaching objectives, more precise in terminology, and more coherent in pedagogical structure.

C.3 Generative Case Study

This section provides concrete examples from the generative outputs of a high school physics lesson on "Atomic Structure," demonstrating how PedagogyBench implements a cognitive difficulty progression from perception to cognition. At the foundational level, Q1 and the SAQ, corresponding to Dimension I, anchor the assessment in visual ground truth, ensuring the model accurately extracts the sequential writing of particles on the blackboard and classifies the subject matter. For intermediate complexity, Q2 and Q5 align with Dimension II, requiring the model to not only interpret the pedagogical intent behind the teacher's gesture toward the screen but also elucidate the logical role of specific examples, such as the photoelectric effect, in constructing the overall knowledge framework.

At the higher-order cognitive level, the generated questions demonstrate complex pedagogical reasoning capabilities. Q3 advances to Dimension III, requiring the model to simulate expert thinking to plan the optimal next step in instruction based on the current progression. Finally, Q4 reaches the pinnacle of Dimension IV, focusing on the deep diagnosis and correction of a student's typical misconception regarding the history of neutron discovery. This comprehensive coverage, progressing from shallow to deep, confirms that the framework successfully guides the model beyond mere multimodal description, enabling the execution of professional assessment tasks with genuine pedagogical depth.

[Q1] Dimension I: Perception & Remembering
Task: Blackboard Extraction
Question: What was the **first particle** the teacher wrote on the blackboard during the lesson?
Options: A. Electron B. Proton C. Neutron D. Photon
Answer: A
Explanation: The teacher first introduced the electron, as evidenced when they wrote it on the board early in the middle phase of the lesson. This reflects the historical order in which particles were discovered.

[Q2] Dimension II: Understanding & Interpretation
Task: Gesture Analysis
Question: Why did the teacher **gesture towards the screen** while adding "electron," "proton," and "neutron" to the list of particles?
Options:
A. To highlight the particles as the most important discovery.
B. To emphasize the visual relationship between the listed particles as part of the particle model of matter.
C. To direct students to take notes on the screen.
D. To introduce the concept of forces between these particles.
Answer: B
Explanation: The gesture indicated the importance of understanding how these particles form the framework of the particle model, underscoring

QA Generation Engine: System Prompt Template

[Role & Input]
You must write all questions, options, and explanations in pure English.
You are the Lead Researcher for 'PedagogyBench', a benchmark that evaluates multimodal large language models on classroom teaching videos.
Your task: Generate 5 difficult, well-structured MCQs that strictly follow the mapping to 4 cognitive dimensions and that require using the concrete details of this lesson (audio + visual), not generic teaching knowledge.
Input Context:
- [pedagogical Phase]: {phase}
- [Teaching Goal]: {teaching_goal}
- [Audio Stream Summary]: {audio_summary}
- [Visual Stream Facts]: {visual_facts}
- [Visual Narrative]: {visual_narrative}

[Cognitive Dimensions & 20 Sub-Tasks]
COGNITIVE DIMENSIONS (ADAPTED FROM BLOOM)
- **Dimension I – Perception & Remembering**: Precise recall of what is shown or said.
Target Sub-tasks: Knowledge Localization, Visual Object Recognition, Transcript Check, Blackboard Extraction, Factual QA.
- **Dimension II – Understanding & Interpretation**: Interpreting why the teacher did something.
Target Sub-tasks: Intent Classification, Concept Reasoning, Cross-modal Consistency, Summary Generation, Keyframe Description.
- **Dimension III – Application & Transfer (HARD)**: Choosing the best next pedagogical action.
Target Sub-tasks: Scaffolding Generation, Similar Problem Gen, Personalized QA, Resource Rec, Step Prediction.
- **Dimension IV – Analysis & Diagnosis (VERY HARD)**: Diagnosing hidden misconceptions.
Target Sub-tasks: Misconception Diagnosis, Strategy Analysis, Difficulty Prediction, Logic Check, Quality Assessment.

[Distractor Design & Difficulty Ladder]
DIFFICULTY ASSIGNMENT:
- Q1 -> Dimension I (easy, factual).
- Q2 -> Dimension II (medium, interpret intent).
- Q3 -> Dimension III (hard, application / next-step choice).
- Q4 -> Dimension IV (very hard, misconception diagnosis).
- Q5 -> Dimension II (medium, concept relation across audio & visual).
DISTRACTOR DESIGN RULES:
- All options must look plausible in the context of this specific lesson.
- For Q3 and Q4: Include at least one distractor that reflects a common misconception and one that is pedagogically reasonable but sub-optimal.

[Chain-of-Thought Enforcement]
LIGHTWEIGHT THOUGHT:
Inside <thought> write a brief bullet plan (<120 tokens) for how you design Q3 and Q4 to be challenging.
- Identify 2–3 key ideas from the data.
- Q3: Application / next-step choice with one distractor that is too direct.
- Q4: Student misconception with deep vs superficial diagnoses.

Table 12: The core system prompt used in the QA Generation Engine.

the integration of audio and visual elements.

[Q3] Dimension III: Application & Transfer
Task: Pedagogical Planning
Question: After completing the **list of particles** on the screen, what would be the most effective **next step** to deepen understanding of the material?
Options:
A. Demonstrate how the listed particles form an atom using a diagram.
B. Move on to explaining the photoelectric effect.
C. Ask students to discuss the differences between the listed particles.
D. Provide historical dates for when each particle was discovered.
Answer: A
Explanation: Demonstrating how particles combine to form an atom provides a natural progression. Options like (B) or (D) shift focus unnecessarily or jump too far ahead.

[Q4] Dimension IV: Analysis & Diagnosis
Task: Misconception Diagnosis
Question: A student claims that the **neutron was discovered before the proton** because it is "more fundamental." How should the teacher respond?
Options:
A. Confirm the neutron is fundamental, but shift to electrons.
B. Highlight the timeline and experimental differ-

ences (e.g., difficulty detecting neutral charge).
C. Emphasize that fundamental particles are unrelated to discovery order.
D. Explain that the neutron's neutral charge made it harder to detect.
Answer: B
Explanation: The best response explains the historical timeline and experimental challenges (neutral charge), refocusing learning on the sequence of discoveries rather than debating "fundamentality."

[Q5] Dimension II: Understanding & Interpretation
Task: Concept Connection
Question: How does the teacher's use of the **photoelectric effect** relate to the overall lesson on fundamental particles?
Options:
A. It introduces the connection between photons and electrons.
B. It serves as a side topic about light.
C. It concludes the lesson showing practical uses.
D. It illustrates the experimental methods by which particle properties are studied.
Answer: D
Explanation: It demonstrates key research methods in understanding particle behaviors, supporting the idea that science is an evolving process of building understanding through experiments.

[SAQ] Dimension I: Perception
Task: Subject Classification
Question: Based on the visual content and teaching context, identify the specific subject being taught.
Ground Truth: Physics

D Full Experimental Results and Analysis

This appendix provides the complete results for all 12 evaluated MLLMs on PedagogyBench (Tables 13–24), including raw scores across 10 subjects, 6 questions, and 4 cognitive dimensions. Based on these results, we summarize four characteristics of current MLLMs in instructional video understanding.

D.1 Disparities in Modality Dependency

Comparing dimensional scores across subjects reveals clear differences in how models rely on different modalities.

Humanities Leverage Greater Parametric Prior Knowledge: In subjects like History, models maintain high higher-order reasoning scores even when their foundational visual perception scores are low, e.g., LLaVA-NeXT scores only 49.72. This suggests that in humanities, models often bypass visual bottlenecks, drawing upon the robust parametric knowledge embedded in their LLM foundation to infer answers, demonstrating an associative problem-solving capability.

Visual Bottleneck in STEM: Conversely, in subjects like Technology and Physics, perception and reasoning exhibit strong coupling. For instance, Qwen2.5's low perception score in Technology directly leads to a collapse in its overall reasoning ability, resulting in a CFS of only 47.83. This suggests that in scientific domains, precise visual grounding is a prerequisite for effective reasoning; a failure of the visual encoder to extract critical details can lead to hallucinations.

D.2 Reasoning Depth Decay

A longitudinal analysis of score progression from Visual Fact Retrieval (MCQ1) to Misconception Diagnosis (MCQ4) verifies the limitations of current models in long-chain reasoning.

Most models excel in single-step visual perception tasks (MCQ1) but experience significant performance degradation in diagnostic tasks (MCQ4) requiring multi-step inference. Taking MiniCPM-V as an example, while it achieves a score of 80.5 in Mathematics intent understanding (MCQ2), its performance drops to 65.5 in Misconception Diagnosis (MCQ4). This indicates that as the reasoning chain lengthens, models are prone to error propagation, making it difficult to maintain the stability of the logical chain.

D.3 The Gap Between Retrieval and Contextual Application

The difference in scores between SAQ and MCQ reveals the gap in the model's memorization and application capabilities.

In knowledge-intensive subjects like Mathematics and Chemistry, models such as Qwen2.5-VL and MiniCPM-V achieve near-perfect scores on SAQs, reflecting a strong capacity for knowledge retrieval and accurate output of specific definitions or formulas.

However, when required to apply this knowledge to specific contexts in MCQ4 within the same subject, scores are notably lower. This phenomenon of high retrieval but low generalization suggests that current MLLMs often behave more like static

Subject	Q1	Q2	Q3	Q4	Q5	SAQ	Avg	D-I	D-II	D-III	D-IV	Total	CFS
Biology	74.44	92.78	82.78	85.56	94.44	100.00	88.33	87.22	93.61	82.78	85.56	87.29	**83.23**
Chemistry	78.19	87.77	80.85	82.98	87.77	99.47	86.17	88.83	87.77	80.85	82.98	85.11	**81.74**
Chinese	77.55	90.31	86.22	85.20	93.37	97.96	88.44	87.76	91.84	86.22	85.20	87.75	**85.19**
English	81.63	90.82	86.22	84.18	94.39	96.94	89.03	89.29	92.61	86.22	84.18	88.07	**84.83**
Geography	70.14	90.28	85.31	91.67	93.75	95.83	87.83	82.99	92.02	85.31	91.67	88.00	**83.97**
History	76.11	90.56	87.78	87.22	92.78	89.44	87.32	82.78	91.67	87.78	87.22	87.36	**84.15**
Math	76.00	91.00	89.50	85.50	90.00	100.00	88.67	88.00	90.50	89.50	85.50	88.38	**86.47**
Physics	81.77	90.62	85.86	86.98	93.75	99.48	89.74	90.63	92.19	85.86	86.98	88.91	**86.29**
Politics	73.89	90.56	89.44	88.33	95.00	90.00	87.87	81.95	92.78	89.44	88.33	88.12	**84.11**
Technology	79.08	92.86	87.76	89.29	96.43	1.02	74.41	40.05	94.65	87.76	89.29	77.94	**51.80**
Average	**76.88**	**90.76**	**86.17**	**86.69**	**93.17**	**87.01**	**-**	**81.95**	**91.96**	**86.17**	**86.69**	**86.69**	**83.07**

Table 13: Detailed performance breakdown for **Gemini-2.5**.

Subject	Q1	Q2	Q3	Q4	Q5	SAQ	Avg	D-I	D-II	D-III	D-IV	Total	CFS
Biology	76.11	93.33	86.67	83.33	95.00	99.44	88.98	87.78	94.17	86.67	83.33	87.99	**83.97**
Chemistry	77.66	94.15	82.45	86.17	92.55	97.87	88.48	87.77	93.35	82.45	86.17	87.43	**83.43**
Chinese	73.98	91.33	88.78	85.20	92.86	91.84	87.33	82.91	92.10	88.78	85.20	87.25	**83.68**
English	89.18	91.75	89.18	79.38	90.21	99.48	89.86	94.33	90.98	89.18	79.38	88.47	**82.72**
Geography	82.73	92.09	89.93	87.77	93.53	91.37	89.57	87.05	92.81	89.93	87.77	89.39	**87.12**
History	82.86	92.57	88.00	83.43	91.43	81.14	86.57	82.00	92.00	88.00	83.43	86.36	**82.33**
Math	76.56	93.23	90.10	82.29	90.10	98.96	88.54	87.76	91.67	90.10	82.29	87.95	**84.33**
Physics	84.04	93.09	84.57	82.45	94.68	96.81	89.27	90.43	93.89	84.57	82.45	87.83	**83.16**
Politics	81.11	90.00	90.56	89.44	90.00	89.44	88.43	85.28	90.00	90.56	89.44	88.82	**86.71**
Technology	72.96	92.86	86.73	84.69	93.88	1.53	72.11	37.25	93.37	86.73	84.69	75.51	**48.77**
Average	**79.72**	**92.44**	**87.70**	**84.41**	**92.42**	**84.79**	**-**	**82.26**	**92.43**	**87.70**	**84.41**	**86.70**	**82.78**

Table 14: Detailed performance breakdown for **GPT-5.1**.

Subject	Q1	Q2	Q3	Q4	Q5	SAQ	Avg	D-I	D-II	D-III	D-IV	Total	CFS
Biology	75.56	84.44	70.00	76.67	80.56	53.89	73.52	64.73	82.50	70.00	76.67	73.47	**66.46**
Chemistry	74.47	80.32	65.43	73.40	76.06	45.21	69.15	59.84	78.19	65.43	73.40	69.22	**61.77**
Chinese	75.00	79.08	73.47	73.47	76.53	69.90	74.58	72.45	77.81	73.47	73.47	74.30	**72.20**
English	78.57	82.65	79.08	73.98	80.10	36.73	71.85	57.65	81.38	79.08	73.98	73.02	**63.11**
Geography	85.42	83.33	76.39	78.47	82.64	44.44	75.12	64.93	82.99	76.39	78.47	75.69	**68.73**
History	77.78	80.56	75.00	76.67	82.78	43.33	72.69	60.56	81.67	75.00	76.67	73.47	**65.18**
Math	70.00	86.50	76.00	72.50	82.50	95.50	80.50	82.75	84.50	76.00	72.50	78.94	**73.90**
Physics	69.27	82.29	71.88	76.56	83.85	47.40	71.88	58.34	83.07	71.88	76.56	72.46	**62.79**
Politics	75.00	83.89	79.44	74.44	80.56	33.33	71.11	54.17	82.23	79.44	74.44	72.57	**60.66**
Technology	72.45	84.69	80.10	76.53	85.20	2.55	66.92	37.50	84.95	80.10	76.53	69.77	**47.61**
Average	**75.35**	**82.78**	**74.68**	**75.27**	**81.08**	**47.23**	**-**	**61.29**	**81.93**	**74.68**	**75.27**	**73.29**	**65.40**

Table 15: Detailed performance breakdown for **InternVL2:8B**.

Subject	Q1	Q2	Q3	Q4	Q5	SAQ	Avg	D-I	D-II	D-III	D-IV	Total	CFS
Biology	79.44	91.67	74.44	78.33	80.00	53.89	76.30	66.67	85.84	74.44	78.33	76.32	**69.09**
Chemistry	71.81	86.70	67.55	76.06	80.85	72.87	75.97	72.34	83.78	67.55	76.06	74.93	**68.77**
Chinese	80.61	81.12	80.10	78.06	79.08	81.12	80.02	80.87	80.10	80.10	78.06	79.78	**78.73**
English	86.73	81.63	79.08	73.98	83.16	64.29	78.15	75.51	82.40	79.08	73.98	77.74	**74.41**
Geography	81.94	84.72	84.72	80.56	84.72	47.22	77.31	64.58	84.72	84.72	80.56	78.65	**69.88**
History	80.56	80.56	77.78	78.89	85.00	36.67	73.24	58.62	82.78	77.78	78.89	74.52	**64.51**
Math	73.00	90.00	79.50	77.00	85.00	85.50	81.67	79.25	87.50	79.50	77.00	80.81	**76.74**
Physics	77.60	86.98	79.69	75.52	84.38	59.90	77.35	68.75	85.68	79.69	75.52	77.41	**70.99**
Politics	75.56	82.78	85.56	78.33	86.11	36.67	74.17	56.12	84.45	85.56	78.33	76.11	**63.20**
Technology	75.51	85.71	81.12	82.14	84.18	0.51	68.20	38.01	84.95	81.12	82.14	71.55	**48.69**
Average	**78.28**	**85.19**	**78.95**	**77.89**	**83.25**	**53.86**	**-**	**66.07**	**84.22**	**78.95**	**77.89**	**76.78**	**69.85**

Table 16: Detailed performance breakdown for **InternVL2.5:8B**.

Subject	Q1	Q2	Q3	Q4	Q5	SAQ	Avg	D-I	D-II	D-III	D-IV	Total	CFS
Biology	84.40	90.60	80.60	77.20	86.10	63.90	80.47	74.15	88.35	80.60	77.20	80.08	**74.61**
Chemistry	80.30	87.20	76.10	76.10	82.40	80.90	80.50	80.60	84.80	76.10	76.10	79.40	**75.70**
Chinese	86.70	89.80	85.20	79.60	90.30	75.00	84.43	80.85	90.05	85.20	79.60	83.93	**79.72**
English	88.80	91.30	84.70	77.00	88.80	87.20	86.30	88.00	90.05	84.70	77.00	84.94	**79.82**
Geography	88.90	88.20	90.30	86.80	90.30	74.30	86.47	81.60	89.25	90.30	86.80	86.99	**83.56**
History	87.80	88.90	84.40	81.70	87.80	63.90	82.42	75.85	88.35	84.40	81.70	82.58	**77.90**
Math	78.50	90.00	81.50	79.50	87.50	89.00	84.33	83.75	88.75	81.50	79.50	83.38	**79.86**
Physics	84.40	89.60	86.90	78.10	91.10	83.90	85.67	84.15	90.35	86.90	78.10	84.88	**80.27**
Politics	82.80	88.90	83.30	78.90	94.40	48.30	79.43	65.55	91.65	83.30	78.90	79.85	**69.83**
Technology	81.00	88.80	82.10	81.10	90.30	1.50	70.80	41.25	89.55	82.10	81.10	73.50	**51.52**
Average	**84.36**	**89.33**	**83.51**	**79.60**	**88.90**	**66.79**	**-**	**75.58**	**89.12**	**83.51**	**79.60**	**81.95**	**76.80**

Table 17: Detailed performance breakdown for **InternVL3:8B**.

Subject	Q1	Q2	Q3	Q4	Q5	SAQ	Avg	D-I	D-II	D-III	D-IV	Total	CFS
Biology	77.22	91.67	76.67	76.11	84.44	93.33	83.24	85.28	88.06	76.67	76.11	81.53	**76.13**
Chemistry	71.81	86.17	74.47	78.19	83.51	99.47	82.27	85.64	84.84	74.47	78.19	80.79	**76.00**
Chinese	73.47	86.73	82.65	79.59	87.24	85.71	82.57	79.59	86.99	82.65	79.59	82.20	**79.12**
English	81.12	86.73	81.63	74.49	86.22	96.94	84.52	89.03	86.48	81.63	74.49	82.91	**77.18**
Geography	88.89	88.89	85.42	85.42	87.50	97.92	89.01	93.41	88.20	85.42	85.42	88.11	**84.79**
History	82.22	88.33	83.33	81.11	90.00	91.11	86.02	86.67	89.17	83.33	81.11	85.07	**81.93**
Math	76.00	88.50	82.50	76.00	86.00	100.00	84.83	88.00	87.25	82.50	76.00	83.44	**78.51**
Physics	79.17	88.02	76.69	77.08	89.06	95.31	84.22	87.24	88.54	76.69	77.08	82.39	**76.68**
Politics	76.67	88.33	90.56	79.44	90.56	56.67	80.37	66.67	89.45	90.56	79.44	81.53	**71.32**
Technology	73.98	89.80	84.69	79.59	87.24	0.00	69.22	36.99	88.52	84.69	79.59	72.45	**47.83**
Average	**78.05**	**88.32**	**81.86**	**78.70**	**87.18**	**81.65**	**-**	**79.85**	**87.75**	**81.86**	**78.70**	**82.04**	**78.48**

Table 18: Detailed performance breakdown for **Qwen2.5-VL:7B**.

Subject	Q1	Q2	Q3	Q4	Q5	SAQ	Avg	D-I	D-II	D-III	D-IV	Total	CFS
Biology	71.70	93.30	82.80	76.70	87.20	98.90	85.10	85.30	90.25	82.80	76.70	83.76	**78.74**
Chemistry	71.30	86.20	73.40	77.70	86.20	95.20	81.67	83.25	86.20	73.40	77.70	80.14	**75.04**
Chinese	80.10	86.70	85.70	83.20	87.80	92.90	86.07	86.50	87.25	85.70	83.20	85.66	**84.13**
English	85.20	89.30	86.20	75.50	88.30	88.80	85.55	87.00	88.80	86.20	75.50	84.38	**79.00**
Geography	83.30	88.90	88.20	85.40	91.00	97.20	89.00	90.25	89.95	88.20	85.40	88.45	**86.50**
History	83.30	91.10	85.60	83.90	87.20	91.10	87.03	87.20	89.15	85.60	83.90	86.46	**84.50**
Math	73.90	91.50	87.40	74.40	88.90	99.50	85.93	86.70	90.20	87.40	74.40	84.68	**78.37**
Physics	80.70	92.20	83.90	76.00	92.20	97.40	87.07	89.05	92.20	83.90	76.00	85.29	**78.94**
Politics	80.60	90.60	85.00	83.90	90.60	81.70	85.40	81.15	90.60	85.00	83.90	85.16	**81.66**
Technology	77.00	92.30	86.70	80.60	88.30	0.50	70.90	38.75	90.30	86.70	80.60	74.09	**49.63**
Average	**78.71**	**90.21**	**84.49**	**79.73**	**88.77**	**84.32**	**-**	**81.52**	**89.49**	**84.49**	**79.73**	**83.81**	**80.03**

Table 19: Detailed performance breakdown for **Qwen3-VL:8B**.

Subject	Q1	Q2	Q3	Q4	Q5	SAQ	Avg	D-I	D-II	D-III	D-IV	Total	CFS
Biology	67.22	87.78	80.56	77.78	86.11	74.44	78.98	70.83	86.95	80.56	77.78	79.03	**73.03**
Chemistry	65.96	83.51	73.94	75.53	79.79	84.04	77.13	75.00	81.65	73.94	75.53	76.53	**73.46**
Chinese	75.00	85.71	80.10	76.02	84.18	97.96	83.16	86.48	84.95	80.10	76.02	81.89	**77.66**
English	77.55	83.67	85.81	73.47	84.69	87.24	82.07	82.40	84.18	85.81	73.47	81.46	**76.55**
Geography	81.25	82.64	83.33	82.64	83.33	68.75	80.32	75.00	82.99	83.33	82.64	80.99	**77.45**
History	73.33	83.33	83.33	81.11	84.44	43.33	74.81	58.33	83.89	83.33	81.11	76.66	**65.20**
Math	71.50	88.50	80.00	77.00	84.50	90.50	82.00	81.00	86.50	80.00	77.00	81.13	**77.62**
Physics	73.44	85.94	83.85	76.56	86.98	85.42	82.03	79.43	86.46	83.85	76.56	81.58	**77.65**
Politics	72.78	76.67	81.67	76.11	84.44	20.56	68.71	46.67	80.56	81.67	76.11	71.25	**55.18**
Technology	66.33	83.16	88.27	80.61	84.18	1.02	67.26	33.68	83.67	88.27	80.61	71.56	**44.88**
Average	**72.44**	**84.09**	**82.09**	**77.68**	**84.26**	**65.33**	**-**	**68.88**	**84.18**	**82.09**	**77.68**	**78.21**	**72.11**

Table 20: Detailed performance breakdown for **Gemma3:12B**.

Subject	Q1	Q2	Q3	Q4	Q5	SAQ	Avg	D-I	D-II	D-III	D-IV	Total	CFS
Biology	71.67	91.67	76.11	71.67	80.56	37.22	71.48	54.45	86.12	76.11	71.67	72.09	**59.67**
Chemistry	68.62	84.04	69.15	73.94	80.32	23.40	66.58	46.01	82.18	69.15	73.94	67.82	**52.88**
Chinese	71.94	82.14	80.61	78.57	82.65	93.37	81.55	82.66	82.40	80.61	78.57	81.06	**79.40**
English	82.14	85.71	84.18	73.98	85.71	68.98	80.12	75.56	85.71	84.18	73.98	79.86	**74.54**
Geography	75.00	86.81	90.28	81.25	86.81	41.67	76.97	58.34	86.81	90.28	81.25	79.17	**65.62**
History	71.11	85.56	83.33	80.00	85.00	11.67	69.45	41.39	85.28	83.33	80.00	72.50	**51.59**
Math	68.50	85.50	76.00	72.50	88.50	66.50	76.25	67.50	87.00	76.00	72.50	75.75	**68.26**
Physics	72.40	83.33	79.69	77.08	90.10	8.85	68.58	40.63	86.72	79.69	77.08	71.03	**50.30**
Politics	72.22	85.56	83.89	73.89	87.22	0.00	67.13	36.11	86.39	83.89	73.89	70.07	**46.16**
Technology	67.86	88.78	81.63	78.06	85.20	1.02	67.09	34.44	86.99	81.63	78.06	70.28	**45.17**
Average	**72.15**	**85.91**	**80.49**	**76.09**	**85.21**	**35.27**	**-**	**53.71**	**85.56**	**80.49**	**76.09**	**73.96**	**60.67**

Table 21: Detailed performance breakdown for **mPLUG-Owl3:7B**.

Subject	Q1	Q2	Q3	Q4	Q5	SAQ	Avg	D-I	D-II	D-III	D-IV	Total	CFS
Biology	71.67	90.56	82.22	76.11	82.78	40.00	73.89	55.84	86.67	82.22	76.11	75.21	**62.39**
Chemistry	72.87	86.17	70.74	75.00	80.32	42.02	71.19	57.45	83.25	70.74	75.00	71.61	**61.64**
Chinese	77.04	84.69	81.12	83.67	83.67	71.94	80.36	74.49	84.18	81.12	83.67	80.87	**76.91**
English	86.22	88.78	84.69	73.98	82.14	76.02	81.97	81.12	85.46	84.69	73.98	81.31	**76.64**
Geography	76.39	86.11	87.50	83.33	89.58	40.97	77.31	58.68	87.85	87.50	83.33	79.34	**66.24**
History	75.00	85.56	82.78	82.22	82.78	24.44	72.13	49.72	84.17	82.78	82.22	74.72	**58.60**
Math	69.00	89.50	78.50	72.50	89.50	90.00	81.50	79.50	89.50	78.50	72.50	80.00	**73.67**
Physics	73.44	86.46	79.69	77.60	88.02	26.04	71.88	49.74	87.24	79.69	77.60	73.57	**57.76**
Politics	76.11	87.22	82.22	76.11	87.78	8.33	69.63	42.22	87.50	82.22	76.11	72.01	**51.67**
Technology	70.92	91.33	83.67	79.59	85.20	6.63	69.56	38.78	88.27	83.67	79.59	72.58	**49.34**
Average	**74.87**	**87.64**	**81.31**	**78.01**	**85.18**	**42.64**	**-**	**58.75**	**86.41**	**81.31**	**78.01**	**76.12**	**64.87**

Table 22: Detailed performance breakdown for **LLaVA-NeXT:7B**.

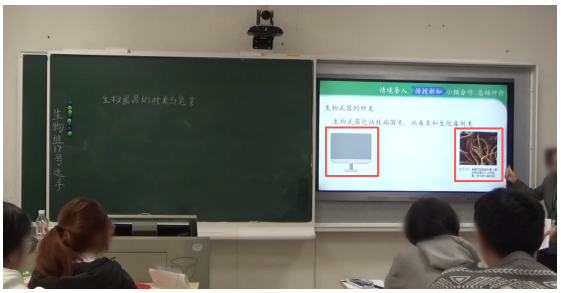
Subject	Q1	Q2	Q3	Q4	Q5	SAQ	Avg	D-I	D-II	D-III	D-IV	Total	CFS
Biology	73.33	78.89	66.11	68.89	76.11	61.11	70.74	67.22	77.50	66.11	68.89	69.93	**65.31**
Chemistry	65.43	75.00	67.02	66.49	76.06	86.17	72.70	75.80	75.53	67.02	66.49	71.21	**66.61**
Chinese	76.53	83.16	73.98	77.04	82.14	86.73	79.93	81.63	82.65	73.98	77.04	78.83	**75.24**
English	77.55	71.43	77.55	66.33	81.63	89.80	77.04	82.66	76.53	77.55	66.33	75.77	**69.60**
Geography	83.33	79.86	72.92	75.69	87.50	66.67	77.66	75.00	83.68	72.92	75.69	76.82	**72.63**
History	82.78	74.44	71.11	72.22	80.56	86.67	77.96	84.73	77.50	71.11	72.22	76.39	**70.82**
Math	74.00	80.50	67.50	65.50	84.50	99.50	78.58	86.75	82.50	67.50	65.50	75.56	**65.79**
Physics	76.56	77.60	73.96	70.31	81.25	84.90	77.43	80.73	79.43	73.96	70.31	76.11	**71.79**
Politics	80.00	78.89	74.44	73.89	85.56	44.44	72.87	62.22	82.23	74.44	73.89	73.19	**65.69**
Technology	74.89	73.98	72.96	69.90	81.63	1.02	62.40	37.96	77.81	72.96	69.90	64.66	**46.63**
Average	**76.44**	**77.38**	**71.75**	**70.63**	**81.69**	**70.70**	**-**	**73.47**	**79.54**	**71.75**	**70.63**	**73.85**	**70.33**

Table 23: Detailed performance breakdown for **MiniCPM-V:8B**.

Subject	Q1	Q2	Q3	Q4	Q5	SAQ	Avg	D-I	D-II	D-III	D-IV	Total	CFS
Biology	81.71	81.14	70.29	55.43	60.00	69.10	69.61	75.41	70.57	70.29	55.43	67.92	**59.98**
Chemistry	83.51	77.13	65.96	57.98	62.23	78.19	70.83	80.85	69.68	65.96	57.98	68.62	**59.90**
Chinese	84.18	71.94	73.98	68.88	66.33	61.73	71.17	72.96	69.14	73.98	68.88	71.24	**68.94**
English	82.14	77.04	75.00	62.76	65.82	72.96	72.62	77.55	71.43	75.00	62.76	71.69	**65.87**
Geography	89.58	75.69	74.31	67.36	77.08	63.89	74.65	76.74	76.39	74.31	67.36	73.70	**69.82**
History	81.67	76.67	71.11	75.00	72.78	66.11	73.89	73.89	74.73	71.11	75.00	73.68	**72.13**
Math	83.00	77.50	74.50	64.50	71.00	92.00	77.08	87.50	74.25	74.50	64.50	75.19	**66.58**
Physics	86.46	74.48	67.19	72.40	77.60	81.25	76.56	83.86	76.04	67.19	72.40	74.87	**68.56**
Politics	86.67	76.67	73.33	64.44	76.11	51.67	71.48	69.17	76.39	73.33	64.44	70.83	**66.20**
Technology	80.61	76.53	70.41	67.35	77.55	5.61	63.01	43.11	77.04	70.41	67.35	64.48	**50.17**
Average	**83.95**	**76.48**	**71.61**	**65.61**	**70.65**	**64.25**	**-**	**74.10**	**73.57**	**71.61**	**65.61**	**71.22**	**67.77**

Table 24: Detailed performance breakdown for **GLM-4.1V-Thinking:9B**.

knowledge bases in these cases and still have limited ability to flexibly apply retrieved knowledge under specific pedagogical contexts.



Dimension	Dim I: Perception & Remembering
Phase	Exposition
Model	InternVL3:8B
Question	What image was shown on the digital screen at the beginning of the lesson?
Options	**A.** A microscopic view of a biological specimen and a computer monitor. **B.** A slide showing a list of biological weapon categories. **C.** Bacillus anthracis under electron microscopy. **D.** A diagram comparing biological and conventional weapons.
Ground Truth	A. A microscopic view of a biological specimen and a computer monitor.
Prediction	A
Analysis	**Visual Grounding Success.** The model correctly identifies the specific tools shown on the screen and aligns them with the lesson context, demonstrating precise multimodal grounding.

Figure 11: A representative success case. InternVL3:8B correctly grounds the relevant visual content and filters out plausible distractors.

D.4 Architectural Evolution

Performance trends across the InternVL series (2.0 $\rightarrow$ 2.5 $\rightarrow$ 3.0) provide a useful view of recent architectural improvements in MLLMs.

InternVL2 exhibits high variance, typifying a specialist model with domain biases. The latest InternVL3 demonstrates significant robustness, not only raising the average CFS to 76.80 but also achieving a more uniform score distribution across subjects. This trend indicates that by incorporating stronger visual encoders and more refined instruction tuning, next-generation models are progressively overcoming early perceptual bottlenecks, evolving from domain-specific experts towards generalist models.

E Additional Case Studies

In this section, we provide additional qualitative examples from PedagogyBench to complement the main-text analysis. These cases further illustrate both the strengths and the limitations of current MLLMs across different subjects and cognitive dimensions.

E.1 Good Case: Precise Visual Grounding

Figure 11 presents a successful visual grounding case from the Biology subject during the Exposition phase. The model is required to accurately identify the content displayed on the digital screen within a complex classroom scene containing blackboard writing and multiple visual elements. InternVL3:8B correctly selects Option A, demonstrating strong visual grounding and fine-grained object recognition. This example shows that high-performing MLLMs can reliably capture pedagogically relevant visual details in PedagogyBench.

E.2 Failure Case: Multimodal Integration in Scientific Reasoning

Figure 12 shows a representative failure case in Dim III from the Exposition phase of a Physics lesson. The model must identify the critical measurement step for estimating molecule size in the oil-film experiment. This requires integrating the blackboard schematic with the teacher's spoken explanation. MiniCPM-V:8B incorrectly selects option A instead of the correct answer B. Despite the presence of a diagram indicating length on the blackboard, the model fails to align this visual cue with the teacher's explanation of microscopic thickness, confusing the size of an oil droplet with that of a molecule. This case highlights the difficulty of transforming perceptual signals into correct scientific reasoning.

Dimension	Dim III: Application & Transfer
Phase	Exposition
Model	MiniCPM-V:8B
Question	After measuring the oil film's surface area in the experiment, what should the students calculate next to estimate the size of a single oleic acid molecule?
Options	**A.** The diameter of the oil droplet before spreading **B.** The thickness of the oil film created **C.** The total number of molecules in the droplet **D.** The mass of the oleic acid used in the experiment
Ground Truth	**B. The thickness of the oil film created**
Prediction	**A**
Analysis	**Multimodal Integration Failure.** The model fails to align the teacher's spoken explanation with the blackboard schematic, incorrectly confusing the macroscopic droplet diameter with the intended film thickness.

Figure 12: A representative failure case in Dim III. MiniCPM-V:8B fails to identify the critical physical variable required for the calculation.

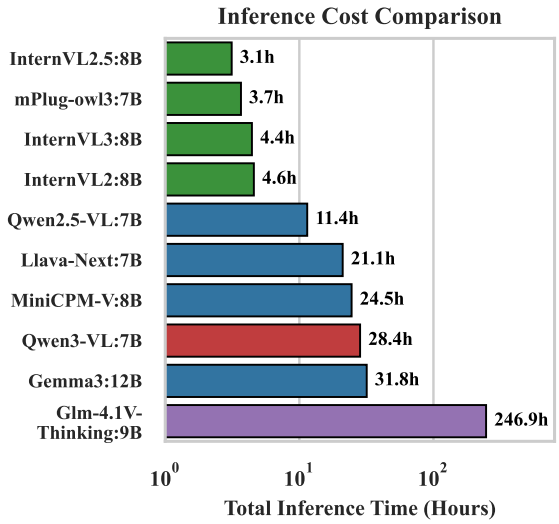


Figure 13: Inference cost comparison across the evaluated open-source models under the same controlled hardware setting.

F Inference Efficiency Analysis

Although the main text focuses on cognitive performance and failure modes, computational efficiency is also relevant for practical deployment. In this section, we report the inference latency of the evaluated open-source models on PedagogyBench under the same controlled hardware setting. Proprietary models are excluded because their latency is not directly comparable under the same controlled hardware setting.

Figure 13 shows a clear efficiency stratification. The InternVL family and mPLUG-Owl3 complete the full benchmark within approximately 3–4 hours, making them comparatively efficient among open-source models. In contrast, Qwen3-VL requires substantially more computation than Qwen2.5-VL, suggesting that part of its performance gain is accompanied by increased inference cost. GLM-4.1V-Thinking exhibits the highest latency overall, indicating that stronger reasoning-oriented generation may come with a substantial computational overhead.

These results suggest that cognitive gains and deployment efficiency remain in tension. Models with stronger reasoning capability do not necessarily provide the best trade-off for practical educational applications, especially in latency-sensitive scenarios such as real-time tutoring or classroom assistance.