

VIRA-Bench: Diagnosing Pedagogical Contingency Understanding in Online Tutoring Videos

Weihang Wang¹ Hansu Gu² Guanming Liu¹ Peng Zhang^{1*} Ning Gu¹ Tun Lu^{1*}

¹Fudan University, Shanghai, China ²Independent Researcher, Seattle, United States

whwang24@m.fudan.edu.cn hansug@acm.org gmliu24@m.fudan.edu.cn

zhangpeng_@fudan.edu.cn ninggu@fudan.edu.cn lutun@fudan.edu.cn

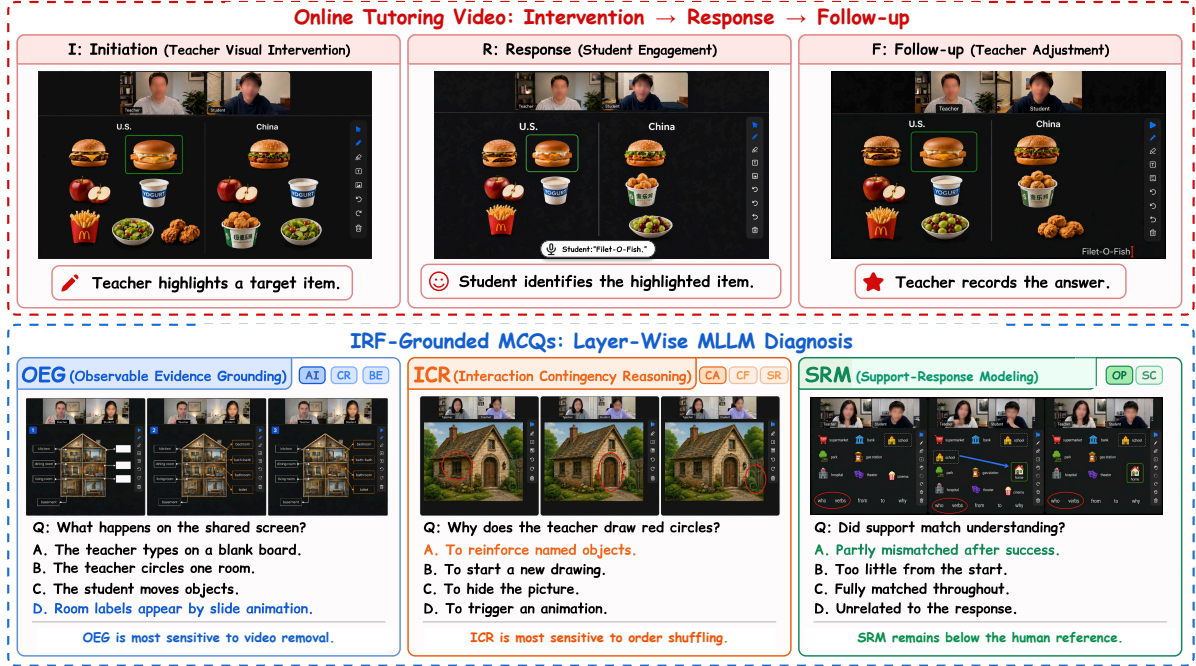


Figure 1: VIRA-Bench overview. An online tutoring clip is organized as an IRF exchange and instantiated as IRF-grounded MCQs across OEG, ICR, and SRM. The controls reveal layer-specific evidence dependence: video removal affects OEG most, order shuffling affects ICR most, while SRM remains below the human reference subset.

Abstract

Understanding online tutoring requires more than recognizing isolated events. A model must connect what a teacher says or does to how the learner responds and how the teacher subsequently adapts. Existing video benchmarks emphasize event and temporal understanding, while educational-dialogue benchmarks are largely transcript-based, leaving this interactional contingency underexplored. We introduce **VIRA-Bench**, a diagnostic benchmark for short-horizon pedagogical contingency in one-to-one tutoring videos. Grounded in the Initiation–Response–Follow-up structure and contingent teaching theory, VIRA-Bench contains 1,005 clips and 1,322 multiple-choice questions organized into three layers: Observable Evidence Grounding (OEG), Interaction Contingency Reasoning (ICR), and Support-

Response Modeling (SRM). Items are built through researcher-led evidence auditing and expert verification to keep questions tied to observable interaction evidence. Across 11 multi-modal large language models, the best model reaches 61.3% accuracy, compared with 91.0% on a stratified human subset, with ICR emerging as the most consistent weakness. Controlled evidence interventions reveal distinct dependencies across layers. Removing video hurts OEG most, while disrupting interaction order hurts ICR most. Additional controls show that this distinction survives free-response evaluation and cannot be explained by question-option shortcuts or video–transcript misalignment. By separating visual grounding from order-sensitive interaction reasoning, VIRA-Bench exposes diagnostic differences that aggregate video-QA scores obscure.

*Corresponding authors.

1 Introduction

Understanding a tutoring video requires more than recognizing its events. When a teacher highlights a word, asks a question, and reacts to the learner’s reply, a model must align language with visible actions, recover event order, identify learner uptake, and determine whether the follow-up is conditioned on that response. This multimodal NLP problem spans dialogue grounding, discourse structure, and interactional pragmatics. We refer to this focused capability as understanding **visually grounded pedagogical contingencies**.

MLLM evaluation has expanded from visual perception and question answering to long-form video comprehension, fine-grained temporal reasoning, and social interaction reasoning (Liu et al., 2023; Chen et al., 2024b; Zhang et al., 2025; Yin et al., 2024; Li et al., 2024; Fu et al., 2025; Mangalam et al., 2023; Liu et al., 2024; Cai et al., 2024; Kong et al., 2025). These benchmarks primarily examine event recognition, temporal relation recovery, or holistic social-state inference. Educational-dialogue resources model teacher and learner turns but typically rely on transcripts (Demszky and Hill, 2022; Macina et al., 2023, 2025). Neither line jointly tests whether a model can connect a visible intervention to the response it elicits and then connect that response to a subsequent adjustment. A model may therefore recognize each action and its order without recovering the interactional contingency that links them.

Online tutoring provides a natural setting for evaluating this capability. Grounded in the zone of proximal development (Vygotsky, 1978), formative assessment (Black and Wiliam, 1998; Ruiz-Primo and Furtak, 2007), and contingent scaffolding (Hattie and Timperley, 2007; van de Pol et al., 2015), effective instruction requires detecting learner understanding and adjusting support in real time. In tutoring interactions, teachers perform visible operations on a shared screen (circling, highlighting, dragging); learners respond, and teachers confirm or correct. A short clip therefore exposes an action, response, and adjustment chain with explicit visual and temporal evidence.

Table 1 uses five binary criteria. Temporal video requires ordered evidence; pedagogical interaction links teaching, learner response, and follow-up; theory anchor denotes an explicit theoretical basis; diagnostic layers separate cognitive operations; and temporal-order perturbation tests event-order

dependence. Existing resources do not combine all five.

We present **VIRA-Bench**, a diagnostic benchmark for evaluating whether video-language models can understand short-horizon pedagogical contingencies, instantiated through one-to-one online tutoring interactions. The design draws on two theoretical foundations. The Initiation, Response, and Follow-up (IRF) structure (Sinclair and Coulthard, 1975) characterizes local classroom exchanges as three-part sequences. Contingent teaching theory (van de Pol et al., 2010; Wood et al., 1976) describes instructional support as adapting to evidence of learner understanding. Based on these foundations, VIRA-Bench decomposes visually grounded interactional contingency understanding into three diagnostic layers distinguished by the cognitive operation required. **Observable Evidence Grounding** tests whether models can ground answers in observable actions and screen content. **Interaction Contingency Reasoning** tests whether models can recover directed relations between observations, such as linking an intervention to the learner response it elicits. **Support-Response Modeling** tests whether models can predict learner responses and judge whether the support level matches the learner state. The benchmark contains 1,005 tutoring clips (15–60 s each) and 1,322 multiple-choice questions across eight diagnostic subtasks, constructed through a multistage pipeline where LLMs generate candidate descriptions and questions, researchers audit evidence against source videos, and domain experts verify a stratified item sample.

We evaluate 11 commercial and open-source MLLMs under the full Video+ASR setting and analyze five representative models with ASR-only and temporally shuffled inputs. Overall accuracy ranges from 20.3% to 61.3%, leaving a 29.7-point gap between the best model and the 91.0% human reference on a stratified subset. Removing video reduces Observable Evidence Grounding accuracy by 12.7–22.0 points, whereas disrupting visual event order reduces Interaction Contingency Reasoning accuracy by 11.9–19.1 points. Free-response, STEM+OPTIONS, and synchronized-shuffle controls further probe dependencies on response format, static question cues, and video–ASR alignment. Together, these results show that the three layers expose distinct behavioral dependencies in visual grounding and temporal-order reasoning rather than collapsing into a single video-QA score.

	Temporal video	Pedagogical interaction	Theory anchor	Diagnostic layers	Temporal-order perturbation
<i>Video Understanding</i>					
MVBench (Li et al., 2024)	✓	✗	✗	✓	✗
Video-MME (Fu et al., 2025)	✓	✗	✗	✗	✗
EgoSchema (Mangalam et al., 2023)	✓	✗	✗	✗	✗
SIV-Bench (Kong et al., 2025)	✓	✓	✓	✓	✗
Social IQa (Zadeh et al., 2019)	✗	✓	✗	✗	✗
<i>Educational NLP</i>					
NCTE Transcripts (Demszky and Hill, 2022)	✗	✓	✓	✓	✗
MathDial (Macina et al., 2023)	✗	✓	✗	✗	✗
MathTutorBench (Macina et al., 2025)	✗	✓	✓	✓	✗
VIRA-Bench (Ours)	✓	✓	✓	✓	✓

Table 1: Comparison with video understanding and educational interaction benchmarks.

Our contributions are:

1. We formulate visually grounded pedagogical contingency understanding as a diagnostic evaluation problem for video-language models and instantiate it through theory-grounded one-to-one online tutoring interactions.
2. We construct **VIRA-Bench**, which comprises 1,005 tutoring clips and 1,322 MCQs across three diagnostic layers and eight subtasks, with LLM-based generation, researcher evidence audit, and expert item verification.
3. We evaluate 11 MLLMs and show, through paired evidence interventions and response-format controls, that the three layers exhibit distinct dependencies on visual content and event order.

2 Related Work

Video benchmarks. Video benchmarks now cover short-form perception, long-form understanding, and domain-specific challenges (Li et al., 2024; Fu et al., 2025; Mangalam et al., 2023; Ataallah et al., 2025; Ray et al., 2025; Shafique et al., 2025). Diagnostic suites further test reliance on temporal evidence or static shortcuts (Liu et al., 2024; Cores et al., 2025; Shangguan et al., 2024; Cai et al., 2024; Yang et al., 2025), motivating controlled input conditions. Adjacent work studies LLM reasoning and agents (Wu et al., 2026a,b,c; Yang et al., 2026b; Liu et al., 2026a,b), and calibration, hallucination, context optimization, and culture-dependent multimodal interpretation (Chen et al., 2024a; Wang et al., 2025; Yang et al., 2026a; Wang et al., 2026). Causal video QA benchmarks, however, still emphasize physical or event causality rather than response-dependent participant contingency (Xiao et al., 2021; Wu et al., 2024; Yi et al.,

2019); Appendix I provides a detailed comparison.

Educational dialogue and scaffolding theory.

Recent work has begun studying social and educational interaction in video (Kong et al., 2025; Li et al., 2025; Selvakumar et al., 2025). In educational NLP, Demszky et al. (2021) operationalize conversational uptake, the NCTE Transcripts (Demszky and Hill, 2022) provide large-scale classroom dialogue data, Demszky and Liu (2023) show NLP-powered feedback improves instruction, and MathDial (Macina et al., 2023) and MathTutorBench (Macina et al., 2025) evaluate LLM tutoring with pedagogically grounded criteria. Demszky et al. (2026) identify the gap between transcript-based NLP and video-based observation. Automated classroom observation tools such as CLASS (Pianta et al., 2008) have been adapted with LLMs (Whitehill and LoCasale-Crouch, 2024; Wang and Demszky, 2023), and ClassMind (Qu et al., 2025) extends this to video-based feedback, though on pre-segmented lessons. Dou et al. (2025) show that user simulators can proxy multi-turn evaluation, highlighting the need for complementary benchmarks that directly test interaction understanding from recorded evidence. VIRA-Bench addresses this gap by translating the IRF exchange structure (Sinclair and Coulthard, 1975) and contingent scaffolding theory (Wood et al., 1976; van de Pol et al., 2010) into a computational benchmark with paired evidence conditions.

3 The VIRA-Bench Benchmark

3.1 Task Formulation

VIRA-Bench evaluates whether video MLLMs can understand short-horizon pedagogical contingencies in one-to-one online tutoring. Given a

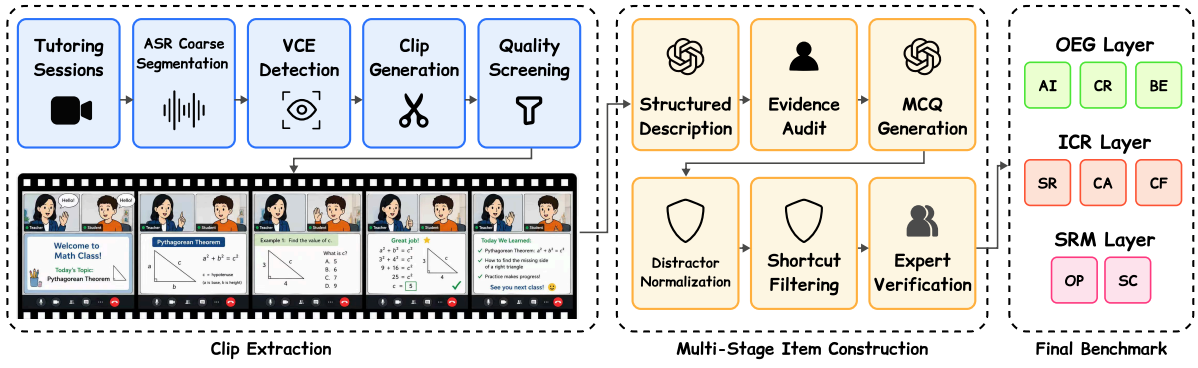


Figure 2: VIRI-Bench construction pipeline. Clip extraction combines ASR-guided segmentation, Visual Content Event detection, and quality screening. Question construction uses separate LLMs for evidence description, MCQ generation, shortcut filtering, and difficulty calibration.

15–60 second tutoring clip containing a shared screen and a student webcam feed, a model answers multiple-choice questions about the interaction. The central construct is **visually grounded interactional contingency**: a short-horizon exchange in which a teacher intervention elicits a learner response and the teacher then follows up based on that response, all inspectable from the clip. Routine responses emerge within seconds, while cognitively demanding responses may take longer (Rowe, 1986); VIRI-Bench focuses on local contingency within a single exchange rather than longitudinal scaffolding fading across sessions.

3.2 Diagnostic Framework

We design a three-layer diagnostic framework grounded in the IRF exchange structure and the contingent teaching cycle. The layers are separated by the *type of cognitive operation* required (Chi and Wylie, 2014), rather than by a simple difficulty ordering. In Observable Evidence Grounding, models must ground the interaction in observable evidence from the screen, teacher action, and student behavior. In Interaction Contingency Reasoning, they must connect these observations through temporal and contingent relations, asking whether the teacher’s intervention plausibly triggered the student’s response and whether the teacher’s follow-up targeted that response. In Support-Response Modeling, they must model the support–response relation that the exchange implies: predicting how a learner is likely to respond to a teacher’s support move, or judging whether that support level matches the learner state revealed by the response.

The eight subtasks operationalize these three cognitive operations while keeping each question targeted at one primary capability. For example,

a causal attribution question may show a teacher circling a word on the shared screen, followed by the student typing a response in the chat box. The model must distinguish whether the student’s response follows from the visual intervention, precedes it, merely coincides with it, or reflects an unsupported alternative explanation. Such questions test whether the model can use temporal evidence to recover a plausible $I \rightarrow R$ contingency, rather than only recognizing isolated events.

3.3 Clip Extraction from Tutoring Sessions

Source videos are drawn from a corpus of over 1,000 professionally recorded online English tutoring sessions, each lasting 1–2 hours, collected through video-conferencing platforms between 2018 and 2021. The released benchmark spans 72 learners and 22 teachers at primary and secondary levels. Most learners speak Mandarin as their first language, while a small number are native English speakers. Teacher speech therefore naturally includes code-switching between English instruction and Mandarin clarification. All faces are blurred to protect participant privacy, with empirically verified impact on pedagogical content visibility below 3%.

Clips are extracted through a coarse-to-fine process. First, ASR-based coarse segmentation identifies candidate instructional intervals, reducing each 1–2 hour session to approximately 15–30 minutes of likely interaction. Second, Visual Content Event (VCE) detection compares consecutive shared-screen frames using structural similarity (Wang et al., 2004) to locate moments where the teacher produces a visual intervention, such as writing, circling, highlighting, or dragging an object. Third, each VCE is used as an anchor point for

clip generation, yielding a 15–60 second window centered on a single teacher intervention and its surrounding interactional context.

The clip-level pipeline starts from 2,760 VCE anchors. After merging nearby anchors and removing duplicate windows, 1,840 candidate clips remain, filtering 33.3% of the anchors. Clip quality screening further removes 19.6% of candidates whose shared screen is unclear, whose student response window is missing, or whose interaction contains multiple overlapping exchanges. This yields 1,480 clips. After structured evidence description and researcher audit, 1,370 clips remain, with 7.4% removed due to unrecoverable missing or inconsistent evidence. Finally, question-level filtering keeps 1,005 clips with at least one valid MCQ, filtering 26.6% of clips at this stage. Extraction parameters are reported in Appendix A.

3.4 Multi-Stage Question Construction

VIRA-Bench follows a construction–evaluation separation protocol with two main stages. First, each clip is converted into an audited structured description recording the observable and temporal evidence needed for pedagogical interpretation. Second, candidate MCQs are generated from these descriptions and filtered through model-based checks and expert audit. To reduce direct construction–evaluation overlap, the two stages use disjoint model sets: GPT-4o (OpenAI, 2024) produces structured descriptions and GPT-5 (OpenAI, 2025) generates candidate MCQs, while held-out validators GLM-5V-Turbo (GLM-V Team et al., 2026) and Claude-Haiku-4.5 (Anthropic, 2025) handle shortcut filtering and diagnostic calibration. All construction models are excluded from the evaluation leaderboard.

3.4.1 Structured Evidence Description

GPT-4o receives each clip as video input together with its ASR transcript and produces a structured JSON description covering five evidence dimensions: (1) *screen events*, including timestamped teacher actions on the shared screen such as drawing, highlighting, dragging, and slide transitions; (2) *student behaviors*, including observable verbal responses, gestures, gaze shifts, and screen interactions; (3) *teacher speech acts*, including ASR-derived utterances categorized by pragmatic function such as instruction, question, feedback, and elicitation; (4) *temporal dynamics*, including the intervention–response pattern and estimated

response latency; and (5) *cognitive demand level*, describing the complexity of the task posed by the teacher.

These descriptions are treated as evidence drafts rather than final annotations. Before question generation, researchers audit each description against the source video, correct factual errors, and remove clips whose key events cannot be reliably recovered. Only audited descriptions are used for MCQ generation. This step provides an explicit evidence interface between raw videos and question construction, reducing the risk that generated items depend on unsupported interpretations.

3.4.2 Question Construction and Filtering

GPT-5 generates candidate questions from the audited descriptions, with the target subtask specified in the prompt. Distractors are produced in a separate pass under three normalization constraints that reduce superficial answer cues: (1) *length matching*, where all four options must fall within 20% of the mean option length; (2) *specificity matching*, where distractors must refer to concrete events in the clip rather than generic alternatives; and (3) *tone consistency*, where all options use the same register and hedging level. This process yields 2,400 candidate MCQs. We remove 5.8% that contain invalid option formats, repeated options, or near-duplicate questions within the same clip, leaving 2,260 questions. Full prompts and templates are provided in Appendix B.

Video-agnostic shortcut filtering. Following adversarial filtering approaches that reduce dataset biases (Bras et al., 2020; Sakaguchi et al., 2020), we test each question under a STEM+OPTIONS condition. GLM-5V and Claude-Haiku-4.5 receive only the question stem and four options, without video frames, ASR transcript, or structured evidence. Items answered correctly by both validators are treated as likely containing option-level shortcuts that allow answering without interaction evidence, and are removed. This filter reduces, but cannot eliminate, surface-level answer cues. It removes 10.6% of the remaining questions, leaving 2,020 candidates.

Evidence-supported diagnostic filtering. The two validators then inspect each question against the audited structured evidence and answer it under the full input condition. A question is removed if its gold answer is not explicitly supported by the evidence, if more than one option is compatible with

the evidence, or if the question requires information outside the clip. A correct answer from both validators does not itself trigger removal. Evidence-supported questions are regenerated only when the validators also identify evidence leakage or weak distractor discrimination. This step is used to calibrate diagnostic difficulty rather than to construct a validator-hard subset. In the final set, neither validator answers 293 items correctly, exactly one answers 587 correctly, and both answer 442 correctly. Together, this filtering stage removes 34.4% of the remaining candidates, leaving 1,326 questions.

Expert audit. Three expert auditors with backgrounds in education and second-language pedagogy independently review a 360-question stratified audit set comprising 300 random questions balanced across all layers and subtasks, and 60 high-risk questions selected from inference-heavy subtasks and cases where the two validators disagreed. Each item is judged on four dimensions: question clarity, gold-answer correctness, evidence sufficiency, and subtask alignment. Of the audited questions, 340 are accepted without change, 16 retained after minor revision, and 4 removed, yielding a 94.4% direct acceptance rate and a 1.1% removal rate. Sampling adequacy analysis and inter-annotator agreement are reported in Appendix C. After expert adjudication, 1,322 MCQs are retained in the final benchmark.

3.5 Dataset Overview

VIRA-Bench contains 1,322 multiple-choice questions over 1,005 clips, distributed across the three diagnostic layers as Observable Evidence Ground-

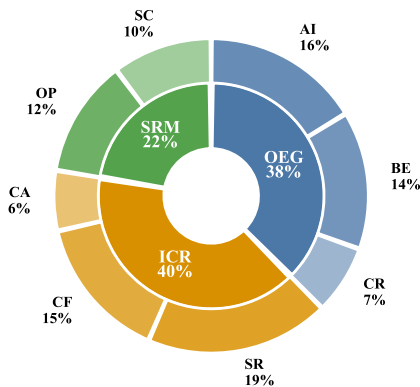


Figure 3: Distribution of 1,322 MCQs across three diagnostic layers and eight subtasks.

ing 497 questions, Interaction Contingency Reasoning 529 questions, and Support-Response Modeling 296 questions (Figure 3). Clip durations range from 15 to 60 seconds (mean 32.4 s), each centered on a single intervention and the student response it elicits. Each question has exactly four options with one gold answer; option length is controlled within 20% of the mean to reduce surface cues. All clips are drawn from 72 learners and 22 teachers, providing natural variation in interaction style and pedagogical approach within the shared one-to-one tutoring format. The benchmark scale reflects a coverage–audit tradeoff: every retained question is tied to a localized interaction window and audited evidence, rather than intended to cover all pedagogical settings. Full per-subtask distributions and cross-lingual characteristics are in Appendix D.

4 Experiments

4.1 Experimental Setup

Models and data. We evaluate 11 multimodal large language models on 1,322 MCQs from 1,005 clips, each containing a complete 15–60 second IRF exchange. The closed-source set includes Gemini 3 Pro (gemini-3-pro-preview) (Google AI for Developers, 2025c), Gemini 3 Flash (gemini-3-flash-preview) (Google AI for Developers, 2025b), Qwen3.6-Plus (Alibaba Cloud Model Studio, 2026c), Qwen3.6-Flash (Alibaba Cloud Model Studio, 2026b), and Qwen3.5-Flash (Alibaba Cloud Model Studio, 2026a). The open-source set includes Kimi-K2.5 (Kimi Team et al., 2026), MiMo-V2.5 (Xiaomi MiMo Team, 2026), and four Qwen3-VL variants—235B-A22B and 32B, each in Instruct and Thinking modes (Bai et al., 2025). All models use default inference settings with video input sampled at 1 frame per second. A human reference based on three expert answerers independently answering a stratified 200-question subset (25 per subtask) with majority voting provides an upper reference point under the same input condition (details in Appendix F; sampling adequacy in Appendix C).

Protocol and controls. Models are evaluated zero-shot under the default Video+ASR condition with video, ASR, a question, and four labeled options. Option order is shuffled for every model-item pair and the gold label is remapped. We parse the returned option letter, using normalized-text matching only as fallback. For evidence-dependency analysis, ASR TEXT ONLY removes

Model	OEG				ICR				SRM			
	AI	BE	CR	Avg	SR	CA	CF	Avg	SC	OP	Avg	All [95% CI]
Human [†]	92.0	88.0	96.0	92.0	92.0	84.0	88.0	88.0	92.0	96.0	94.0	91.0 [86.2, 94.2]
<i>Closed-source</i>												
Gemini 3 Pro	67.4	58.0	67.0	63.8	64.0	51.2	52.3	57.7	67.6	60.6	63.9	61.3 [58.7, 63.9]
Gemini 3 Flash	58.6	61.7	70.2	62.0	56.4	61.0	50.8	55.0	57.4	70.0	64.2	59.7 [57.0, 62.3]
Qwen3.6-Plus	54.9	51.6	62.8	55.1	66.0	48.8	41.6	54.3	51.5	67.5	60.1	55.9 [53.2, 58.6]
Qwen3.6-Flash	48.8	36.7	46.8	43.9	36.0	34.1	38.6	36.7	40.4	46.2	43.6	40.9 [38.3, 43.6]
Qwen3.5-Flash	24.7	18.6	22.3	21.9	18.0	40.2	24.9	24.0	41.9	20.6	30.4	24.7 [22.4, 27.1]
<i>Open-source</i>												
Kimi-K2.5	63.7	50.5	74.5	60.8	53.2	53.7	44.7	50.1	61.0	68.1	64.9	57.4 [54.7, 60.1]
MiMo-V2.5	54.0	62.8	56.4	57.7	55.2	41.5	56.9	53.7	72.8	51.2	61.1	56.9 [54.2, 59.5]
Qwen3-VL-235B-Think	39.1	48.4	52.1	45.1	39.2	42.7	38.1	39.3	44.9	46.9	45.9	43.0 [40.3, 45.7]
Qwen3-VL-235B-Inst	43.3	35.1	39.4	39.4	42.8	29.3	26.4	34.6	52.9	32.5	41.9	38.0 [35.5, 40.7]
Qwen3-VL-32B-Think	17.7	26.6	19.1	21.3	27.2	20.7	17.8	22.7	20.6	23.8	22.3	22.1 [19.9, 24.4]
Qwen3-VL-32B-Inst	22.3	18.1	14.9	19.3	13.6	32.9	23.9	20.4	31.6	13.8	22.0	20.3 [18.3, 22.6]

Table 2: Leaderboard results under the Video+ASR condition. Accuracy and 95% Wilson intervals are reported. [†] The human reference uses a stratified 200-question subset, while every model is evaluated on all 1,322 questions.

video to isolate language-only performance. VISUAL SHUFFLE WITH FIXED ASR retains all frames and the original transcript but permutes VCE-detected visual segments, disrupting the intervention–response–follow-up order and potentially introducing local video–text mismatch. SYNCHRONIZED VIDEO+ASR SHUFFLE instead permutes each segment with its aligned ASR span, preserving local cross-modal alignment. These controls do not affect leaderboard ranking. We report overall, layer-level, and subtask-level accuracy; full details are in Appendix E.

4.2 Human–Model Gap and Layer Profiles

Table 2 reports leaderboard results under the Video+ASR condition. The best model, Gemini 3 Pro, reaches 61.3% overall accuracy, leaving a 29.7-point gap below the 91.0% human reference on the stratified subset. The five leading models cluster between 55.9% and 61.3% with overlapping confidence intervals, while the remaining models range from 20.3% to 43.0%. We therefore focus on the broad human–model gap and recurring layer profiles rather than fine-grained model ranks.

Across diagnostic layers, all top-tier models follow a consistent ordering of SRM > OEG > ICR. Among the top five models, the OEG–ICR gap ranges from 0.8 to 10.7 points. This recurring profile, rather than any individual model comparison, motivates the evidence-dependency analyses below. The human reference follows the same layer ordering (SRM 94.0% > OEG 92.0% > ICR 88.0%) but at uniformly high accuracy. Its 95% interval

is 86.2–94.2%, leaving its lower bound about 25 points above the best model. This subset result supports a robust human-model gap but should not be read as a precise human estimate for all 1,322 questions. The three below-chance models do not collapse onto a single answer position. Their predicted labels span A–D, with each label accounting for 22.0–26.6% of outputs, and all 1,322 responses per model are parsed successfully. Their scores therefore reflect incorrect item-level choices rather than invalid responses or a dominant-label parsing artifact. Together, the leaderboard and layer scores establish the first experimental claim: current models remain far from the human reference, with the most consistent weakness appearing in interaction contingency reasoning.

4.3 Response-Format and Shortcut Controls

The leaderboard could be inflated if models exploit labeled choices or static question cues rather than interaction evidence. We test these alternatives with two controls. Under free response, models answer without labeled choices and a deterministic normalized-text matcher scores the output; it agrees with manual judgments on 96.0% of a 50-item audit. Under STEM+OPTIONS, models receive the question and answer options but no video, ASR, description, or construction metadata.

Figure 4 shows that free response lowers overall accuracy by 12.1 and 11.7 points, so answer options provide useful scaffolding, yet the same layer ordering persists: both models remain stronger on OEG than ICR. More decisively, STEM+OPTIONS

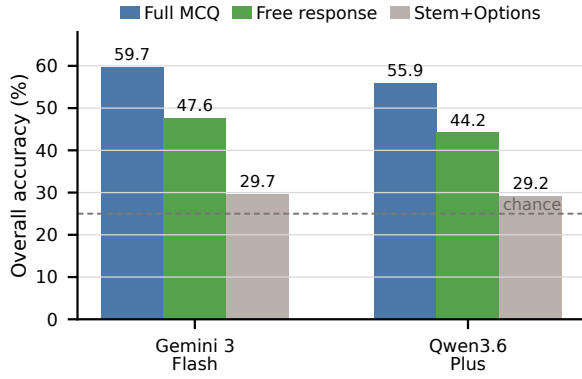


Figure 4: Overall accuracy under full multiple-choice input, free response, and STEM+OPTIONS. The dashed line marks chance accuracy; layer-level scores are reported in Appendix H.

reaches only 29.2–29.7%, compared with 55.9–59.7% under full Video+ASR input. Its layer scores are nearly flat, unlike the layer-specific profiles under full input. Gold labels are balanced across A–D, and all full-condition outputs are parsed successfully (Appendix E). Thus, response format affects absolute accuracy, but neither answer position nor static question cues explains the main model separation or the OEG–ICR pattern.

4.4 Visual Evidence for OEG

We next test the source of each layer’s evidence by evaluating five representative models under the perturbation conditions defined in §4. Table 3 first isolates what is lost when the video is removed while the complete ASR transcript remains available.

Removing video reduces OEG by 12.7–22.0 points across all five models. Action Identification drops by 16–25 points and Content Recognition by 23–40 points across the four stronger models, whereas Behavior Extraction drops by only 4–9 points. Teacher actions and screen content therefore require visual grounding, while verbal cues partially support behavior extraction. ASR-only overall accuracy remains 35.7–48.9%, confirming that transcripts are informative but cannot replace video. The intervention thus localizes the incremental value of visual evidence to OEG. Paired McNemar tests sharpen this result: every OEG decline under ASR-only input remains significant after Holm correction, whereas the corresponding ICR and SRM effects are smaller and less consistent. The no-video intervention therefore removes information preferentially required for observable

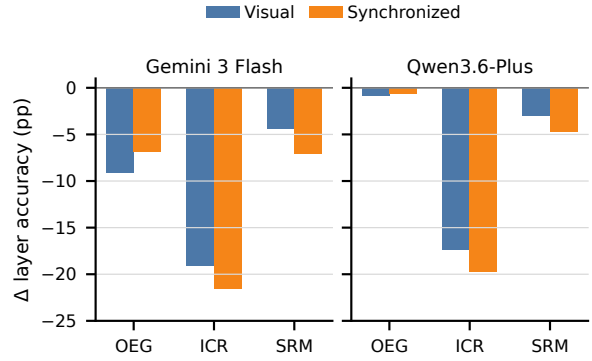


Figure 5: Layer-level changes under fixed-ASR visual shuffle and synchronized video–ASR shuffle. The synchronized control reproduces the ICR-selective degradation while preserving local cross-modal alignment.

grounding rather than imposing a uniform increase in difficulty.

4.5 Temporal-Order Evidence for ICR

Table 3 shows that VISUAL SHUFFLE WITH FIXED ASR causes the largest drops on ICR, especially Sequence Recognition (16–30 points; e.g., Gemini 56.4→32.8, MiMo 55.2→30.0). Causal Attribution and Contingent Feedback show model-dependent drops, indicating that the strongest temporal-order effect concentrates in Sequence Recognition. Accuracy therefore declines sharply when the intervention–response–follow-up order is disrupted, even though the same visual frames remain available. Across the five models, ICR falls by 11.9–19.1 points, and all five declines survive Holm correction. By contrast, four of five OEG changes and four of five SRM changes are not significant under the same perturbation. This selective pattern isolates order-bearing relations rather than uniform video degradation (Appendix G).

Figure 5 shows that the synchronized control preserves local video–text alignment and yields the same pattern. Relative to full input, ICR falls by 21.5 points for Gemini 3 Flash and 19.7 points for Qwen3.6-Plus, while OEG changes by 6.9 and 0.6 points. Cross-modal mismatch therefore cannot explain the ICR decline by itself. The measured surface-complexity factors do not account for the ICR-specific effect: humans reach 88.0%, experts accept 131/140 unchanged, stems are shortest (13.0 words), option lengths match OEG/SRM (20.0 vs. 18.8/19.5 words), and clip durations differ by at most 1.1 seconds across layers. Other unmeasured differences among layers may remain.

Model	Setting	OEG				ICR				SRM			
		AI	BE	CR	Avg	SR	CA	CF	Avg	SC	OP	Avg	All
Gemini 3 Flash	Full	58.6	61.7	70.2	62.0	56.4	61.0	50.8	55.0	57.4	70.0	64.2	59.7
	ASR Text Only	33.5	52.7	29.8	40.0	48.4	46.3	48.2	48.0	51.5	60.0	56.1	46.8
	Visual Shuffle	55.3	47.9	57.4	52.9	32.8	34.1	40.6	35.9	57.4	61.9	59.8	47.7
Kimi-K2.5	Full	63.7	50.5	74.5	60.8	53.2	53.7	44.7	50.1	61.0	68.1	64.9	57.4
	ASR Text Only	39.1	46.8	43.6	42.9	52.8	50.0	38.6	47.1	63.2	61.9	62.5	48.9
	Visual Shuffle	64.7	58.0	62.8	61.8	37.6	35.4	36.5	36.9	59.6	51.2	55.1	50.3
MiMo-V2.5	Full	54.0	62.8	56.4	57.7	55.2	41.5	56.9	53.7	72.8	51.2	61.1	56.9
	ASR Text Only	38.1	54.8	31.9	43.3	44.0	41.5	47.2	44.8	63.2	54.4	58.4	47.3
	Visual Shuffle	52.1	52.1	63.8	54.3	30.0	40.2	41.6	35.9	61.8	58.8	60.1	48.3
Qwen3.6-Plus	Full	54.9	51.6	62.8	55.1	66.0	48.8	41.6	54.3	51.5	67.5	60.1	55.9
	ASR Text Only	31.6	45.2	39.4	38.2	54.0	47.6	44.7	49.5	48.5	65.6	57.8	47.1
	Visual Shuffle	59.1	48.9	54.3	54.3	35.6	34.1	39.6	36.9	60.3	54.4	57.1	48.0
Qwen3-VL-235B-Think	Full	39.1	48.4	52.1	45.1	39.2	42.7	38.1	39.3	44.9	46.9	45.9	43.0
	ASR Text Only	32.6	34.6	27.7	32.4	36.0	41.5	27.4	33.6	45.6	44.4	44.9	35.7
	Visual Shuffle	42.3	43.1	44.7	43.1	28.0	25.6	27.4	27.4	47.8	42.5	44.9	37.2

Table 3: Evidence dependency analysis on representative models. Subtask-level accuracy is reported under Full, ASR Text Only, and Visual Shuffle with Fixed ASR. Shaded columns show layer averages.

4.6 Diagnostic Implications

The paired interventions turn layer scores into testable development targets. For OEG, the large no-video drop points to event-level alignment among speech, teacher actions, and screen changes. For ICR, the selective shuffle effect identifies preservation of actor, target, and order across intervention, response, and follow-up as the relevant evaluation target. For SRM, the best model remains 29.1 points below the human reference, motivating further tests of learner-state tracking beyond partial cues from individual channels.

These targets also define how to interpret future gains. Stronger OEG should reflect better recovery of actions and screen state from video; stronger ICR should appear under intact order and diminish when the exchange is permuted; and stronger SRM should not be attributable to static question cues. Reporting full-input scores together with perturbation deltas therefore distinguishes gains in grounding, order-sensitive reasoning, and pedagogical judgment more clearly than aggregate accuracy alone.

5 Conclusion

We introduced VIRA-Bench, a focused benchmark that decomposes short-horizon pedagogical contingency understanding into three theory-anchored layers in one-to-one online tutoring. Across 11 models, the best reaches 61.3% against a 91.0% human reference on a stratified subset, and ICR generally has the lowest layer accuracy among leading

models. Removing video most strongly degrades OEG, whereas disrupting event order most strongly degrades ICR; this ICR-selective decline persists when aligned video-ASR segments are shuffled together. These results identify distinct behavioral evidence dependencies rather than an internal architectural mechanism. Within its one-to-one tutoring scope, VIRA-Bench separates progress in observable grounding from progress in order-sensitive contingency reasoning.

6 Limitations

VIRA-Bench focuses on 15–60-second excerpts from one-to-one online English tutoring, so its findings may not transfer directly to multi-party, cross-domain, or longer-horizon interactions. Although its LLM-assisted construction is checked by multiple models and experts, occasional annotation artifacts may remain. Finally, the multiple-choice video-plus-ASR protocol targets reproducible multimodal understanding rather than open-ended generation, action, or multi-turn adaptation.

Ethical Considerations

The source videos are online tutoring sessions involving child and adolescent learners, recorded by a commercial educational institution between 2018 and 2021 with parental or guardian consent for educational and research use. The original data-collection protocol was approved by the source institution’s ethics review committee. The institution provided de-identified recordings under a

data-sharing agreement limited to non-commercial academic research and prohibiting re-identification. Before analysis or release, faces are blurred, identifying metadata are removed, and person mentions in transcripts are replaced with role placeholders; no question relies on identity, and the research team has no access to unblurred recordings or participant identities. Only processed clips and transcripts are distributed under a restricted license, and VIRABench is intended for model evaluation rather than high-stakes assessment of individual teachers or students.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62672121 and the Major Project of the National Social Science Foundation of China (NSSFC) under Grant No. 25&ZD260. Peng Zhang and Tun Lu are with the College of Computer Science and Artificial Intelligence, Fudan University. Tun Lu is also affiliated with the Silver-X MOE Philosophy & Social Sciences Laboratory, the Fudan Institute on Aging, and the Shanghai Key Laboratory of Data Science, Fudan University. We thank Jiahao Liu for valuable feedback on this work.

References

- Alibaba Cloud Model Studio. 2026a. [qwen3.5-flash model information](#). Alibaba Cloud Model Studio documentation. API model identifier: qwen3.5-flash; accessed 2026-08-30.
- Alibaba Cloud Model Studio. 2026b. [qwen3.6-flash model information](#). Alibaba Cloud Model Studio documentation. API model identifier: qwen3.6-flash; accessed 2026-08-30.
- Alibaba Cloud Model Studio. 2026c. [qwen3.6-plus model information](#). Alibaba Cloud Model Studio documentation. API model identifier: qwen3.6-plus; accessed 2026-08-30.
- Anthropic. 2025. Claude Haiku 4.5 system card. <https://www.anthropic.com/claude-haiku-4-5-system-card>.
- Kirolos Ataallah, Eslam Mohamed Bakr, Mahmoud Ahmed, Chenhui Gou, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. 2025. [InfiniBench: A benchmark for large multi-modal models in long-form movies and TV shows](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19485–19512, Suzhou, China. Association for Computational Linguistics.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. [Qwen3-VL technical report](#). Preprint, arXiv:2511.21631.
- Paul Black and Dylan Wiliam. 1998. [Assessment and classroom learning](#). *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74.
- Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. [Adversarial filters of dataset biases](#). In *International Conference on Machine Learning*.
- Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. 2024. [Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models](#). *ArXiv*, abs/2410.10818.
- Tianshui Chen, Weihang Wang, Tao Pu, Jinghui Qin, Zhijing Yang, Jie Liu, and Liang Lin. 2024a. [Dynamic correlation learning and regularization for multi-label confidence calibration](#). *IEEE Transactions on Image Processing*, 33:4811–4823.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. [Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.
- Micheline T. H. Chi and Ruth Wylie. 2014. [The icap framework: Linking cognitive engagement to active learning outcomes](#). *Educational Psychologist*, 49(4):219–243.
- Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. 2025. [Lost in time: A new temporal benchmark for videollms](#). In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24–27, 2025*. BMVA.
- Dorottya Demszky, Edith Bouton, Alison Twiner, Sara Hennessy, and Richard Correnti. 2026. [Mapping the methodological space of classroom interaction research: Scale, duration, and modality in an age of ai](#). Preprint, arXiv:2604.28098.
- Dorottya Demszky and Heather C. Hill. 2022. [The ncte transcripts: A dataset of elementary math classroom transcripts](#). *ArXiv*, abs/2211.11772.
- Dorottya Demszky and Jing Liu. 2023. [M-powering teachers: Natural language processing powered feedback improves 1:1 instruction and student outcomes](#). In *Proceedings of the Tenth ACM Conference on Learning @ Scale*.

- Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather C. Hill, Dan Jurafsky, and Tatsunori B. Hashimoto. 2021. [Measuring conversational uptake: A case study on student-teacher interactions](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Yao Dou, Michel Galley, Baolin Peng, Chris Kedzie, Weixin Cai, Alan Ritter, Chris Quirk, Wei Xu, and Jianfeng Gao. 2025. [Simulatorarena: Are user simulators reliable proxies for multi-turn evaluation of ai assistants?](#) *ArXiv*, abs/2510.05444.
- Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiaowu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2025. [Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24108–24118.
- Timnit Gebru, Jamie H. Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2021. [Datasheets for datasets](#). *Communications of the ACM*, 64(12):86–92.
- GLM-V Team, Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, Xijun Liu, Wenmeng Yu, Weihang Wang, Wei Li, Shuaiqi Duan, Sheng Yang, Ruiliang Lv, Mingdao Liu, and 78 others. 2026. [GLM-5V-Turbo: Toward a native foundation model for multimodal agents](#). *Preprint*, arXiv:2604.26752.
- Google AI for Developers. 2025a. [Gemini 2.5 Pro model information](#). Google AI for Developers documentation. Model identifier: gemini-2.5-pro; accessed 2026-08-30.
- Google AI for Developers. 2025b. [Gemini 3 Flash Preview model information](#). Google AI for Developers documentation. Model identifier: gemini-3-flash-preview; accessed 2026-08-30.
- Google AI for Developers. 2025c. [Gemini 3 Pro Preview model information](#). Google AI for Developers documentation. Model identifier: gemini-3-pro-preview; accessed 2026-08-30.
- John A. C. Hattie and Helen S. Timperley. 2007. [The power of feedback](#). *Review of Educational Research*, 77(1):81–112.
- Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, and 318 others. 2026. [Kimi k2.5: Visual agentic intelligence](#). *Preprint*, arXiv:2602.02276.
- Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. 2025. [Siv-bench: A video benchmark for social interaction understanding and reasoning](#). *ArXiv*, abs/2506.05425.
- Hengzhi Li, Megan Tjandrasuwita, Yi R. Fung, Armando Solar-Lezama, and Paul Pu Liang. 2025. [Mimeqa: Towards socially-intelligent nonverbal foundation models](#). *ArXiv*, abs/2502.16671.
- Jiangtong Li, Li Niu, and Liqing Zhang. 2022. [From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21241–21250.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. [Mvbench: A comprehensive multi-modal video understanding benchmark](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.
- Guanming Liu, Yuqi Ren, Hansu Gu, Peng Zhang, Weihang Wang, Jiahao Liu, Ning Gu, and Tun Lu. 2026a. [StreamMemBench: Streaming evaluation of agent memory for future-oriented assistance](#). *Preprint*, arXiv:2606.14571.
- Guanming Liu, Meng Wu, Peng Zhang, Yu Zhang, Yubo Shu, Xianliang Huang, Kainan Tu, Ning Gu, Liuxin Zhang, Qianying Wang, and Tun Lu. 2026b. [IntPro: A proxy agent for context-aware intent understanding via retrieval-conditioned inference](#). *Preprint*, arXiv:2603.03325.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). *Advances in neural information processing systems*, 36:34892–34916.
- Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Shihuo Chen, Xu Sun, and Lu Hou. 2024. [Tempcompass: Do video llms really understand videos?](#) In *Annual Meeting of the Association for Computational Linguistics*.
- Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. [Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems](#). *ArXiv*, abs/2305.14536.
- Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. [Mathtutorbench: A benchmark for measuring open-ended pedagogical capabilities of llm tutors](#). *ArXiv*, abs/2502.18940.
- Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. [Egoschema: A diagnostic benchmark for very long-form video language understanding](#). *ArXiv*, abs/2308.09126.

- OpenAI. 2024. [GPT-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI. 2025. GPT-5 system card. <https://cdn.openai.com/gpt-5-system-card.pdf>.
- Robert C. Pianta, Karen M. La Paro, and Bridget K. Hamre. 2008. *Classroom Assessment Scoring System (CLASS) Manual, K–3*. Paul H. Brookes Publishing Company, Baltimore, MD.
- Ao Qu, Yuxiao Wen, Jiayi Zhang, Yunge Wen, Yibo Zhao, Alok Prakash, Andrés F. Salazar-Gómez, Paul Liang, and Jinhua Zhao. 2025. [Classmind: Scaling classroom observation and instructional feedback with multimodal ai](#). *ArXiv*, abs/2509.18020.
- Sourjyadip Ray, Shubham Sharma, Somak Aditya, and Pawan Goyal. 2025. [Eduvidqa: Generating and evaluating long-form answers to student questions based on lecture videos](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Mary Rowe. 1986. [Wait time: Slowing down may be a way of speeding up!](#) *Journal of Teacher Education*, 37(1):43–50.
- Maria Araceli Ruiz-Primo and Erin Marie Furtak. 2007. [Exploring teachers’ informal formative assessment practices and students’ understanding in the context of scientific inquiry](#). *Journal of Research in Science Teaching*, 44(1):57–84.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. [WinoGrande: An adversarial winograd schema challenge at scale](#). In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.
- Ramaneswaran Selvakumar, Ashish Seth, Nishit Anand, Utkarsh Tyagi, Sonal Kumar, Sreyan Ghosh, and Dinesh Manocha. 2025. [Multivox: A benchmark for evaluating voice assistants for multimodal interactions](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Bhuiyan Sanjid Shafique, Ashmal Vayani, Muhammad Maaz, Hanoona Abdul Rasheed, Dinura Dissanayake, Mohammed Irfan Kurpath, Yahya Hmaiti, Go Inoue, Jean Lahoud, Md. Safirur Rashid, Shadid Intisar Quasem, Maheen Fatima, Franco Vidal, Mykola Maslych, Ketan More, Sanoojan Baliah, Hasindri Watawana, Yuhao Li, Fabian Farestam, and 10 others. 2025. [A culturally-diverse multilingual multimodal video benchmark & model](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2024. [Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models](#). *ArXiv*, abs/2410.23266.
- John McHardy Sinclair and Malcolm Coulthard. 1975. *Towards an Analysis of Discourse: The English Used by Teachers and Pupils*. Oxford University Press, London.
- Janneke van de Pol, Monique Volman, and Jos Beishuizen. 2010. [Scaffolding in teacher–student interaction: A decade of research](#). *Educational Psychology Review*, 22:271–296.
- Janneke van de Pol, Monique L.L. Volman, Frans J Oort, and Jos Beishuizen. 2015. [The effects of scaffolding in the classroom: support contingency and student independent working time in relation to student achievement, task effort and appreciation of support](#). *Instructional Science*, 43:615–641.
- Lev S. Vygotsky. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA.
- Rose E. Wang and Dorottya Demszky. 2023. [Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction](#). *ArXiv*, abs/2306.03090.
- Weihang Wang, Xinhao Li, Ziyue Wang, Yan Pang, Jielei Zhang, Peiyi Li, Qiang Zhang, and Longwen Gao. 2025. [Diving into mitigating hallucinations from a vision perspective for large vision-language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17271–17289, Suzhou, China. Association for Computational Linguistics.
- Weihang Wang, Kainan Tu, Jielei Zhang, Run Yang, Boheng Sheng, Yuchen He, Yu Xie, Pengyu Chen, Peiyi Li, Huyang Sun, Longwen Gao, and Zhouhui Lian. 2026. [MemeBench: What LVLMs miss when interpreting culture-dependent memes](#). *Preprint*, arXiv:2607.27798.
- Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. [Image quality assessment: From error visibility to structural similarity](#). *IEEE Transactions on Image Processing*, 13(4):600–612.
- Jacob Whitehill and Jennifer LoCasale-Crouch. 2024. [Automated evaluation of classroom instructional support with llms and bows: Connecting global predictions to specific feedback](#). *Preprint*, arXiv:2310.01132.
- David J. Wood, Jérôme Seymour Bruner, and Gail P. Ross. 1976. [The role of tutoring in problem solving](#). *Journal of Child Psychology and Psychiatry*, 17(2):89–100.
- Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2024. [Star: A benchmark for situated reasoning in real-world videos](#). *ArXiv*, abs/2405.09711.
- Jinyang Wu, Mingkuan Feng, Shuai Zhang, Feihu Che, Zhengqi Wen, Chonghua Liao, Ling Yang, Haoran Luo, Zheng Lian, and Jianhua Tao. 2026a. [Beyond examples: Towards automated thought-level in-context reasoning for large language models](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

- Long Papers*), pages 2955–2995, San Diego, California, United States. Association for Computational Linguistics.
- Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, and Jianhua Tao. 2026b. [SEED: Self-evolving on-policy distillation for agentic reinforcement learning](#). *Preprint*, arXiv:2607.14777.
- Jinyang Wu, Shuo Yang, Yuhao Shen, Shuai Zhang, Zhengqi Wen, and Jianhua Tao. 2026c. [SPARK: Strategic policy-aware exploration via dynamic branching for long-horizon agentic learning](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23981–24004, San Diego, California, United States. Association for Computational Linguistics.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat seng Chua. 2021. [Next-qa: Next phase of question-answering to explaining temporal actions](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9772–9781.
- Xiaomi MiMo Team. 2026. [MiMo-V2.5](#). Official Hugging Face model card. Model identifier: mimo-v2.5; accessed 2026-08-30.
- Run Yang, Weihang Wang, Boheng Sheng, Yuchen He, Jielei Zhang, Pengyu Chen, Zhiyu Wu, Qiang Sun, Huyang Sun, and Longwen Gao. 2026a. [MemeMind: Reference-guided trace construction for offline context optimization](#). *Preprint*, arXiv:2608.09316.
- Shuo Yang, Jinyang Wu, Zhengxi Lu, Yuhao Shen, Fan Zhang, Lang Feng, Shuai Zhang, Haoran Luo, Zheng Lian, Zhengqi Wen, and Jianhua Tao. 2026b. [OPID: On-policy skill distillation for agentic reinforcement learning](#). *Preprint*, arXiv:2606.26790.
- Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. 2025. [Audio-centric video understanding benchmark without text shortcut](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2019. [Clevrer: Collision events for video representation and reasoning](#). *ArXiv*, abs/1910.01442.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. [A survey on multimodal large language models](#). *National Science Review*, 11(12):nwae403.
- Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis philippe Morency. 2019. [Social-iq: A question answering benchmark for artificial social intelligence](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8799–8809.
- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025. [Llava-video: Video instruction tuning with synthetic data](#). *Transactions on Machine Learning Research*.

A Clip Extraction Details

We extract clips from long online tutoring sessions through three stages. First, ASR-guided coarse segmentation transcribes each session with Whisper-large-v3, identifies teacher utterances containing instructional functions (explanation, elicitation, correction, feedback, task prompting), and merges adjacent instructional intervals separated by less than 10 seconds. This reduces each 1–2 hour session to approximately 15–30 minutes of likely interaction.

Second, Visual Content Event (VCE) detection samples shared-screen frames at 2 fps within the ASR-selected intervals and compares consecutive frames using structural similarity (Wang et al., 2004) and pixel-level difference ratio. A VCE is triggered when either the SSIM score drops below 0.88 or the changed-pixel ratio exceeds 3% of the frame area. Events are classified by magnitude: *large* changes (SSIM < 0.88 and diff > 30%) indicate slide transitions or major screen rewrites, while *local* changes (SSIM < 0.88 and diff > 3%) capture teacher drawing, highlighting, or annotation. A minimum gap of 2 s is enforced between consecutive events to avoid duplicate triggers from gradual animations, and low-level cursor-only motion is removed unless accompanied by a stable pointing, drawing, highlighting, dragging, or slide-transition pattern.

Third, clip windows are generated by merging detected VCE events within a 5-second window into a single anchor group, then extending 3 s before the first event and 12 s after the last event. The window is expanded to nearby ASR turn boundaries when this helps include the student response or teacher follow-up. Clips are truncated at 60 s and discarded if shorter than 15 s.

Quality screening and face blurring. Candidate clips are screened by Gemini 2.5 Pro (Google AI for Developers, 2025a) on five dimensions: visual clarity, teaching activity, interaction completeness, audio quality, and benchmark value. A clip is rejected if any dimension fails. All faces are blurred using MediaPipe face detection with a 99×99 Gaussian kernel prior to any downstream processing; a 50-clip comparison before and after blurring confirmed accuracy differences below 3 percentage points, indicating that pedagogical

Stage	Parameter	Value
VCE detection	Sampling rate	2 fps
	SSIM threshold	0.88
	Changed-pixel ratio threshold	3%
	Minimum event gap	2 s
Clip generation	Event merge window	5 s
	Pre-onset window	3 s
	Post-onset window	12 s
	Min clip duration	15 s
	Max clip duration	60 s
Face blurring	Detection	MediaPipe
	Blur kernel	99×99 Gaussian

Table 4: Clip extraction pipeline parameters.

signals reside primarily in the shared screen and speech.

Pipeline yield. Starting from 2,760 detected VCE anchors, anchor merging and duplicate removal yield 1,840 candidate clips. Quality screening retains 1,480 clips, and structured evidence audit retains 1,370 clips. After item-level filtering, 1,005 clips with at least one valid MCQ are kept in the final benchmark.

B Prompting and Item Construction Details

VIRA-Bench uses a construction-evaluation separation protocol. GPT-4o drafts structured evidence descriptions, GPT-5 generates candidate MCQs and distractors, and GLM-5V and Claude-Haiku-4.5 serve as held-out validators. None of these construction or validation models is included in the leaderboard.

B.1 Structured Evidence Description

GPT-4o receives the tutoring clip and its ASR transcript, and is instructed to output a structured JSON record. The prompt specifies five evidence fields with typed sub-fields and timestamp requirements. The model is instructed to focus on observable evidence and to mark uncertain observations explicitly rather than infer hidden mental states.

You are an expert educational video analyst specializing in online tutoring sessions. Your task is to produce a detailed, structured description of the teaching interaction in this video clip.

The video shows an online classroom with two main visual areas: (1) webcam feeds showing teacher and student (body language, gestures, and posture are visible); (2) teacher’s shared screen showing instructional content.

Respond in JSON with the following fields:

screen_events: list of {timestamp_start, timestamp_end, action_type, action_description, content_context, pedagogical_tool}. Action

types: drawing_writing, pointing_highlighting, slide_transition, drag_drop_interactive, animation_reveal, none_ambient.

student_behaviors: list of {timestamp_start, timestamp_end, behavior_type, behavior_description, verbal_content, body_language}. Behavior types: verbal_active, verbal_passive, gestural_physical, screen_interaction, no_observable_response.

teacher_speech: list of {timestamp, content, speech_act}. Speech acts: instruction, question, feedback_positive, feedback_corrective, explanation, elicitation, social_talk.

temporal_dynamics: {intervention_response_pattern, response_latency_seconds, causal_chain_description}. Patterns: immediate (<3s), delayed (3–10s), cumulative, interleaved, no_response.

cognitive_demand: one of recall_recognize, apply_practice, analyze_create.

B.2 MCQ Generation

GPT-5 receives the audited evidence record, a target layer/subtask, and per-subtask instructions specifying the cognitive operation to test. It generates one four-option MCQ per subtask in a single pass, including the stem, four options, the gold answer, and a distractor strategy label.

The system prompt enforces the following global constraints:

Global Rules. (1) Each MCQ has exactly 4 options (A–D) with exactly one correct answer. (2) Distractors must be plausible but clearly wrong given the video evidence. (3) Anti-information-leakage: the question stem + options alone must not be sufficient to determine the correct answer without watching the video. (4) Anti-shortcut: distribute correct answers roughly evenly across A–D; include distractors that would fool a model relying only on audio, only on vision, or only on common sense. (5) Evidence-based: the correct answer must be verifiable from the structured description. (6) Questions must require understanding the video, not just text.

Option Length Parity. All four options must be approximately equal in length (within $\pm 20\%$ of the average option length in characters). The correct answer must not be systematically longer or more detailed than distractors.

Distractor Quality. Distractors must share the same grammatical structure, level of specificity, and concreteness as the correct answer. Never use absolute language (“no”, “never”, “always”) or extreme scenarios. Never use hedging language that contrasts with concrete correct answers.

Anti-Transcript-Leakage. The question stem must not quote or closely paraphrase specific speech content from the clip. The correct answer should depend on visual information that cannot be inferred from the audio transcript alone.

Each subtask receives dedicated instructions specifying the target cognitive operation, the evidence fields to focus on, and the distractor design strategy. For example, the Action Identification (AI) subtask instructs the model to focus on `screen_events` and generate distractors from plausible but non-occurring action types; the Causal Attribution (CA) subtask instructs the model to focus on `temporal_dynamics` and generate distractors that reverse causal order or confuse co-occurrence with causation.

B.3 Generate-Judge-Regenerate Loop

Each clip undergoes up to three iterations of a closed-loop quality cycle:

1. **Generate:** GPT-5 produces one MCQ per pending subtask.
2. **Judge:** Each MCQ passes through three sequential checks (described below). Failed MCQs receive structured feedback identifying the specific violation.
3. **Regenerate:** Failed MCQs are regenerated with the original prompt augmented by the judge feedback, which specifies what to fix.

MCQs that pass all three checks are accepted; those that fail after three iterations are discarded.

Phase 1: Programmatic quality checks. Before any API call, each MCQ is checked for: (a) option length ratio ($\max/\min \leq 1.2$); (b) absence of absolute language in distractors; (c) valid JSON structure with exactly four options; (d) correct answer present among options. Items failing these checks are immediately returned for regeneration with specific feedback.

Phase 2: Shortcut detection. GLM-5V and Claude-Haiku-4.5 receive only the question stem and four options, without video, ASR transcript, or evidence description. Each model selects an answer and reports confidence and any detected shortcut type (length bias, absolute language, semantic outlier, common sense, grammatical cue). If both validators answer correctly, the item is flagged as containing an option-level shortcut and returned for regeneration.

You are taking a multiple-choice test about educational video content. You have NOT seen any video, transcript, or context. You ONLY see the question stem and four options.

Select the answer you think is most likely correct. Rate your confidence: high, medium, or low. If you detect a shortcut (e.g., one option is

Stage	Removed	Remaining
Initial generation (GPT-5)	—	2,400
Format + duplicate checks	5.8%	2,260
Shortcut filtering (STEM+OPTIONS)	10.6%	2,020
Evidence-supported difficulty filtering	34.4%	1,326
Expert adjudication	0.3%	1,322

Table 5: Item filtering pipeline yield. Percentages are relative to the input of each stage.

obviously different, uses absolute language, or is much longer/shorter), describe it.

Phase 3: Evidence-supported diagnostic filtering. GLM-5V and Claude-Haiku-4.5 receive the question stem, four options, and the full audited evidence description. If both validators answer correctly and judge the item as trivial or weakly discriminative for the intended subtask, the item is returned for regeneration. This step calibrates diagnostic difficulty without reference to any specific input modality.

B.4 Item Filtering Statistics

Table 5 summarizes the multi-stage filtering pipeline. After expert adjudication of the remaining items, the final benchmark contains 1,322 MCQs.

C Expert Verification

We conduct a single-round expert audit to estimate residual item errors after full-set automatic filtering. Three domain experts independently review an audit set of 360 items (300 stratified random items covering all layers and subtasks, plus 60 high-risk items selected from inference-heavy subtasks and validator-disagreement cases). Each item is judged along four dimensions: question clarity, gold-answer correctness, evidence sufficiency, and subtask alignment, each annotated as ACCEPT, MINOR REVISION, or REJECT.

Audit instructions. The auditors received the following instructions:

Review the complete de-identified video clip and its ASR transcript, followed by the question, four answer options, gold answer, and assigned subtask. Evaluate each item independently along four dimensions: question clarity, gold-answer correctness, evidence sufficiency, and subtask alignment. Mark ACCEPT when the item is clear, has exactly one evidence-supported answer, and matches the assigned subtask. Mark MINOR REVISION when a limited wording, option-specificity, or evidence-target correction can resolve the issue without changing the intended reasoning. Mark REJECT when the question is ambiguous, the gold answer

Layer	Audited	Accept	Revise/Remove
Observable Evidence			
Grounding	135	130 (96.3%)	4 / 1
Interaction Contingency			
Reasoning	140	131 (93.6%)	7 / 2
Support-Response			
Modeling	85	79 (92.9%)	5 / 1
Total	360	340 (94.4%)	16 / 4

Table 6: Expert audit results stratified by layer. Accept = no change needed. Revise = minor wording adjustment. Remove = item excluded from final benchmark.

is unsupported, multiple options are valid, the required evidence is absent from the clip, or the item does not match its assigned subtask. Base all judgments only on evidence available in the video and transcript.

Recruitment and compensation. The three auditors were recruited based on their graduate-level training in applied linguistics or educational technology and their prior experience with classroom-interaction analysis. They were compensated at the institution’s standard research-assistant rate for comparable expert annotation work.

Among the 360 audited items, 340 are accepted without change, 16 are retained after minor revision, and 4 are removed, yielding a 94.4% direct acceptance rate and 98.9% final acceptance rate. For the stratified random split, 288/300 are accepted without change (2 removed); for the high-risk split, 52/60 are accepted without change (2 removed). Among the 16 revised items, 7 involved tightening question wording, 5 involved rebalancing option specificity, and 4 involved clarifying the evidence target. Among the 4 removed items, 2 had insufficient visual evidence, 1 had an ambiguous gold answer, and 1 had a subtask mismatch.

Agreement and stratified results. Fleiss’ κ on the binary acceptability decision is 0.76 (pairwise agreement 96.7%). Per-dimension κ : question clarity 0.82, gold-answer correctness 0.79, evidence sufficiency 0.68, subtask alignment 0.74. Evidence sufficiency shows the lowest agreement, particularly for Support-Response Modeling items where experts differ on whether short clips provide sufficient context. Table 6 reports audit outcomes by layer; acceptance rates are uniformly high, with Interaction Contingency Reasoning showing slightly more revisions due to the difficulty of judging temporal evidence sufficiency.

Sampling adequacy. The 360-item audit set represents a 27% sampling rate of the final benchmark. At the observed 1.1% removal rate (4/360), the 95%

Layer	Subtask	MCQs	%
Observable Evidence Grounding	AI (Action Identification)	215	16.3
	BE (Behavior Extraction)	188	14.2
	CR (Content Recognition)	94	7.1
Interaction Contingency Reasoning	SR (Sequence Recognition)	250	18.9
	CA (Causal Attribution)	82	6.2
	CF (Contingent Feedback)	197	14.9
Support-Response Modeling	SC (Support Calibration)	136	10.3
	OP (Outcome Prediction)	160	12.1
Total		1,322	100.0

Table 7: MCQ distribution across subtasks.

Wilson confidence interval is [0.4%, 2.8%], bounding the expected number of residual defective items in the full 1,322-item set to fewer than 37. This human audit is the final layer in a multi-stage quality pipeline: automated shortcut filtering removes 10.6% of candidates, difficulty calibration removes trivially easy items, and model-based evidence validation removes 34.4% of candidates before expert review.

D Benchmark Statistics

Clip and item statistics. Across 1,005 clips, the mean duration is 32.4 s (SD = 11.4 s; 25th/50th/75th percentiles: 23/30/40 s; 94% within the 15–60 s design window). Each clip has between 1 and 3 MCQs (mean 1.32): 72% have exactly 1, 24% have 2, and 4% have 3. Multiple MCQs on the same clip always target different subtasks.

Cross-lingual characteristics. All source sessions are online English tutoring, where the teacher instructs in English while adapting explanations to students with diverse first-language backgrounds (primarily Mandarin Chinese, with some English L1 students). Teacher speech naturally includes code-switching between English instruction and native-language clarification. The shared screen content is predominantly in English (instructional materials, exercises, vocabulary).

The source corpus contains over 1,000 recorded sessions (each 1–2 hours); the 1,005 clips are drawn from sessions selected for instructional richness and interaction density. Students include both Chinese L1 and English L1 learners at primary and secondary levels. Teacher speech is approximately 75% English and 25% other languages by word count (estimated from ASR output). ASR is performed with Whisper-large-v3, which handles code-switched multilingual input.

Because all sessions share the same instructional format (English-medium online tutoring), the pri-

mary variation is in student L1 background and teacher code-switching patterns rather than in the language of instruction itself. We acknowledge the limited diversity of instructional contexts as a scope limitation in §6.

E Evaluation Protocol

All models use provider-default generation parameters (temperature and top- p are not overridden). Video clips are uploaded at native resolution (720p, 15–60 s duration) and sampled at 1 frame per second across all providers. The ASR transcript is provided as a separate text field in the prompt. All models are queried via their respective provider APIs in a zero-shot setting with no system prompt.

Inference prompt templates. Under the default Video+ASR condition, models receive the video clip together with the ASR transcript provided as a separate text input. The transcript is not burned into the video frames; models read it from the prompt text. The prompt is:

```
Watch the video clip and read the speech transcript, then answer the question.
Use the transcript as spoken-language evidence;
use the video for visual context.

## Speech Transcript
{transcript}

## Question
{stem}

## Options
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}

Return ONLY valid JSON: {"answer":
"<A/B/C/D>", "confidence": "high|medium|low",
"reasoning": "brief explanation"}
```

Under the ASR TEXT ONLY condition, the video is removed entirely and only the speech transcript is provided:

```
Read the speech transcript from a tutoring clip,
then answer the question. You do NOT have access to the video.

## Speech Transcript
{transcript}

## Question
{stem}

## Options
A. {option_A} ... D. {option_D}

Return ONLY valid JSON: {"answer":
"<A/B/C/D>", "confidence": ... }
```

Under the VISUAL SHUFFLE WITH FIXED ASR condition, the video segments are permuted at

VCE-detected scene boundaries before upload while the original transcript remains fixed. The prompt is identical to the Video+ASR condition; models are not informed that the temporal order has been altered.

Under the SYNCHRONIZED VIDEO+ASR SHUFFLE condition, each visual segment is paired with its timestamp-aligned ASR span and the pairs are permuted together. Order within each pair is preserved. This retains local video-text alignment while disrupting the original intervention, response, and follow-up sequence.

Under the STATIC FRAME condition, 5 evenly-spaced key frames are extracted from the clip and presented as images:

```
Look at these 5 key frames extracted from a video
clip of an online tutoring session, then answer the
question.

## Question
{stem}

## Options
A. {option_A} ... D. {option_D}

Return ONLY valid JSON: {"answer":
"<A/B/C/D>", "confidence": ... }
```

Response parsing and scoring. Models are instructed to return a JSON object containing an answer letter (A–D), a confidence level, and a brief reasoning string. We first attempt JSON extraction via regex; if parsing fails, we fall back to extracting a bare option letter or matching the normalized response to the option text. Responses that match neither pattern are scored as incorrect. After fall-back parsing, all models have zero invalid outputs under the full condition. Option order is randomized independently for each model–item pair using a fixed random seed, and the gold label is remapped accordingly. For API-based models, responses indicating the model did not receive the video trigger up to 3 automatic retries before being marked as failures. Accuracy is computed as the proportion of items matching the gold answer (no partial credit; random baseline 25%). We report accuracy at three levels: overall (1,322 items), layer-level (Observable Evidence Grounding, Interaction Contingency Reasoning, Support-Response Modeling), and subtask-level (eight subtasks), computed directly over items in each group.

Answer-label diagnostics. The remapped gold labels are balanced with 330 A labels, 331 B labels, 331 C labels, and 330 D labels. Table 8 reports model output labels under the full condition. The

distributions show no output collapse and zero invalid responses. On the 924 errors shared by the two 32B Qwen3-VL variants, the models choose the same distractor content in 47.6% of cases after independent option shuffling. A permutation test gives $p < 0.001$, which is consistent with related models favoring the same plausible distractors rather than a fixed answer position.

Statistical testing. For the evidence dependency analysis, we apply McNemar’s test for paired binary outcomes, comparing item-level accuracy under the Video+ASR condition against each perturbation condition at the layer level. We apply Holm correction for multiple comparisons and report adjusted p -values with significance thresholds at $\alpha = 0.05, 0.01$, and 0.001 .

F Human Reference Details

The 200-question human evaluation set is stratified by subtask (25 questions $\times$ 8 subtasks), sampled uniformly at random from the full 1,322-question set. Three domain experts with graduate training in applied linguistics independently answered all 200 questions after watching the corresponding video clips with ASR transcripts. Table 9 reports individual and majority-vote accuracy.

Using the Wilson score interval at 95% confidence: Observable Evidence Grounding: 92.0% [83.6%, 96.3%]; Interaction Contingency Reasoning: 88.0% [78.7%, 93.6%]; Support-Response Modeling: 94.0% [83.8%, 97.9%]; Overall: 91.0% [86.2%, 94.2%]. The lower bound of the overall CI (86.2%) remains substantially above the best model (61.3%), confirming the human–model gap is robust. Individual expert totals are 175, 170, and 177 out of 200 (mean 87.0%); majority voting raises accuracy to 91.0%, indicating that disagreements are resolved by at least two experts agreeing. Pairwise agreement on the binary correct/incorrect decision exceeds 90% across all expert pairs, with agreement highest on Observable Evidence Grounding items and lowest on Interaction Contingency Reasoning Causal Attribution items. This sample size is comparable to human evaluation protocols in recent video understanding benchmarks (Li et al., 2024; Fu et al., 2025).

G Paired Significance Analysis of Evidence Dependency

To determine whether the observed performance changes are statistically reliable, we conduct two

sided McNemar tests using paired item level correctness outcomes. For each model and diagnostic layer, we compare predictions under the **Full** setting against predictions under **ASR Text Only** and **Visual Shuffle with Fixed ASR** on the same set of questions. ASR Text Only removes visual evidence while retaining the standardized ASR input. Visual Shuffle with Fixed ASR retains the ASR input and visual content, but disrupts the temporal organization of the visual stream. We apply Holm correction across all comparisons reported in Table 10.

The significance analysis supports a selective evidence dependency pattern across diagnostic layers. Removing visual input produces the largest degradation on OEG for every evaluated model, with drops ranging from 12.7 to 21.9 percentage points. All OEG declines under ASR Text Only remain significant after correction. This result indicates that observable teacher actions, learner behaviors, and shared screen content cannot be reliably recovered from ASR input alone.

Visual Shuffle with Fixed ASR produces a complementary effect. Its largest degradation consistently occurs on ICR, with drops ranging from 11.9 to 19.1 percentage points, all significant after correction. Because the ASR input remains unchanged in this condition, the ICR decline reflects the loss of visual temporal organization required to connect teacher intervention, learner response, and teacher follow up. In contrast, Visual Shuffle with Fixed ASR has no significant adverse effect on OEG for four of the five evaluated models, suggesting that the perturbation largely preserves access to local observable evidence while disrupting relation bearing temporal structure.

SRM exhibits a less uniform response to the two perturbations. Most SRM changes are not significant after correction, with the exception of larger model specific effects such as the degradation under ASR Text Only for Gemini 3 Flash and under Visual Shuffle with Fixed ASR for Kimi-K2.5. This mixed pattern is consistent with the design of SRM, which requires predictive or evaluative judgments supported by multiple forms of evidence, including visual context, language cues, learner state, and broader instructional priors.

Effect of temporal shuffling. Visual Shuffle with Fixed ASR is constructed by permuting temporally coherent event segments while preserving the internal order of frames within each segment. It

Model	Accuracy	A	B	C	D	Invalid
Gemini 3 Pro	61.3	293	345	342	342	0
Gemini 3 Flash	59.7	337	320	317	348	0
Kimi-K2.5	57.4	291	348	344	339	0
MiMo-V2.5	56.9	291	353	336	342	0
Qwen3.6-Plus	55.9	328	332	330	332	0
Qwen3-VL-235B-Think	43.0	307	334	343	338	0
Qwen3.6-Flash	40.9	295	340	338	349	0
Qwen3-VL-235B-Inst	38.0	302	352	331	337	0
Qwen3.5-Flash	24.7	291	340	340	351	0
Qwen3-VL-32B-Think	22.1	326	324	338	334	0
Qwen3-VL-32B-Inst	20.3	332	334	319	337	0

Table 8: Model output-label distributions under the full condition. Accuracy is reported in percent and label columns report counts.

	Exp. 1	Exp. 2	Exp. 3	Majority
Observable Evidence				
Grounding (75 Qs)	90.7	89.3	92.0	92.0
Interaction Contingency				
Reasoning (75 Qs)	84.0	80.0	84.0	88.0
Support-Response				
Modeling (50 Qs)	88.0	86.0	90.0	94.0
Overall (200 Qs)	87.5	85.0	88.5	91.0

Table 9: Human reference accuracy (%) by layer and expert. Majority = majority vote across three experts.

therefore disrupts the ordering among intervention, response, and follow up without reducing each segment to visually incoherent frames. The resulting performance pattern further indicates that the primary effect is not generic visual corruption. OEG remains largely stable for most models, whereas ICR declines substantially and consistently. At the subtask level, the strongest degradation is expected on Sequence Recognition, followed by relation focused tasks such as Causal Attribution Plausibility and Contingent Feedback Recognition. Together, these results show that VIRA-Bench requires models to use visual temporal evidence to recover contingent pedagogical relations rather than merely recognize isolated events or rely on language cues.

H Detailed Control Results

I Positioning Against Causal Video QA

VIRA-Bench differs from existing causal video QA benchmarks in three respects. First, causal video QA benchmarks (Xiao et al., 2021; Wu et al., 2024; Yi et al., 2019; Li et al., 2022) primarily test physical causality or event-level “why” questions in third-person observation settings. VIRA-Bench tests *interactional* causality: whether one participant’s action triggers another’s response and shapes subsequent behavior. Second, VIRA-Bench grounds its diagnostic layers in established peda-

gogical theory (IRF structure, contingent teaching), providing a principled decomposition rather than ad-hoc question categories. Third, VIRA-Bench includes controlled evidence perturbation (ASR-only, temporal shuffle) as part of the evaluation protocol, enabling construct validity analysis beyond aggregate accuracy.

J Transcript De-identification

Spoken names can remain in ASR transcripts after visual and metadata de-identification. We therefore process all 1,005 released transcripts with multilingual named-entity recognition and rules for vocatives and honorific-plus-name constructions in English and Mandarin. The process detects 1,340 person mentions. Each mention is replaced with [TEACHER], [STUDENT], or [SPEAKER] when the role cannot be determined. Replacement rather than deletion preserves who addresses whom, which can be relevant to Sequence Recognition and Causal Attribution. The released transcripts are de-identified versions of the raw Whisper-large-v3 output used during construction. No subtask asks for identity and no name is used as answer-bearing evidence.

K Datasheet

We follow the Datasheet for Datasets framework (Geburu et al., 2021).

Motivation. VIRA-Bench was created to evaluate whether video LLMs can understand pedagogical interaction contingency in educational videos. Existing video QA benchmarks test physical causality or social perception but do not assess whether models can track the causal chain from teacher intervention through student response to instructional follow-up. The benchmark was created by

Model	Condition	Observable Evidence Grounding		Interaction Contingency Reasoning		Support-Response Modeling	
		Δ	p_{adj}	Δ	p_{adj}	Δ	p_{adj}
Gemini 3 Flash	ASR Text Only	-21.9	<.001***	-7.0	<.001***	-8.1	<.01**
	Visual Shuffle	-9.1	<.001***	-19.1	<.001***	-4.4	n.s.
Kimi-K2.5	ASR Text Only	-17.9	<.001***	-3.0	n.s.	-2.4	n.s.
	Visual Shuffle	+1.0	n.s.	-13.2	<.001***	-9.8	<.001***
MiMo-V2.5	ASR Text Only	-14.5	<.001***	-8.9	<.001***	-2.7	n.s.
	Visual Shuffle	-3.4	n.s.	-17.8	<.001***	-1.0	n.s.
Qwen3.6-Plus	ASR Text Only	-16.9	<.001***	-4.7	<.05*	-2.4	n.s.
	Visual Shuffle	-0.8	n.s.	-17.4	<.001***	-3.0	n.s.
Qwen3-VL-235B-Thinking	ASR Text Only	-12.7	<.001***	-5.7	<.01**	-1.0	n.s.
	Visual Shuffle	-2.0	n.s.	-11.9	<.001***	-1.0	n.s.

Table 10: Paired significance analysis of evidence dependency. Δ denotes the change in accuracy, measured in percentage points, relative to the Full setting. Negative values indicate performance degradation. Adjusted p values are obtained from two sided McNemar tests over paired item level correctness outcomes with Holm correction for multiple comparisons. *n.s.* indicates that the change is not significant after correction. * $p_{adj} < .05$, ** $p_{adj} < .01$, *** $p_{adj} < .001$.

Model	Setting	OEG	ICR	SRM	All
Gemini 3 Flash	Full MCQ	62.0	55.0	64.2	59.7
	Free response	50.5	41.8	53.0	47.6
	STEM+OPTIONS	29.8	29.5	29.7	29.7
Qwen3.6-Plus	Full MCQ	55.1	54.3	60.1	55.9
	Free response	43.9	41.6	49.3	44.2
	STEM+OPTIONS	29.0	29.3	29.4	29.2

Table 11: Response-format and no-evidence controls. Accuracy is reported in percent.

	Causal reasoning	Participant interaction	Theory-grounded	Evidence control
NExT-QA (Xiao et al., 2021)	✓	✗	✗	✗
STAR (Wu et al., 2024)	✓	✗	✗	✗
CLEVRER (Yi et al., 2019)	✓	✗	✗	✓
Causal-VidQA (Li et al., 2022)	✓	✗	✗	✗
Social-IQ 2.0 (Zadeh et al., 2019)	✓	✓	✗	✗
VIRA-Bench	✓	✓	✓	✓

Table 12: Comparison with causal video QA benchmarks. VIRA-Bench uniquely combines causal reasoning over participant interaction with theory-grounded diagnostic layers and controlled evidence perturbation.

the authors at their respective institutions.

Composition. The benchmark contains 1,322 four-option multiple-choice questions spanning 1,005 video clips from online tutoring sessions (duration statistics in Appendix D). Questions are distributed across three diagnostic layers (Observable Evidence Grounding: 497, Interaction Contingency Reasoning: 529, Support-Response Modeling: 296) and eight subtasks. All clips depict real instructional interactions; no synthetic or staged content is included. The dataset does not contain

offensive content. All faces are blurred.

Collection process. Source videos are online tutoring sessions recorded by a commercial educational institution between 2018 and 2021. Sessions involve English-medium instruction with Chinese L1 students at primary and secondary levels. Informed consent for educational and research use was obtained from all participants’ parents or legal guardians. The authors selected sessions from a corpus of over 1,000 recorded sessions based on instructional richness and interaction density.

Preprocessing. All faces are blurred prior to any downstream processing. Clips are extracted via the VCE-based pipeline described in Appendix A. ASR transcription uses Whisper-large-v3. Released transcripts are de-identified as described in Appendix J.

Uses. VIRA-Bench is intended for evaluating multimodal language models on pedagogical interaction understanding. It should not be used for teacher evaluation, student assessment, or any high-stakes educational decision-making. The benchmark is not designed for training; using it as training data would compromise its diagnostic validity.

Distribution. The benchmark will be released under a research-only license. Video clips are distributed as de-identified MP4 files. MCQ metadata (stems, options, gold answers, layer/subtask labels) are distributed as JSON. Raw tutoring sessions are not released. Access requires agreeing to terms pro-

hibiting re-identification attempts and commercial use.

Maintenance. The benchmark is maintained by the authors. Version updates will be documented in the release repository. We do not plan scheduled updates but will issue corrections if errors are identified post-release.