DYNAMICDX: EVALUATING EVIDENCE ACQUISITION IN VIDEO-BASED DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diagnosing a patient from video requires more than recognizing the sign: a vision–language model must turn what it sees into hypotheses, questions and tests. DYNAMICDX evaluates each step in 71 neurological consultations across 11 sign categories, linking authentic patient videos to confirmed diagnoses and fixed charts built from the same case reports, so that every model queries the same evidence. Across five such models, video improves accuracy by 9.9–22.5 percentage points over blind input, but neither recognition alone nor temporal order explains the gain: the cause is usually missing from the model's video-only differential diagnosis even when the sign is recognized, and shuffling the frames produces no reliable accuracy loss. Instead, a trajectory replay traces most of the gain to the investigation results the video prompts. Evidence acquisition is the bottleneck: supplying the decisive investigations raises accuracy to 73.2–93.0%. Two interventions act on it. A post-trained 4B video describer improves sign descriptions, especially from a short, densely sampled segment, and source-clean literature retrieval expands initial hypotheses; both bring the tests a model orders closer to those the treating clinicians documented and, through them, raise accuracy. For video-based diagnosis, seeing better helps when it leads to asking better.

1 INTRODUCTION

Clinical diagnosis links presenting signs to hypotheses, evidence acquisition, and a final decision. In neurology, a sign's phenomenology, what the movement looks like, narrows its plausible causes (aetiologies) and the investigations to order; irregular flowing movements (chorea) and sustained postures (dystonia), for example, point to different disorders (Abdo et al., 2010; Bhatia et al., 2018; Albanese et al., 2013). Recognizing a sign is only the beginning: a model must list candidate causes (the differential diagnosis), question the patient (history-taking), order tests (the work-up), and integrate the returned evidence. A correct description can therefore coexist with an incomplete differential or an incorrect diagnosis. Evaluating only the final answer hides these steps.

Existing benchmarks capture parts of this process. Text-based systems assess differential diagnosis and patient interaction (McDuff et al., 2025; Tu et al., 2025), while multimodal evaluations use static images, procedural videos, or actor-based encounters (Gupta et al., 2023; Schmidgall et al., 2026; Shah et al., 2026). Two questions remain open. Do models use how a patient's sign unfolds over time, not only how it looks? And how does what they see shape the evidence they acquire?

We introduce DYNAMICDX, a process-level diagnostic environment driven by real patient videos: 71 clips across 11 neurological phenomenological categories, each paired with the confirmed diagnosis and a fixed chart from the same source report. Consultations proceed in two stages, so that a missed diagnosis can be traced to what the model saw or to the evidence it sought: Stage 1 measures reactive perception (recognizing the sign and proposing a differential), and Stage 2 proactive evidence acquisition (taking a history and ordering investigations) before the diagnosis. Every model queries the same fixed chart, so differences in outcome reflect the model rather than a simulated patient. Charts are semi-synthetic: expected values fill in the test results a report left out, so the chart is complete and which tests return results does not reveal the diagnosis. Seven controlled conditions vary visual input, textual representation, and supplied evidence; single-frame and shuffled inputs separate additional views from their temporal order.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

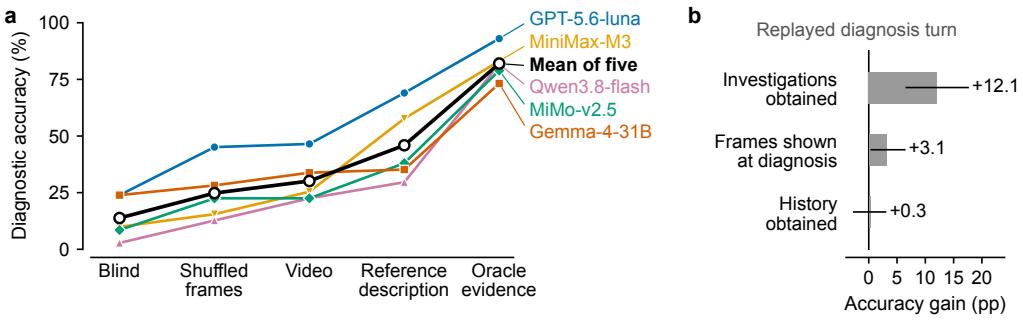


Figure 1: Video helps diagnosis through the evidence it prompts, not through its dynamics. (a) Diagnostic accuracy on the 71 cases as the input grows (Table 2); black, mean over the five models. (b) Accuracy gained at a replayed diagnosis turn from each component, mean over the five models with 95% bootstrap intervals (Table 11).

Across five vision–language models, video improves accuracy by 9.9–22.5 percentage points over blind input (Figure 1). On the first question, current models use what a clip shows rather than how it unfolds: shuffling the frames produces no reliable accuracy loss, a gap that single-answer video benchmarks (Fu et al., 2025) do not expose. On the second, what they see helps through the evidence it prompts rather than through recognition: recognizing the sign rarely puts its cause in the differential, and replaying consultations with swapped trajectories traces most of the gain at the diagnosis step to the investigation results obtained, not to the history. Evidence acquisition is the bottleneck: clinician-written descriptions recover more of the documented work-up and raise accuracy, and supplying the decisive investigations raises it to 73.2–93.0%. Two interventions act on it: a post-trained video describer improves sign descriptions, especially from a short, densely sampled segment, and source-clean retrieval expands hypotheses; both raise accuracy by moving the work-up toward the documented one. Our contributions are:

- **A benchmark for evidence acquisition from patient video.** DynamicDx builds interactive consultations on authentic patient video, pairing each clip with the confirmed diagnosis and a fixed chart from the same case report, so that models are compared on identical evidence and a missed diagnosis can be traced to the stage where it failed.
- **Evidence acquisition as the bottleneck.** We show that video helps diagnosis through the investigations it prompts rather than through recognition or temporal order, and that accuracy is limited by acquiring the decisive investigations rather than by interpreting them.
- **Two ways to act on it.** A post-trained video describer and source-clean retrieval each raise accuracy by moving the work-up toward the documented one, and controls separate sampling density from window choice and list breadth from retrieved content.

2 Related Work

Sequential clinical reasoning and conversational diagnosis. Clinical reasoning links presenting findings to diagnostic hypotheses and evidence acquisition (Elstein et al., 1978; Elstein & Schwartz, 2002; Croskerry, 2009). Medical VQA and question answering evaluate answers from fixed inputs (Lau et al., 2018; Singhal et al., 2023; 2025; Moor et al., 2023; Saab et al., 2024; Jeong et al., 2024), whereas interactive systems assess question asking, diagnostic dialogue, multi-agent deliberation, and tool use (Li et al., 2024; Johri et al., 2024; 2025; Tu et al., 2025; Kim et al., 2024; Schmidgall et al., 2026; Liévin et al., 2026); selective evidence acquisition also improves clinical signal classification (Li et al., 2026a). Recent work also evaluates differential diagnosis from case descriptions (McDuff et al., 2025), multimodal artifacts (Saab et al., 2026), and live video consultations with patient actors (Shah et al., 2026). The interactive multimodal evaluations in Table 1 span 20–120 cases or scenarios, none drawn from patient video; DynamicDx instead links 71 authentic patient clips to fixed source-case history and test results and scores each stage from recognition to diagnosis.

Table 1: Benchmark comparison (✓: included; Diff.: differential; Hx: history; Dx: diagnosis). Cases: evaluation cases or scenarios (MedVidQA: videos; AgentClinic: multimodal + text).

Benchmark	Source	Cases	Video	Sign	Diff.	Hx	Tests	Dx
MedVidQA	Instructional video	899	✓	–	–	–	–	–
MediQ	Simulated cases	1,413	–	–	–	✓	–	✓
AgentClinic	Simulated cases/records	120 + 415	–	–	–	✓	✓	✓
Multimodal AMIE	Actors + artifacts	105	–	–	–	✓	–	✓
AI co-clinician	Actors	20	✓	–	–	✓	–	✓
DynamicDx (ours)	**Patient clips + records**	**71**	✓	✓	✓	✓	✓	✓

Dynamic medical video understanding and temporal sampling. MedVidQA evaluates medical-video classification and temporal answer localization (Gupta & Demner-Fushman, 2022); video analysis supports tremor, Parkinson's disease, dystonia and seizure assessment (Friedrich et al., 2024; Di Biase et al., 2025; Haberfehlner et al., 2023; Ahmedt-Aristizabal et al., 2024) and remote movement-disorder examination (Srinivasan et al., 2020; Newton et al., 2022; Angelopoulou et al., 2024). In general-domain video, many questions can be answered from a single frame and temporal understanding remains weak (Buch et al., 2022; Liu et al., 2024), echoing the reliance on priors in image VQA (Goyal et al., 2017; Agrawal et al., 2018). General video sampling methods use sequential selection (Wu et al., 2019), relevance–coverage optimization (Tang et al., 2025), adaptive selectors (Buch et al., 2025), or generative relevance supervision (Yao et al., 2025) to address redundancy and context limits. Clinical clips present a complementary challenge: brief or high-frequency signs may be poorly captured by uniform sampling, and instruction tuning can adapt language models to physiological signals (Li et al., 2026b). We evaluate temporal-window selection through its effects on downstream evidence acquisition and diagnosis, beyond video recognition alone.

Retrieval-augmented clinical reasoning. Clinicians routinely consult the literature (Westbrook et al., 2004; Weller et al., 2023; 2026). MedRAG benchmarks retrieval configurations for medical QA (Xiong et al., 2024), i-MedRAG supports iterative information seeking (Xiong et al., 2025), and RULE examines retrieved-context reliability in multimodal QA (Xia et al., 2024). In DynamicDx, retrieval supplies candidate aetiologies from predicted phenomenology before history-taking.

3 Method

3.1 Benchmark Construction and Case Representation

DynamicDx contains 71 consultations from 66 open-access case reports with patient videos, spanning 11 neurological phenomenological categories: chorea (10), paroxysmal events (9), central vertigo (9), functional movement disorder (7), ataxia (6), dystonia (6), parkinsonism (6), ptosis (6), tremor (6), myoclonus (5), and facial palsy (1). Reports retrieved from Europe PMC (Europe PMC Consortium, 2015) were retained when the video visibly demonstrated the presenting sign and the report documented a confirmed diagnosis and workup. Videos were trimmed to at most 30 seconds around the sign; findings not established visually were represented as history or investigation evidence. Sources and licences are listed in Appendix A.

Each case is $c = (V, H, E, D, T)$: video V, documented positive and negative history findings H, investigation results E, diagnosis D, and decisive entries $T \subseteq E$ that support the diagnosis or exclude an alternative. Free-form questions and orders map to H and E. A case report lists only the tests its clinicians ran, so a chart built from it alone would leave reasonable tests unanswered, even when other reports establish their results for the condition. It would also give the answer away: any test that returned a result would point toward the diagnosis. E is therefore semi-synthetic. All cases of a category share one menu of tests; a test the report ran returns its reported result, and every other test returns a derived value expected for the patient's condition, usually normal or not performed. Derived values were set by annotators or by DeepSeek-V4.1-Flash given the confirmed diagnosis, and fixed before evaluation (Appendix C.2). A second, blinded annotator independently labeled decisive entries (Cohen's $\kappa = 0.75$), with disagreements resolved by consensus.

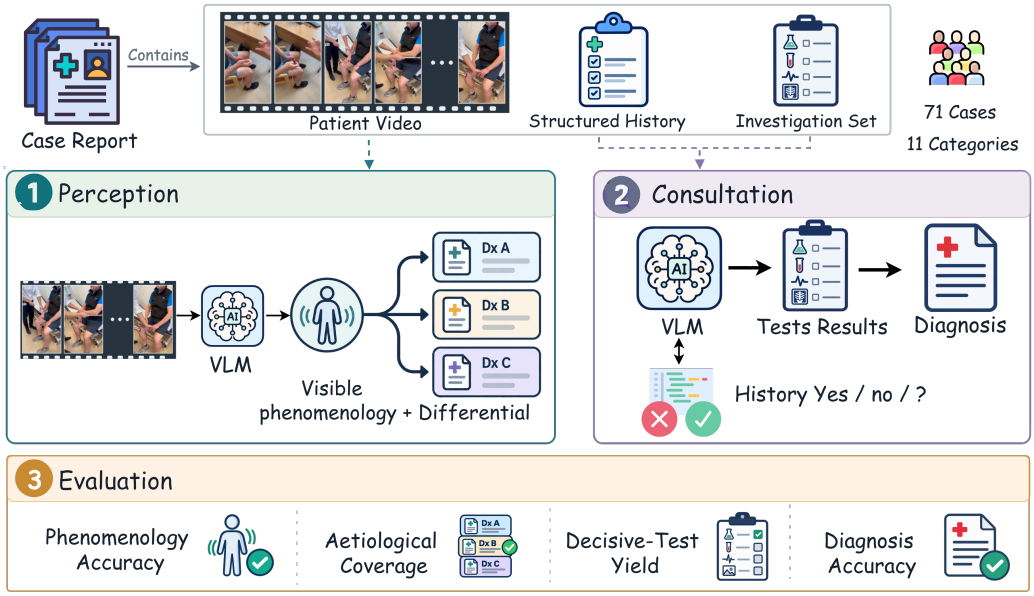


Figure 2: Overview of DYNAMICDX. Stage 1 evaluates reactive perception (sign recognition and hypotheses from video); Stage 2 evaluates proactive evidence acquisition (history and investigations) and diagnosis from the case chart.

3.2 SEQUENTIAL DIAGNOSTIC ENVIRONMENT

Consultations proceed in two stages (Figure 2), so that a missed diagnosis can be traced to what the model saw or to the evidence it sought. Stage 1 measures reactive perception: from K temporally ordered frames alone, the model describes the phenomenology and gives an open-ended differential. Stage 2 measures proactive evidence acquisition: the model asks all its history questions in one turn and orders all investigations in a second, then gives its diagnosis. By default, Stage 2 starts from the video alone; own-words and retrieval conditions supply Stage 1 output instead.

Fixed prompts on DeepSeek-V4.1-Flash (DeepSeek-AI, 2026) map free-text questions and orders to the chart, so that every model is answered from the same fixed record (Appendix H). A history question is answered `yes` or `no` only when the record documents presence or absence, and `unknown` otherwise. An order releases the fixed results of the entries it matches, and unmatched orders return `not performed / not available`. Orders may name therapeutic trials, such as a levodopa challenge. The model gives one primary diagnosis and up to three alternatives; only the primary diagnosis is scored (Section 3.3).

3.3 PROCESS-LEVEL MEASUREMENTS

We evaluate four outcomes. **Phenomenology recognition** credits the reference visible sign, written independently by two annotators from reports and videos ($\kappa = 0.83$), or equivalent terminology, distinguishing partial from incorrect descriptions. **Aetiological coverage** considers the full video-only differential, including entries beyond the first three; compatible syndromes or broader aetiological categories qualify without requiring the confirmed disease. **Decisive test yield** asks whether a consultation obtains the results that settled the case. It is measured by source-workup coverage, $\tau = |T \cap E_{\text{acq}}|/|T|$, the share of a case's decisive entries T found among the chart entries E_{acq} a consultation obtains, pooled over the 71 cases; an entry reporting the same finding from the same kind of test also counts (Appendix C). τ measures how much of the documented work-up is recovered. **Final diagnosis** uses case-specific acceptance lists: ACCURATE includes accepted disease or syndrome formulations; PARTIAL meets partial-credit criteria or identifies only the broad disease family; NOT ACCURATE meets neither criterion. Clinical experts independently built the acceptance lists, blinded to all model outputs. Diagnostic accuracy is the proportion graded ACCURATE.

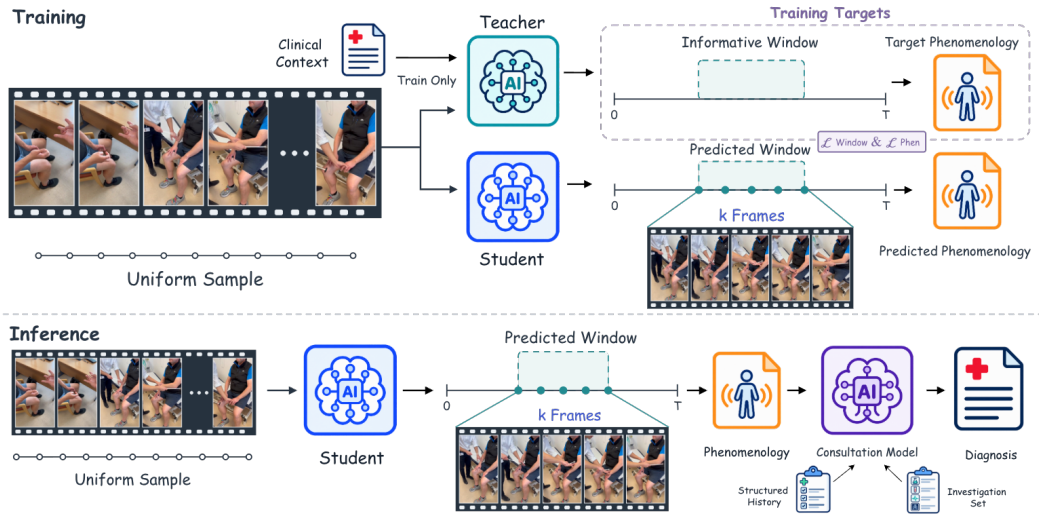


Figure 3: Teacher-guided temporal-window sampling. Clinical context supervises the offline teacher only; benchmark phenomenology supplies the recognition target. The student predicts at most one window from 32 survey frames and describes K window frames.

All outcomes are graded automatically by separate DeepSeek-V4.1-Flash prompts, given with the grading criteria in Appendix H. Independent human review agrees with the automated grades on 89.6% of phenomenology and 96.9% of aetiology assessments (96 Stage 1 responses) and on 91.5% of diagnostic grades for all 71 GPT-5.6-luna video consultations ($\kappa_w = 0.947$). Of 100 random orders across five models, 95 match correctly, with three over-releases and two misses (Appendix J).

3.4 Controlled Diagnostic Conditions

Seven conditions locate where the benefit of video arises; each changes one input and keeps the rest of the consultation fixed. **Video** provides $K = 32$ temporally ordered frames throughout the consultation, and **Blind** removes them, stating that the model has not yet seen or examined the patient; their difference is the total benefit of visual access. **Single frame** keeps one frame, testing whether multiple views matter, and **Shuffled** presents the same frames in randomly permuted order (three permutations per case), testing whether temporal order matters. **Own words** replaces the video with the model's own phenomenology description, testing whether the clip carries anything the model cannot put into words, and **Reference** substitutes a clinician-written description restricted to visible findings, testing whether better perception would help. **Oracle evidence** keeps the Video condition's frames and history but replaces investigation selection with the adjudicated decisive investigations and their results, so $E_{\text{acq}} = T$ and $\tau = 1$; it measures how well models diagnose once evidence acquisition is solved.

3.5 Targeted Interventions

Diagnosis can fail before any evidence is sought, if the sign is misdescribed or a correct description does not suggest its cause. Temporal-window training (Section 3.5.1) targets the first step, describing the phenomenology, and literature retrieval (Section 3.5.2) the second, turning it into a differential.

3.5.1 Teacher-Guided Temporal-Window Training

A dynamic sign often occupies a short segment of a clip, which uniform sampling at small K can miss, so the describer first chooses where to look. An offline teacher (Figure 3) uses video, reference phenomenology, and diagnosis in one call per clip to label at most one window by start time and duration, or choose no window. We train Qwen3.5-4B (Qwen Team, 2026a) with LoRA (Hu et al., 2022) to predict this window from 32 survey frames and describe $K \in \{8, 16, 32\}$ window frames, with reference phenomenology as the recognition target. At evaluation, Window

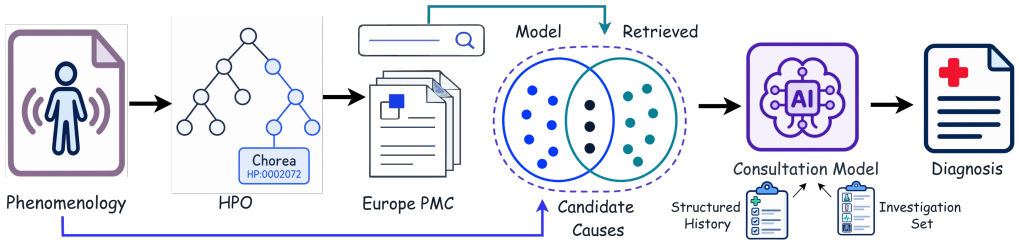


Figure 4: Retrieval pipeline from predicted phenomenology to candidate causes.

uses the predicted position; Random holds the adapter, survey, duration, and K fixed but randomizes position; responses without a valid window fall back to the survey. A sign-only adapter, trained and tested on $32{+}K$ frames spread over the clip, uses no window. Five-fold cross-validation groups cases by source article. Out-of-fold descriptions initialize GPT-5.6-luna consultations. Teacher outputs and clinical fields only build training targets; at test time the student sees frames alone. Appendix I gives settings, including the sign-only adapter's one-epoch schedule.

3.5.2 Literature Retrieval and Decontamination

DeepSeek-V4.1-Flash normalizes the predicted phenomenology (Figure 4) to Human Phenotype Ontology terms (Gargano et al., 2024), which query Europe PMC, and extracts candidate causes from the retrieved titles and abstract excerpts; these are appended to the model's video-only differential before history-taking. Queries use the predicted sign, not the confirmed diagnosis. Against the standard consultation without retrieval, three retrieval conditions test whether gains come from leaked answers. **Original** keeps every record; the primary **Source-clean** condition removes the source article, shared-DOI records and near-duplicate titles (token Jaccard similarity ≥ 0.85); and **Strict no-answer** also removes records whose titles or excerpts name the confirmed diagnosis or an accepted equivalent. Two controls test whether gains come from the retrieved content or from the list itself: **Own candidates** supplies the model's Stage 1 differential alone, and **Mismatched retrieval** appends source-clean causes retrieved for a case from another category, truncated to the same length, so that the list matches in length and form but not in content. Outcomes are candidate coverage of the true cause, source-workup coverage, investigation counts and accuracy; Appendix F also audits retained records for same-patient reports.

3.6 Models and Statistical Analysis

We evaluate Gemma-4-31B (Gemma Team, 2026), MiMo-v2.5 (Xiaomi MiMo Team, 2026), MiniMax-M3 (MiniMax, 2026), GPT-5.6-luna (OpenAI, 2026), and Qwen3.8-flash (Qwen Team, 2026b) through a common serving infrastructure, with a fixed endpoint per model and temperature 0. Stage 1 uses eight uniformly sampled budgets, $K \in \{1, 2, 4, 8, 16, 32, 64, 128\}$; standard Stage 2 uses $K = 32$, with temporal training budgets specified in Section 3.5.1. We report 95% percentile intervals from 10,000 case-level bootstrap resamples. Between-condition comparisons resample the same 71 cases under both conditions, reporting percentage-point differences and whether their intervals exclude zero. Appendix G reports a stress test with corrupted history responses.

4 Results

4.1 Visual Recognition and Hypothesis Formation

Stage 1 shows that correctly describing a movement rarely places its cause in the initial differential. Recognition rises unevenly with the frame budget and peaks at 15.5–32.4%, depending on the model (Figure 5); at each model's best budget for aetiological coverage, the true cause is still absent from at least 84% of video-only differentials. Even conditional on correct phenomenology, coverage reaches only 18.9% overall (13.5–22.5% across models). Recognizing the sign and forming a useful aetiological hypothesis are thus distinct hurdles (Appendix E), which leaves open whether video helps later, through the questions and tests it prompts.

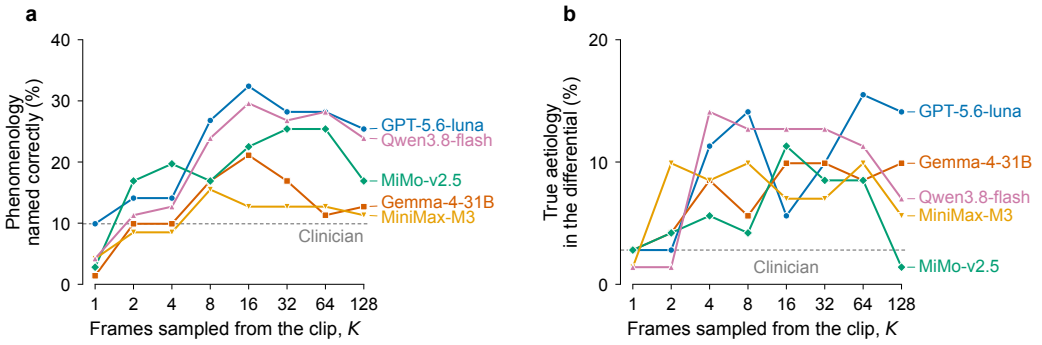


Figure 5: Stage 1 performance on 71 clips. (a) Phenomenology recognition. (b) True aetiology in the differential. Model curves vary the sampled-frame budget; dashed lines show one neurologist's clip-only reference (9.9% recognition; 2.8% aetiological coverage).

Table 2: Consultation outcomes (%; 95% bootstrap intervals in brackets). Seven primary conditions are defined in Section 3.4; multi-turn uses up to ten adaptive rounds. The clinician completed only the video condition (—: not run).

		GPT-5.6-luna	Gemma-4-31B	MiMo-v2.5	MiniMax-M3	Qwen3.8-flash	Clinician
Diagnosis Accuracy (%)	Blind	23.9 [14.1,33.8]	23.9 [14.1,33.8]	8.5 [2.8,15.5]	9.9 [4.2,16.9]	2.8 [0.0,7.0]	—
	Single Frame	36.6 [25.4,47.9]	28.2 [18.3,39.4]	21.1 [12.7,31.0]	15.5 [7.0,23.9]	19.7 [11.3,29.6]	—
	Shuffled	45.1 [33.8,56.3]	28.2 [18.3,39.4]	22.5 [12.7,32.4]	15.5 [7.0,23.9]	12.7 [5.6,21.1]	—
	Video	46.5 [35.2,57.7]	33.8 [22.5,45.1]	22.5 [12.7,32.4]	25.4 [15.5,35.2]	22.5 [12.7,32.4]	35.2 [23.9,46.5]
	Multi-Turn	52.1 [40.8,63.4]	26.8 [16.9,36.6]	19.7 [11.3,29.6]	33.8 [22.5,45.1]	23.9 [14.1,33.8]	—
	Own Words	47.9 [36.6,59.2]	33.8 [22.5,45.1]	22.5 [12.7,32.4]	22.5 [14.1,32.4]	21.1 [12.7,31.0]	—
	Reference	69.0 [57.7,78.9]	35.2 [23.9,46.5]	38.0 [26.8,49.3]	57.7 [45.1,69.0]	29.6 [19.7,40.8]	—
	Oracle Evidence	**93.0** [85.9,98.6]	**73.2** [62.0,83.1]	**78.9** [69.0,87.3]	**83.1** [74.6,91.5]	**81.7** [71.8,90.1]	—
Source-Workup Coverage (%)	Blind	19.3 [13.1,26.1]	22.3 [15.3,29.4]	8.3 [4.2,13.1]	13.6 [8.9,18.8]	2.0 [0.3,4.0]	—
	Single Frame	28.6 [22.3,35.4]	17.3 [11.6,23.6]	16.9 [11.1,23.4]	11.0 [7.2,15.4]	16.3 [10.3,22.8]	—
	Shuffled	32.9 [27.0,39.2]	15.0 [10.8,19.5]	18.9 [13.5,24.7]	15.6 [11.0,20.9]	13.6 [9.3,18.3]	—
	Video	28.9 [22.1,35.9]	20.3 [14.9,26.2]	16.3 [11.4,21.5]	19.3 [14.8,24.2]	16.3 [11.2,21.9]	18.5 [13.1,24.7]
	Multi-Turn	39.2 [32.5,45.9]	17.6 [12.5,23.2]	14.0 [9.1,19.2]	21.6 [15.6,28.3]	10.6 [6.8,14.8]	—
	Own Words	39.2 [32.7,45.8]	18.6 [12.9,24.8]	15.3 [10.7,20.3]	18.3 [13.7,23.2]	15.3 [11.6,19.1]	—
	Reference	50.8 [43.7,57.9]	19.9 [15.3,24.9]	22.3 [17.5,27.2]	30.6 [25.1,36.3]	17.9 [13.4,22.6]	—

The clip-only clinician reference underscores both the difficulty of this step and the effect of differential breadth. One neurologist recognized the phenomenology in $7/71$ clips (9.9%), judged 22 clips unidentifiable, and included the correct aetiology in $2/71$ differentials (2.8%), listing 2.3 diagnoses per clip against 5.1–6.4 for the models; most model configurations exceed both rates.

4.2 Diagnostic Outcomes under Controlled Conditions

Video improves diagnosis by turning what the model sees into better investigations. Across the five models, video raises accuracy over blind input by 9.9–22.5 percentage points, and source-workup coverage rises with it (Table 2): the two models with the largest accuracy gains also gain most in coverage, from 19.3% to 28.9% for GPT-5.6-luna and from 2.0% to 16.3% for Qwen3.8-flash, whereas Gemma-4-31B, the only model whose coverage falls (22.3% to 20.3%), gains least in accuracy. A replay of the diagnosis turn with swapped trajectories confirms this route (Figure 1b; Table 11): swapping in the blind consultation's investigations costs 12.1 points, swapping its history 0.3 (the records settle few of the questions asked), and showing the frames at diagnosis adds 3.1. Part of the benefit runs through derived chart values: neutralising them narrows the video–blind gap from 14.9 to 10.1 points (Appendix C). The benefit of video is thus realised in the work-up it prompts rather than in the model's description of the sign.

Better evidence helps most. Clinician descriptions raise accuracy over video by 1.4–32.4 points and raise coverage most where accuracy rises most, and supplying the decisive investigations raises accuracy to 73.2–93.0% (+39.4–59.2 points): correct interpretation of the sign and acquisition of

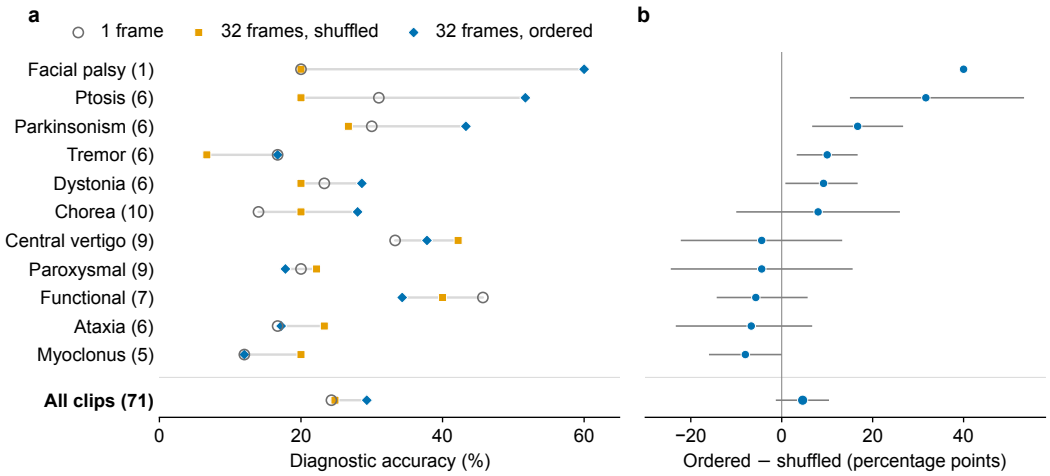


Figure 6: Frame-order effects across clinical categories. Five systems pooled over 71 clips; clip counts in parentheses. (a) Accuracy for one frame, 32 shuffled frames, and 32 ordered frames. (b) Ordered-minus-shuffled accuracy with bootstrap intervals (10,000 clip resamples).

the right evidence both improve diagnosis, acquisition more. By contrast, video, shuffled frames, multi-turn consultation and the model's own words perform alike, consistent with models that read the clip once, largely ignore frame order, and rely on what they can put into words. The model's own description changes accuracy by -2.8 to $+1.4$ points relative to video; ordered video exceeds 32 shuffled frames by a pooled 5.4 points ($[-0.3, +11.3]$; Figure 6); and up to ten adaptive rounds raise accuracy for three models but lower it for two, so single-round consultation stays primary. A single frame trails video by 1.4–9.9 points. The clinician's video consultation reaches 35.2% accuracy, second only to GPT-5.6-luna, with fewer questions (4.6 versus 6.3–16.2) and investigations (3.1 versus 3.5–8.5) than every model. The main contrasts also hold on the 37 cases published in 2025–2026, where prior exposure is least likely (Appendix A), on the 49 clips the neurologist judged identifiable from video alone (Appendix B), and on cases whose records answer more of the history questions the models asked (Appendix C).

4.3 Temporal-Window Training and Downstream Diagnosis

Post-training improves the sign description, and the better description improves the consultation. Every arm feeds the same GPT-5.6-luna consultation, so only the description differs. A blinded judge prefers the adapter's sentence over the untrained model's for 48 clips against 14, mainly because the body part (40% to 66%) and the character of the movement (10% to 54%) are named correctly more often (Appendix D). On identical survey-plus-window inputs, the adapter raises source-workup coverage over untrained Qwen3.5-4B by 3.9–5.2 points, with intervals excluding zero at every budget, and accuracy by 5.2, 8.0 and 6.6 points at $K = 8, 16, 32$ (Table 3; Table 23). Sampling densely within a short segment matters more than where the segment lies: Window minus Random accuracy is only -0.5, $+1.4$ and $+3.3$ points, yet both windowed arms lead a sign-only adapter that spreads the same $32+K$ frame budget over the whole clip by 2.3–6.6 and 1.4–3.3 points at similar coverage (Table 3). A short segment seen at a higher frame rate thus appears more useful than the same frames spread thinly; as with the adapter's accuracy gain, these differences stay within sampling error.

4.4 Literature Retrieval and Evidence Acquisition

Like post-training, retrieval improves diagnosis by acting on evidence acquisition. It first expands the initial hypotheses: the true cause appears in 7.0–9.9% of the models' own video-only differentials but in 62.0–64.8% of the combined lists supplied to them (Appendix F). With these lists, source-clean retrieval moves the work-up toward the treating clinicians' documented one in four of five systems and significantly raises accuracy for MiMo-v2.5 and MiniMax-M3 (Figures 4 and 7). Much of the accuracy effect comes from the breadth of the list rather than its content: the model's own differential alone (**Own candidates**) lowers accuracy for three models, appending an equal-

Table 3: Matched LoRA sampling outcomes (Section 3.5.1; 71 cases; three-run per-clip means). Accuracy and source-workup coverage τ are percentages; brackets give 95% clip-level bootstrap intervals (Appendix D). No window: sign-only adapter on $32+K$ uniform frames (Table 24). W/R/N: window, random and no-window arms.

K	Window accuracy	Random accuracy	No-window accuracy	τ (W / R / N)
8	29.1 [20.7, 38.0]	29.6 [21.6, 38.5]	26.8 [19.2, 34.3]	17.3 / 16.4 / 19.0
16	30.0 [21.1, 39.4]	28.6 [20.2, 37.6]	27.2 [19.2, 35.7]	18.6 / 17.9 / 18.3
32	33.8 [24.4, 43.7]	30.5 [21.6, 39.9]	27.2 [19.2, 35.7]	19.4 / 20.9 / 18.9

Baselines (accuracy; coverage). No adapter (K=8/16/32): 23.9 / 22.1 / 27.2; 13.4 / 13.4 / 14.3 (Table 23).

length list of unrelated causes (**Mismatched retrieval**) raises it again, and only MiMo-v2.5 gains from the retrieved content itself, while GPT-5.6-luna gains from neither appended list (Appendix F). Leaked answers do not explain the gains: **Strict no-answer** retrieval lowers candidate coverage of the true cause by about 13 points, yet the gains of MiMo-v2.5 and MiniMax-M3 remain, and no retained abstract reports the same patient (Figure 7a; Appendix F). Human re-grading of the source-clean diagnoses reproduces these gains (Appendix J.3).

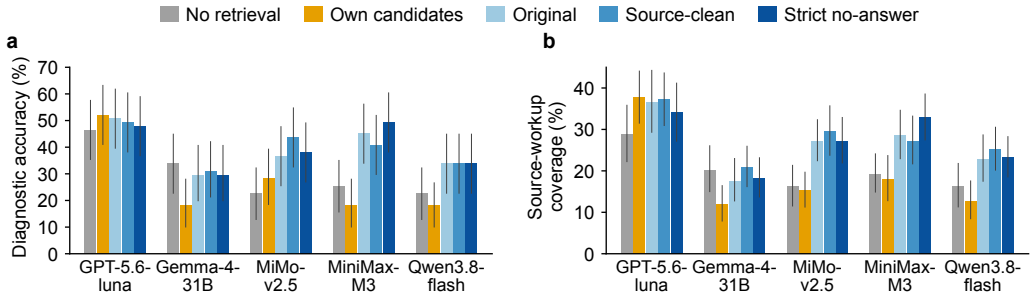


Figure 7: Retrieval acts on evidence acquisition. (a) Diagnostic accuracy and (b) source-workup coverage on the 71 cases, with 95% bootstrap intervals. Own candidates supplies the model's Stage 1 differential alone; retrieval conditions append retrieved causes (Section 3.5.2): Original keeps all records, Source-clean removes the source article and near duplicates, and Strict no-answer also removes records naming the diagnosis.

5 Conclusion

DynamicDx evaluates video-based diagnosis as a consultation, scoring each step from recognizing the sign to ordering tests against a fixed record from the same case report. Across five vision–language models, video raises accuracy over blind input, yet recognizing the sign rarely places its cause in the initial differential and shuffling the frames costs little; the gain arrives instead through the investigations the video prompts. Models thus appear to read a clip once, as a description of what it shows, and the decisive step is turning that description into the right tests: once the decisive investigations are supplied, accuracy reaches 73.2–93.0%. The two interventions improve what enters this step: post-training helps a small describer name the sign, especially from a short, densely sampled segment, and source-clean retrieval widens the initial differential; both bring the work-up closer to the one clinicians documented and, through it, improve diagnosis.

For video-based diagnostic models, seeing better helps when it leads to asking better. Final-answer accuracy hides this route; scoring the work-up against a fixed record reveals it and compares models on identical evidence. These findings rest on 71 published case reports, whose fixed and partly derived records cannot reproduce a live consultation and raise absolute accuracy. Extending the benchmark to other specialties and live recordings would test whether evidence acquisition remains the bottleneck and whether models that use how a sign unfolds, not only how it looks, can ease it.

DISCLOSURE OF AI USE

We used large language models as the systems under evaluation and as tools: GPT-5.6-luna generated the temporal-window supervision, and DeepSeek-V4.1-Flash mapped free-text questions and orders to the case records, ran the retrieval steps and graded outcomes, as described in the methods and appendices. DeepSeek-V4.1-Flash also generated synthetic data: given the confirmed diagnosis and the reported chart, it filled the generic panel of expected investigation values in each chart before any model was evaluated (Section 3.1; Appendix C). We also used generative AI tools to assist with translation, provide feedback on methodology, interpret results, check consistency across the manuscript, revise its structure, wording and LaTeX formatting, and write plotting code.

The authors reviewed the AI-assisted material and take responsibility for the final manuscript, including its data, analyses, claims, and artifacts.

ETHICS STATEMENT

DynamicDx evaluates diagnostic reasoning using patient presentations from published case reports. Patient videos may contain identifiable information; their use and any redistribution require attention to consent, privacy, and source-specific permissions. The benchmark is intended for research evaluation, not clinical deployment or autonomous medical decision-making. Its small, uneven, case-report-derived sample limits the representativeness of the findings. Human-expert evaluations and automated grading procedures are described in the paper.

REPRODUCIBILITY STATEMENT

The main text describes benchmark construction, the consultation protocol, evaluation metrics, and statistical procedures. The appendices provide case-source attribution, prompts, retrieval and filtering procedures, temporal-window training settings, and grading and order-matching audits. Together, these materials document the experimental setup and support inspection of the reported evaluations. An anonymized repository with the data, prompts and evaluation code is available at `https://anonymous.4open.science/r/dynamicDx-53C8`.

REFERENCES

Wilson F Abdo, Bart PC Van De Warrenburg, David J Burn, Niall P Quinn, and Bastiaan R Bloem. The clinical approach to movement disorders. *Nature Reviews Neurology*, 6(1):29–37, 2010.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018.

David Ahmedt-Aristizabal, Mohammad Ali Armin, Zeeshan Hayder, Norberto Garcia-Cairasco, Lars Petersson, Clinton Fookes, Simon Denman, and Aileen McGonigal. Deep learning approaches for seizure video analysis: A review. *Epilepsy & Behavior*, 154:109735, 2024.

Alberto Albanese, Kailash Bhatia, Susan B Bressman, Mahlon R DeLong, Stanley Fahn, Victor SC Fung, Mark Hallett, Joseph Jankovic, Hyder A Jinnah, Christine Klein, et al. Phenomenology and classification of dystonia: a consensus update. *Movement Disorders*, 28(7):863–873, 2013.

Efthalia Angelopoulou, Christos Koros, Evangelia Stanitsa, Ioannis Stamelos, Dionysia Kontaxopoulou, Stella Fragkiadaki, John D Papatriantafyllou, Evangelia Smaragdaki, Kalliopi Vourou, Dimosthenis Pavlou, et al. Neurological examination via telemedicine: an updated review focusing on movement disorders. *Medicina*, 60(6):958, 2024.

Kailash P Bhatia, Peter Bain, Nin Bajaj, Rodger J Elble, Mark Hallett, Elan D Louis, Jan Raethjen, Maria Stamelou, Claudia M Testa, Guenther Deuschl, et al. Consensus statement on the classification of tremors. from the task force on tremor of the International Parkinson and Movement Disorder Society. *Movement Disorders*, 33(1):75–87, 2018.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2917–2927, 2022.

Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29071–29082. IEEE, 2025.

Pat Croskerry. A universal model of diagnostic reasoning. *Academic Medicine*, 84(8):1022–1028, 2009.

DeepSeek-AI. DeepSeek-V4.1-Flash. Large language model. Model string: `deepseek/deepseek-v4.1-flash`, 2026. URL `https://openrouter.ai/deepseek/deepseek-v4.1-flash`.

Lazzaro Di Biase, Pasquale Maria Pecoraro, and Francesco Bugamelli. AI video analysis in Parkinson's disease: a systematic review of the most accurate computer vision tools for diagnosis, symptom monitoring, and therapy management. *Sensors*, 25(20):6373, 2025.

Arthur S Elstein and Alan Schwartz. Clinical problem solving and diagnostic decision making: selective review of the cognitive literature. *BMJ*, 324(7339):729–732, 2002.

Arthur S Elstein, Lee S Shulman, and Sarah A Sprafka. *Medical problem solving: An analysis of clinical reasoning*. Harvard University Press, 1978.

Europe PMC Consortium. Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Research*, 43(D1):D1042–D1048, 2015.

Maximilian U Friedrich, Anna-Julia Roenn, Chiara Palmisano, Jane Alty, Steffen Paschen, Guenther Deuschl, Chi Wang Ip, Jens Volkmann, Muthuraman Muthuraman, Robert Peach, et al. Validation and application of computer vision algorithms for video-based tremor analysis. *npj Digital Medicine*, 7(1):165, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 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*, pp. 24108–24118, 2025.

Michael A Gargano, Nicolas Matentzoglu, Ben Coleman, Eunice B Addo-Lartey, Anna V Anagnostopoulos, Joel Anderton, Paul Avillach, Anita M Bagley, Eduard Bakštein, James P Balhoff, et al. The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346, 2024.

Gemma Team. Gemma 4 technical report, 2026. URL `https://arxiv.org/abs/2607.02770`.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913, 2017.

Deepak Gupta and Dina Demner-Fushman. Overview of the MedVidQA 2022 shared task on medical video question-answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pp. 264–274, 2022.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158, 2023.

Helga Haberfehlner, Shankara S Van De Ven, Sven A Van Der Burg, Florian Huber, Sonja Georgievska, Ignazio Aleo, Jaap Harlaar, Laura A Bonouvrié, Marjolein M Van Der Krogt, and Annemieke I Buizer. Towards automated video-based assessment of dystonia in dyskinetic cerebral palsy: A novel approach using markerless motion tracking and machine learning. *Frontiers in Robotics and AI*, 10:1108114, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. Medical adaptation of large language and vision-language models: Are we making progress? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12143–12170, 2024.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. CRAFT-MD: A conversational evaluation framework for comprehensive assessment of clinical LLMs. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, 31(1):77–86, 2025.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251, 2018.

Andrea Gurmankin Levy, Aaron M Scherer, Brian J Zikmund-Fisher, Knoll Larkin, Geoffrey D Barnes, and Angela Fagerlin. Prevalence of and factors associated with patient nondisclosure of medically relevant information to clinicians. *JAMA Network Open*, 1(7):e185293, 2018.

Jiahui Li, Ruili Fang, Zishuai Liu, WenZhan Song, Jin Lu, and Fei Dou. DeepArrhythmia: Segment-contextualized ECG arrhythmia classification via selective evidence acquisition. *arXiv preprint arXiv:2605.16441*, 2026a.

Jiahui Li, Yida Zhang, Zixuan Zeng, Jiayu Chen, Yingjian Song, Yin Xiao, Nishan Dong, Junjie Lu, Younghoon Kwon, Xiang Zhang, Jin Lu, Wenzhan Song, and Fei Dou. Peak-Detector: Explainable peak detection via instruction-tuned large language models in physiological signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 10(2): 1–44, June 2026b. doi: 10.1145/3810224.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28858–28888, 2024.

Valentin Liévin, Anil Palepu, Wei-Hung Weng, Khaled Saab, David Stutz, Yong Cheng, Kavita Kulkarni, S Sara Mahdavi, Joëlle Barral, Dale R Webster, et al. Towards conversational artificial intelligence for disease management. *Nature*, 655(8125):1292–1299, 2026. doi: 10.1038/s41586-026-10764-5.

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

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. Towards accurate differential diagnosis with large language models. *Nature*, 642(8067):451–457, 2025.

MiniMax. MiniMax-M3. Large language model. Model string: `minimax/minimax-m3`, 2026. URL `https://huggingface.co/MiniMaxAI/MiniMax-M3`. Open-weight model card. Accessed July 2026.

Salhadin Mohammed, Selam Kifelew, Fikru Tsehayneh, Abel Teklit Haile, and E. H. Gebrehiwot. Genetically confirmed Charcot–Marie–Tooth disease type 2A manifesting with postural tremor: a case report. *Journal of Medical Case Reports*, 18:571, 2024. doi: 10.1186/s13256-024-04945-x.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Danielle Newton, Margaret McGurn, Daniella I Hernandez, Nora C Hernandez, Mazen Elkurd, and Elan D Louis. Through the looking glass: remote versus in-person videotaped neurologic assessment of essential tremor. *Movement Disorders Clinical Practice*, 9(1):87–90, 2022.

OpenAI. GPT-5.6 system card. OpenAI Deployment Safety Hub, July 2026. URL `https://deploymentsafety.openai.com/gpt-5-6`. Covers GPT-5.6 Sol, Terra and Luna.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, 2026a.

Qwen Team. Qwen3.8-Flash. Large language model. Model string: `qwen/qwen3.8-flash`, 2026b. URL `https://openrouter.ai/qwen/qwen3.8-flash`.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of Gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Khaled Saab, Chunjong Park, Tim Strother, Jan Freyberg, David GT Barrett, Yong Cheng, Wei-Hung Weng, David Stutz, Nenad Tomasev, Anil Palepu, et al. Advancing conversational diagnostic AI with multimodal reasoning. *Nature Medicine*, 32(5):1726–1736, 2026. doi: 10.1038/s41591-026-04371-0.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: a multimodal benchmark for tool-using clinical AI agents. *npj Digital Medicine*, 9:499, 2026. doi: 10.1038/s41746-026-02674-7.

Meet Shah, Jason Gusdorf, Anil Palepu, Chunjong Park, Jack W O'Sullivan, Vishnu Ravi, Tim Strother, Pavel Dubov, Aliya Rysbek, Toshiyuki Fukuzawa, et al. Towards conversational medical AI with eyes, ears and a voice. *arXiv preprint arXiv:2605.09272*, 2026.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

Ragini Srinivasan, Hilla Ben-Pazi, Marieke Dekker, Esther Cubo, Bas Bloem, Emile Moukheiber, Josefa Gonzalez-Santos, and Mark Guttman. Telemedicine for hyperkinetic movement disorders. *Tremor and Other Hyperkinetic Movements*, 10, 2020.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128. IEEE, 2025.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450, 2025.

Floris S Weller, Jaap F Hamming, Sjoerd Repping, and Leti van Bodegom-Vos. What information sources do Dutch medical specialists use in medical decision-making: a qualitative interview study. *BMJ Open*, 13(10):e073905, 2023.

Floris S Weller, Sjoerd Repping, Jaap F Hamming, and Leti van Bodegom-Vos. Systematic review and meta-analysis of information source usage: do medical specialists use the best evidence for clinical decision-making? *BMJ Open*, 16(3):e099887, 2026.

Johanna I Westbrook, A Sophie Gosling, and Enrico Coiera. Do clinicians use online evidence to support patient care? a study of 55,000 clinicians. *Journal of the American Medical Informatics Association*, 11(2):113–120, 2004.

Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. AdaFrame: Adaptive frame selection for fast video recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1278–1287. IEEE, 2019.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. RULE: Reliable multimodal RAG for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1081–1093, 2024.

Xiaomi MiMo Team. MiMo-V2.5. Model release, April 2026. URL https://mimo.xiaomi.com/mimo-v2-5.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6233–6251, 2024.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Improving retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pp. 199–214. World Scientific, 2025.

Linli Yao, Haoning Wu, Kun Ouyang, Yuanxing Zhang, Caiming Xiong, Bei Chen, Xu Sun, and Junnan Li. Generative frame sampler for long video understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17900–17917, 2025.

Appendix Contents

A	Source Attribution and Dataset Characteristics	16
B	Decidability of the Visual Task	19
C	Evidence Acquisition and Investigation Selection	21
D	Temporal-Window Training: A Paired Analysis	28
E	Recognition across Frame Budgets and Clinical Categories	30
F	Retrieval: Candidate Coverage and a Consultation Example	32
G	Robustness to Corrupted History Responses	34
H	Prompts	37
I	Temporal-Window Training: Reproducibility Details	42
J	Grading and Order-Matching Audits	47

A Source Attribution and Dataset Characteristics

The 71 cases are drawn from 66 open-access articles and their supplementary patient videos. Table 7 lists the source articles and their licences, retrieved from Europe PMC at build time, together with the country of the first author's affiliation and the category, age and sex of each patient filmed. Each PMCID links to its record at `https://europepmc.org/article/PMC/<PMCID>`.

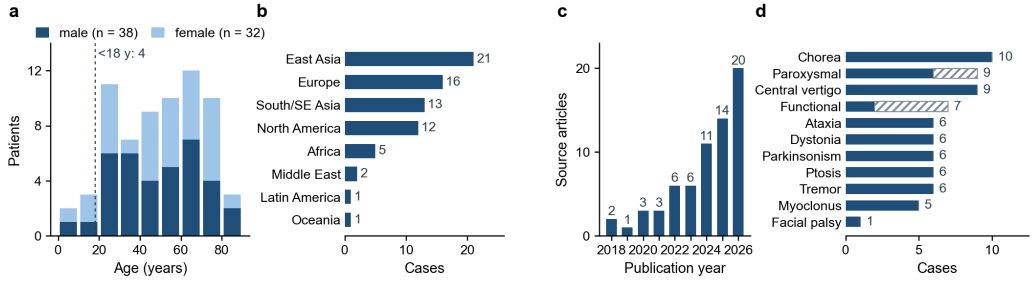


Figure 8: **Who is in the benchmark. a**, Age of the 67 patients whose age the source states, stacked by sex; four further patients are described only as adults and one has no stated sex. The dashed line marks 18 years. **b**, Region of the first author's affiliation, by case (23 countries). **c**, Publication year of the 66 source articles. **d**, Cases per category; hatched segments are the eight cases drawn from the three articles that contribute more than one patient.

Clustering. Resampling articles instead of cases leaves the intervals essentially unchanged. Sixty-three articles contribute one case each; three contribute more than one, each clip showing a different patient: a two-patient functional-head-tremor series (`PMC12999016`), a three-patient functional-gait video series (`PMC7455329`) and a three-patient ictal-body-turning series (`PMC13101308`). Intervals that resample the 66 articles have endpoints within 1 percentage point of the case-level intervals and nearly the same widths, for every system and the clinician (Table 5).

Demographics. Figure 8 summarises the sample. Age is stated for 67 patients (median 52, IQR 30–66, range 2–90); four are described only as adults. Four patients are under 18 (ages 2, 4, 15 and 17), 19 are aged 18–39, 19 are 40–59, 19 are 60–74 and 6 are 75 or older. Thirty-eight are male, 32 female and one is not stated; for three, sex is not given in the text and was read off the video. The sample is thus mainly adult, with too few patients under 18 for a subgroup estimate.

Geography and venue. By first-author affiliation the 66 articles come from 23 countries: China (13), India (11), the United States (9 articles, 12 cases), Japan (4), South Korea (4), Italy (3), and 17 countries with one or two each. By region the cases are East Asia 21, Europe 16, South and South-East Asia 13, North America 12, Africa 5, Middle East 2, Latin America 1 and Oceania 1. The articles appear in 36 journals, most often *Cureus* (9), *Annals of Indian Academy of Neurology* (5) and *Journal of Movement Disorders* (4). Publication years (2018–2026) skew recent by construction of the search: 34 of 66 articles are from 2025–2026, 17 from 2023–2024 and 15 earlier.

Diagnostic labels. The 71 confirmed diagnoses are almost all distinct entities. The only repeated labels are three ictal-body-turning generalized epilepsies from one series, two functional head tremors from one series, four hyperglycaemia-related choreas, two Graves'-related choreas and two myasthenia gravis presentations; every other case is the sole example of its diagnosis. The case-report sample emphasizes rare and atypical presentations that often require a specific confirming test. The facial-palsy category has one clip, so its statistics carry no interval.

Performance by subgroup. Publication period and region show the largest gradients in video-condition accuracy (Table 6): every model solves articles from 2022 or earlier more often (29–71%) than articles from 2023–2024 (12–29%), and so does the clinician (47% vs. 18%). Pre-training exposure and presentation typicality could both contribute. Cases from South and South-East Asian articles ($n = 13$) are also easier for every system and the clinician (46–69%, against 16–43% for the

Table 4: **Main contrasts by publication period of the source article.** Stage-2 diagnosis accuracy (%) averaged over the five systems; differences in percentage points with 95% paired case-level bootstrap intervals. Clean: source-clean retrieval.

period	n	video	video − blind	oracle − video	reference − video	video − single	video − shuffled	clean − video
All	71	30.1	+16.3 [+9.3, +23.7]	+51.8 [+43.4, +60.0]	+15.8 [+8.7, +22.8]	+5.9 [−0.3, +12.4]	+5.4 [−0.3, +11.3]	+9.6 [+2.5, +16.6]
≤2022	17	49.4	+32.9 [+14.1, +49.4]	+38.8 [+23.5, +55.3]	+12.9 [−1.2, +27.1]	+16.5 [+3.5, +31.8]	+7.1 [−5.9, +21.2]	+2.4 [−15.3, +21.2]
2023–24	17	22.4	+12.9 [+2.4, +23.6]	+60.0 [+43.5, +76.5]	+3.5 [−8.2, +16.5]	+3.5 [−7.1, +14.1]	+12.9 [+5.9, +21.2]	+7.1 [−5.9, +21.2]
2025–26	37	24.9	+10.3 [+1.6, +18.9]	+54.1 [+42.2, +65.4]	+22.7 [+13.5, +32.4]	+2.2 [−5.9, +10.8]	+1.1 [−7.0, +9.2]	+14.1 [+5.4, +22.7]

Table 5: **Article-clustered bootstrap intervals** for diagnosis accuracy in the video condition, 10,000 resamples. Clustered resampling draws the 66 source articles with replacement, keeping all their cases; the width ratio divides the clustered interval's width by the case-level one in Table 2.

	GPT-5.6-luna	Gemma-4-31B	MiMo-v2.5	MiniMax-M3	Qwen3.8-flash	Clinician
95% CI by article	[34.8, 58.1]	[22.9, 45.3]	[13.0, 33.3]	[15.7, 35.7]	[13.6, 32.4]	[24.2, 46.5]
width ratio	1.04	1.00	1.03	1.01	0.95	0.99

other $n = 58$), but these groups differ in case mix, including Graves'-related and hyperglycaemic chorea and dopa-responsive dystonia. No sex difference is consistent across systems. Because subgroups differ in difficulty, the main analyses compare conditions within the same cases.

Main contrasts by publication period. Earlier publication makes cases easier once the video is shown, not from the diagnosis alone. For articles from 2022 or earlier, blind accuracy is 16.5%, close to 14.6% for 2025–2026, whereas video raises it by 32.9 points against 10.3 (Table 4). The main contrasts hold in the most recent period, where prior exposure is least likely: on the 37 cases from 2025–2026, video exceeds blind input by 10.3 points, oracle evidence exceeds video by 54.1 points and the reference description adds 22.7 points. Source-clean retrieval adds 14.1 points in this period and 2.4 points for articles from 2022 or earlier, consistent with retrieval widening hypotheses most where models know least. Prior exposure may contribute to the advantage of older cases, but it does not produce the main conclusions.

Distribution. The benchmark distributes annotations, structured patient and chart tables, and a script for retrieving the supplementary videos and reproducing the excerpts locally. It does not redistribute the trimmed and re-encoded clips: these are derivative works, which the CC BY-NC-ND licences of 18 source articles do not permit us to distribute.

Table 7: **Source articles.** Licence, country of the first author's affiliation, and the category, age and sex of each patient filmed (–: not stated in the source).

Source	PMCID	Licence	Country	Patient(s)
Abbassi O et al., Cureus, 2024	PMC11283634	CC BY	Morocco	chorea, 66F
Antova I et al., Cureus, 2025	PMC12596229	CC BY	Bulgaria	chorea, 62M
Bartl M et al., Neurological research and practice, 2020	PMC7713151	CC BY	Germany	functional, 48F
Batot C et al., Tremor and other hyperkinetic movements, 2022	PMC9122005	CC BY	France	chorea, 90M
Calvo PA et al., Cureus, 2024	PMC11586875	CC BY	Portugal	chorea, 84F
Chen W et al., Frontiers in neurology, 2022	PMC9815763	CC BY	China	chorea, 44F
Cutellè R et al., Epileptic disorders : international epilepsy journal with videotape, 2026	PMC13276695	CC BY	Italy	paroxysmal, 48M
Davis K et al., Journal of education & teaching in emergency medicine, 2023	PMC10414977	CC BY	USA	ptosis, 46F
Dentoni M et al., Cerebellum, 2024	PMC11585521	CC BY	Italy	ataxia, 71F
Doumbouya I et al., Cureus, 2023	PMC10051035	CC BY	Guinea	chorea, 64F
He Y et al., Frontiers in pediatrics, 2025	PMC12440885	CC BY	China	paroxysmal, 4F
Horinouchi T et al., Clinical case reports, 2018	PMC6230673	CC BY	Japan	paroxysmal, 23M

Table 7, continued

Source	PMCID	Licence	Country	Patient(s)
Hu Q et al., Frontiers in genetics, 2025	PMC12558638	CC BY	China	tremor, 28M
Inzirillo K et al., Cureus, 2025	PMC12572359	CC BY	USA	facial palsy, 53M
Jimsheleishvili S et al., Neurological sciences : official journal of the Italian Neurological Society and of the Italian Society of Clinical Neurophysiology, 2025	PMC12678565	CC BY	USA	ataxia, 59F
Liang W et al., Medicine, 2026	PMC13008232	CC BY	China	ataxia, 53M
Liu C et al., Medicine, 2026	PMC13166699	CC BY	China	tremor, 66F
Liu Y et al., Frontiers in immunology, 2026	PMC12979508	CC BY	China	vertigo, 26F
Long X., JPRAS open, 2026	PMC13049410	CC BY	China	dystonia, 59F
Nambiar SV et al., Clinical parkinsonism & related disorders, 2026	PMC13080482	CC BY	India	parkinsonism, 50M
Nie J et al., BMC neurology, 2026	PMC13085407	CC BY	China	ataxia, 38M
Oliveira D et al., Cureus, 2023	PMC10656108	CC BY	Portugal	dystonia, 70M
Pham K et al., Cureus, 2020	PMC7055013	CC BY	USA	chorea, 76F
Rohatgi SJ et al., Cureus, 2024	PMC11098550	CC BY	India	chorea, 48M
Roy S et al., Tremor and other hyperkinetic movements, 2026	PMC12904113	CC BY	India	dystonia, 17M
Song DJ et al., Frontiers in immunology, 2025	PMC12833086	CC BY	China	vertigo, 15F
Srimanan W et al., Cureus, 2026	PMC13186686	CC BY	Thailand	vertigo, 34M
Stephens T et al., Clinical practice and cases in emergency medicine, 2026	PMC13135425	CC BY	USA	vertigo, 53F
Stone J et al., Movement disorders : official journal of the Movement Disorder Society, 2026	PMC13206381	CC BY	UK	functional, —-
Tai XY et al., Movement disorders : official journal of the Movement Disorder Society, 2025	PMC12371612	CC BY	UK	parkinsonism, 54M
Tchopev ZN et al., Frontiers in neurology, 2018	PMC5786569	CC BY	USA	paroxysmal, 36M
Tomic S et al., Frontiers in neurology, 2024	PMC11021691	CC BY	Croatia	parkinsonism, 70M
Totuk Ö et al., Case reports in psychiatry, 2025	PMC12297136	CC BY	Turkey	myoclonus, 73M
Vaghi G et al., Cerebellum, 2024	PMC11102397	CC BY	Italy	vertigo, 71F
Wang RY et al., Journal of medical case reports, 2026	PMC13224678	CC BY	China	ataxia, 52F
Xie N et al., Tremor and other hyperkinetic movements, 2024	PMC11639689	CC BY	China	tremor, 20M
Zhang Q et al., Journal of medical case reports, 2019	PMC6883617	CC BY	China	vertigo, 62M
Chang HJ et al., Journal of clinical neurology, 2023	PMC9982172	CC BY-NC	Korea	dystonia, 42M
Lee GJ et al., Brain & NeuroRehabilitation, 2022	PMC9833462	CC BY-NC	Korea	myoclonus, 29F
Lim SY et al., Journal of movement disorders, 2024	PMC11082598	CC BY-NC	Malaysia	parkinsonism, 73F
Simma K et al., Oxford medical case reports, 2025	PMC12741441	CC BY-NC	Morocco	ataxia, 39M
Sugiyama A et al., Case reports in neurology, 2023	PMC10359687	CC BY-NC	Japan	tremor, 68M
Surisetti BK et al., Journal of movement disorders, 2021	PMC8490194	CC BY-NC	India	vertigo, 2M
Turgunkhujaev O et al., Journal of movement disorders, 2026	PMC13175740	CC BY-NC	Russia	parkinsonism, 58F
Youn J et al., Journal of movement disorders, 2023	PMC10548079	CC BY-NC	Korea	myoclonus, 60M
Zhou LC et al., The Journal of international medical research, 2025	PMC12745546	CC BY-NC	China	vertigo, 63M

Table 7, continued

Source	PMCID	Licence	Country	Patient(s)
Baizabal-Carvallo JF et al., Movement disorders clinical practice, 2025 (2 clips)	PMC12999016	CC BY-NC-ND	USA	functional, 61M; functional, 83M
Değirmenci MFK et al., BMC ophthalmology, 2024	PMC11495012	CC BY-NC-ND	Turkey	ptosis, 40F
Hormaza-Jaramillo A et al., Clinical neurophysiology practice, 2026	PMC12933468	CC BY-NC-ND	Colombia	myoclonus, 30M
Jogi H et al., Annals of Indian Academy of Neurology, 2026	PMC12962376	CC BY-NC-ND	India	dystonia, 62F
Koutsis G et al., Journal of the peripheral nervous system : JPNS, 2026	PMC12960836	CC BY-NC-ND	Greece	ptosis, 45F
Kuma A et al., Clinical Case Reports, 2026	PMC13093780	CC BY-NC-ND	Ethiopia	tremor, 32F
Lee D et al., BMC ophthalmology, 2025	PMC12606889	CC BY-NC-ND	Korea	ptosis, 24F
Mathuram D et al., Annals of Indian Academy of Neurology, 2025	PMC12798911	CC BY-NC-ND	India	paroxysmal, 78F
Mohammed S et al., Journal of medical case reports, 2024	PMC11600610	CC BY-NC-ND	Ethiopia	tremor, 34M
Nagabushana D et al., BMC neurology, 2026 (3 clips)	PMC13101308	CC BY-NC-ND	USA	paroxysmal, 26F; paroxysmal, 24F; paroxysmal, 29M
Nakatake N et al., AACE clinical case reports, 2025	PMC11973688	CC BY-NC-ND	Japan	chorea, 73F
Nonnekes J et al., Neurology, 2020 (3 clips)	PMC7455329	CC BY-NC-ND	Netherlands	functional, –M; functional, –F; functional, –M
Okadome T et al., Epilepsy & behavior reports, 2022	PMC9062418	CC BY-NC-ND	Japan	paroxysmal, 26M
Rodin RE et al., Annals of clinical and translational neurology, 2024	PMC11093232	CC BY-NC-ND	USA	chorea, 74F
Saibaba J et al., Annals of Indian Academy of Neurology, 2026	PMC12962431	CC BY-NC-ND	India	parkinsonism, 52M
Verma R et al., Journal of neurosciences in rural practice, 2022	PMC9357509	CC BY-NC-ND	India	dystonia, 19F
Vinny PW et al., Journal of neurosciences in rural practice, 2021	PMC8064859	CC BY-NC-ND	India	ptosis, 75M
Yeow D et al., Annals of clinical and translational neurology, 2026	PMC13071100	CC BY-NC-ND	Australia	myoclonus, 41M
Maramattom BV., Annals of Indian Academy of Neurology, 2021	PMC8061496	CC BY-NC-SA	India	vertigo, 64M
Pedapati R et al., Annals of Indian Academy of Neurology, 2022	PMC9350803	CC BY-NC-SA	India	ptosis, 24M

B Decidability of the Visual Task

Low sign recognition could mean that models miss a visible sign, that a clip does not show the sign well enough for any observer, or that the reference description draws on the report rather than the video; the inclusion criterion (Section 3.1) was applied with the report in hand. A diagnosis-blind decidability label and an audit of the reference descriptions separate the three. Recognition stays low even on clips a neurologist could read, harder clips lower accuracy, and the contrasts between visual conditions do not depend on them.

A diagnosis-blind decidability label. The clinician's Stage-1 record supplies an independent judgement of decidability. The clinician viewed each clip before receiving any history, investigation results or diagnosis, and for 22 of 71 clips recorded the sign as not identifiable (uncertain, normal or none). We designate these 22 clips as *clinician-undecidable* and the remaining 49 as *clinician-decidable*. The undecidable clips are distributed across nine of the eleven categories (paroxysmal 4; dystonia, functional, ptosis and central vertigo 3 each; chorea and myoclonus 2 each; ataxia and

Table 6: **Diagnosis accuracy (%) in the video condition by subgroup.** The denominator is the number of cases in the subgroup; a case without a usable answer counts as wrong. Strata with $n < 10$ are shown for completeness only.

subgroup	n	GPT-5.6-luna	Gemma-4-31B	MiMo-v2.5	MiniMax-M3	Qwen3.8-flash	Clinician
male	38	47.4	34.2	23.7	28.9	18.4	39.5
female	32	43.8	34.4	21.9	21.9	25.0	28.1
age <18	4	50.0	50.0	25.0	50.0	25.0	50.0
age 18–59	38	42.1	34.2	21.1	18.4	23.7	28.9
age ≥60	25	44.0	28.0	28.0	28.0	16.0	36.0
East Asia	21	42.9	33.3	19.0	28.6	9.5	19.0
South/SE Asia	13	61.5	53.8	46.2	46.2	53.8	69.2
Europe	16	50.0	37.5	18.8	18.8	31.2	43.8
North America	12	41.7	25.0	16.7	16.7	16.7	25.0
Africa, Middle East, Latin America, Oceania	9	33.3	11.1	11.1	11.1	0.0	22.2
published ≤2022	17	70.6	52.9	47.1	47.1	29.4	47.1
published 2023–24	17	29.4	29.4	23.5	17.6	11.8	17.6
published 2025–26	37	43.2	27.0	10.8	18.9	24.3	37.8
single-case article	63	42.9	34.9	25.4	23.8	22.2	31.7
multi-case article	8	75.0	25.0	0.0	37.5	25.0	62.5

parkinsonism 1 each; tremor and facial palsy 0). The list is released with the benchmark. We retain all 71 clips in the primary analysis and use this label for stratified results.

Phenomenology recognition. Table 8 reports Stage-1 recognition and Stage-2 accuracy by decidability. Recognition does not differ between decidable and undecidable clips in any system (differences of -8.4 to $+7.9$ points), and stays low (10.2–30.6%) even on clinician-decidable clips.

Diagnostic accuracy. Accuracy is higher on the 49 clinician-decidable clips than on the 22 undecidable clips in every visual condition: 32.2% versus 25.5% with video, 26.9% versus 20.0% with shuffled frames, and 27.3% versus 17.3% with a single frame (means over the five systems), with the same direction in 13 of 15 system–condition pairs. The clinician, who also had the history and investigations, scores similarly on the two subsets (34.7% and 36.4%). Part of the low model accuracy may therefore reflect hard-to-read clips.

The contrasts between visual conditions do not depend on these clips (Table 9). On the 49 decidable clips, video exceeds shuffled frames by 5.3 points and a single frame by 4.9 points, close to the 5.4 and 5.9 points on all 71 clips; shuffled frames and a single frame differ by 0.4 points. The small advantage of video over a single frame, and the lack of a reliable ordering effect, hold when undecidable clips are removed.

Visual grounding of the reference descriptions. The reference condition supplies a clinician-written description of the visible sign, prepared with access to the source report. We decomposed all 71 descriptions into atomic claims and classified each as observable or not observable in a silent video. Of 286 claims, 34 (11.9%), in 27 descriptions, were classified as not observable. Most are temporal or volitional qualifiers such as continuous, involuntary, or sudden. Nine phrases in eight descriptions exceed the visual evidence: reported symptoms (double vision, painful), contextual information (arising from sleep with vocalisation, bed-bound, alert and able to speak), the label seizure in two descriptions, and one editorial remark about lumbar hyperlordosis. We removed these phrases from the released descriptions and updated the reference-condition results and student training targets accordingly. The complete audit is released with the descriptions.

Table 8: **Results by clinician-judged decidability.** The neurologist judged 22 clips unidentifiable from the video alone, before seeing case material. Recognition is the Stage-1 sign grade at 32 frames (*correct*); accuracy is Stage-2 video-condition accuracy. Cells give decidable (n=49) / undecidable (n=22); the gap is their difference with a 10,000-resample bootstrap interval.

	sign recognition (%)		diagnosis accuracy (%)	
	dec. / undec.	gap [95% CI]	dec. / undec.	gap [95% CI]
GPT-5.6-luna	30.6 / 22.7	+7.9 [−14.4, +29.2]	46.9 / 45.5	+1.5 [−23.7, +26.9]
Gemma-4-31B	14.3 / 22.7	−8.4 [−28.7, +11.3]	36.7 / 27.3	+9.5 [−13.7, +31.7]
MiMo-v2.5	26.5 / 22.7	+3.8 [−18.5, +24.0]	24.5 / 18.2	+6.3 [−14.8, +26.1]
MiniMax-M3	10.2 / 18.2	−8.0 [−27.7, +9.7]	30.6 / 13.6	+17.0 [−2.8, +35.8]
Qwen3.8-flash	26.5 / 27.3	−0.7 [−23.5, +21.1]	22.4 / 22.7	−0.3 [−22.1, +19.9]

Table 9: **Visual conditions on the 49 clinician-decidable clips.** Stage-2 diagnosis accuracy (%) with 32 ordered frames (video), 32 shuffled frames, and a single frame, and paired differences with 95% case-level bootstrap intervals. The last row averages the five systems, resampling cases jointly.

	video	shuffled	single	video − shuffled	video − single	shuffled − single
GPT-5.6-luna	46.9	51.0	44.9	−4.1 [−16.3, +6.1]	+2.0 [−12.2, +16.3]	+6.1 [−8.2, +20.4]
Gemma-4-31B	36.7	28.6	28.6	+8.2 [−4.1, +20.4]	+8.2 [−6.1, +22.4]	+0.0 [−14.3, +14.3]
MiMo-v2.5	24.5	26.5	22.4	−2.0 [−18.4, +16.3]	+2.0 [−14.3, +18.4]	+4.1 [−10.2, +18.4]
MiniMax-M3	30.6	14.3	18.4	+16.3 [+2.0, +30.6]	+12.2 [−2.0, +28.6]	−4.1 [−16.3, +8.2]
Qwen3.8-flash	22.4	14.3	22.4	+8.2 [−2.0, +20.4]	+0.0 [−14.3, +14.3]	−8.2 [−22.4, +6.1]
Mean of five	32.2	26.9	27.3	+5.3 [−1.6, +12.2]	+4.9 [−3.3, +13.5]	−0.4 [−7.8, +6.9]

C Evidence Acquisition and Investigation Selection

This appendix supports the claim that video helps mainly through the investigation results it obtains (Section 4.2). Section C.1 traces the video gain, Section C.2 describes what the records contain and how τ is scored, and Section C.3 compares model-selected work-ups with fixed and random lists.

C.1 How the Video Gain Reaches the Diagnosis

Where the video gain arises. The video gain is concentrated in consultations where video changes which decisive evidence is obtained. For each system we split the 71 cases by whether the video consultation obtains more, the same, or fewer decisive entries than the blind consultation, and attribute the accuracy change to each group (Table 10). Pooled over systems, 15.8 of the 16.3-point gain comes from the 35% of consultations in which video obtains a decisive entry that blind input missed; within this group accuracy rises by 44.4 points. Order counts do not rise with video except for Qwen3.8-flash (Table 15), yet consultations with at least one decisive entry rise from 115 to 188 of 355. Consultations that recover more documented history but the same decisive evidence contribute 0.6 points, whereas those with more decisive evidence contribute 16.0 points whatever the change in history. The replay below tests these paths directly.

Trajectory replay. At the diagnosis step, the video gain runs mainly through the investigation results the video consultation obtained. We re-ran only the final diagnosis turn, always with the 32 video frames, while crossing the source of the history (questions and answers) with the source of the investigations (orders and returned results): each came either from the video or from the blind consultation of the same case (Table 11). Replaying the video trajectory reproduces the original video accuracy (29.3% versus 30.1%, means over systems). Replacing the investigations of the video consultation with the blind ones lowers accuracy by 12.1 points whether the history comes from the video or the blind consultation, whereas replacing the history changes accuracy by 0.3 points with either set of investigations. Showing the frames at diagnosis to an otherwise blind trajectory adds 3.1 points. Most of the 15.5-point replayed gain therefore enters the diagnosis through the returned investigations rather than through the history or a second look at the video.

Table 10: **Where the video gain arises.** For each system the 71 cases are split by whether the video consultation obtains more, the same, or fewer distinct decisive entries than the blind consultation. Each cell gives the number of cases and that group's contribution to the video − blind accuracy gain (percentage points of the 71-case rate); contributions sum to the gain. The last column counts consultations with at least one decisive entry, blind → video.

system	gain	more decisive	same decisive	fewer decisive	cases with a decisive entry
GPT-5.6-luna	$+22.5$	24: $+16.9$	39: $+9.9$	8: -4.2	$34 \rightarrow 45$
Gemma-4-31B	$+9.9$	18: $+11.3$	40: $+4.2$	13: -5.6	$34 \rightarrow 38$
MiMo-v2.5	$+14.1$	27: $+14.1$	33: $+2.8$	11: -2.8	$16 \rightarrow 33$
MiniMax-M3	$+15.5$	29: $+16.9$	29: $+2.8$	13: -4.2	$26 \rightarrow 42$
Qwen3.8-flash	$+19.7$	28: $+19.7$	40: $+1.4$	3: -1.4	$5 \rightarrow 30$
Pooled	$+16.3$	126: $+15.8$	181: $+4.2$	48: -3.7	$115 \rightarrow 188$

Table 11: **Trajectory replay of the diagnosis turn.** (a) Accuracy (%). The original runs are the blind and video consultations; each replayed diagnosis turn sees the 32 video frames and combines the history of one consultation with the investigations and returned results of another. (b) Paired differences averaged over the five systems, with 95% case-level bootstrap intervals.

(a)

	original run		replayed diagnosis turn			
history from investigations from system	blind	video	video video VV	video blind VB	blind video BV	blind blind BB
GPT-5.6-luna	23.9	46.5	42.3	31.0	39.4	32.4
Gemma-4-31B	23.9	33.8	35.2	28.2	31.0	28.2
MiMo-v2.5	8.5	22.5	18.3	8.5	26.8	8.5
MiniMax-M3	9.9	25.4	28.2	12.7	26.8	11.3
Qwen3.8-flash	2.8	22.5	22.5	5.6	21.1	4.2
Mean	**13.8**	**30.1**	**29.3**	**17.2**	**29.0**	**16.9**

(b)

path	contrast	Δ accuracy (pp)
Investigations	VV − VB (video history)	$+12.1$ $[+6.5, +17.7]$
	BV − BB (blind history)	$+12.1$ $[+5.9, +18.3]$
History	VV − BV (video investigations)	$+0.3$ $[-2.8, +3.1]$
	VB − BB (blind investigations)	$+0.3$ $[-1.7, +2.5]$
Frames at diagnosis	BB − blind	$+3.1$ $[+0.3, +6.5]$
Replay fidelity	VV − video	-0.8 $[-4.2, +2.5]$

Evidence-return patterns in incorrect consultations. Table 12 partitions incorrect video-condition consultations by the investigation results returned: (i) at least one decisive entry; (ii) other results but no decisive entry; and (iii) no returned result. Group (i) accounts for 29–49% of incorrect consultations, group (ii) for 34–67% and group (iii) for 4–29%. Missing decisive evidence is therefore common, yet in roughly a third to a half of incorrect consultations at least one decisive entry was returned without leading to the diagnosis.

Table 12: **Evidence returned in incorrect video-condition consultations.** Consultations not graded ACCURATE, split by the recorded investigation results into three exclusive categories.

system	incorrect (of 71)	of which (%)		
		decisive entry returned	no decisive entry, some results returned	no results returned
GPT-5.6-luna	38	37	34	29
Gemma-4-31B	47	34	60	6
MiMo-v2.5	55	35	45	20
MiniMax-M3	53	49	45	6
Qwen3.8-flash	55	29	67	4

C.2 EVIDENCE AVAILABLE IN THE RECORDS AND THE SCORING OF τ

Chart composition. Most chart entries are derived rather than reported. Each case chart combines the results its source article reports with the investigation menu of its category. Every case in a category offers the same menu, so that a reasonable test the report omits still returns a result, and the presence of a result does not reveal which tests the report performed; entries a case does not report carry a value expected for its presentation (`derived`), in most cases normal, not performed or not recorded (82% of the $5{,}972$ derived entries). Derived values were fixed before any model was evaluated: category-specific entries were written by the annotators, and a generic panel of commonly ordered tests was filled per case with DeepSeek-V4.1-Flash given the confirmed diagnosis and the reported chart, under the instruction to answer normal unless the diagnosis specifically changes the test. The 45 derived decisive entries are menu tests whose expected result bears on the diagnosis; the annotators flagged them together with the reported ones (44 from the category menus, one from the generic panel). Table 13 gives the counts.

Table 13: **Chart composition** over the 71 cases. Neutral: normal, negative, not performed, not recorded or not tried; specific: any other value.

source of the entry	entries	decisive	specific values
reported in the source article	426	256	394
derived, category menu (annotators)	2,473	44	808
derived, generic panel (DeepSeek-V4.1-Flash)	3,499	1	278
total	6,398	301	1,480

Sensitivity to derived values. Most of the video benefit, and all of the oracle and retrieval effects, survive the removal of derived values. We replaced each of the $1{,}086$ derived entries that hold a specific value with the neutral value a sibling case returns (normal / non-contributory, not performed, not recorded or not tried) and re-ran only the diagnosis turn, holding each consultation's questions, orders and matched entries fixed; because every order is placed before any result is returned, this equals a full re-run of the consultation. Neutral values lower accuracy by about ten points with video and five without, so derived values raise absolute accuracy. Video stays ahead of blind input by 10.1 points (14.9 with the original chart), the oracle reference is unchanged, and retrieval keeps its advantage over video (Table 14). Derived values therefore carry about a third of the video benefit and none of the oracle or retrieval effects.

Table 14: **Neutral-value sensitivity.** Accuracy (%), mean of five systems, with the diagnosis turn re-run on fixed trajectories under the original chart and with every derived specific value neutralised. Differences are pooled over systems with 95% paired case-level bootstrap intervals; all rows use the 71 cases, with an empty reply counted as incorrect.

	original	neutral	neutral − original
Blind	14.4	9.3	−5.1 [−8.2, −2.3]
Video	29.3	19.4	−9.9 [−13.5, −6.2]
Oracle evidence	81.4	80.0	−1.4 [−5.4, +2.5]
Source-clean retrieval	39.7	30.4	−9.3 [−13.5, −5.1]
Video − blind	+14.9	+10.1	−4.8 [−9.3, −0.3]
Retrieval − video	+10.4	+11.0	+0.6 [−4.8, +5.9]

Availability of requested evidence. The records settle few of the questions asked (Table 15). Depending on system and condition, 67–92% of questions are answered *unknown* and 7–52% of orders return nothing. The reference condition has lower unavailability than the video condition for all five systems. Compared with the video condition, the multi-turn protocol reduces order unavailability for GPT-5.6-luna, MiMo-v2.5, and Qwen3.8-flash, but increases it for Gemma-4-31B and MiniMax-M3. These rates depend on both the model's requests and the record: the same 71 records settle between 8% and 33% of the questions and answer between 48% and 93% of the orders.

Answerability of each record. Record sparsity does not drive the main contrasts. We define a case's answerability as the share of all questions asked about it, pooled over the five systems and six conditions of Table 15, that its record settles. Answerability ranges from 5.7% to 28.2% across cases (median 14.7%) and is unrelated to the number of documented history features ($r = -0.02$), so sparsity reflects a mismatch between the questions asked and what reports record rather than short records. On the 35 cases with above-median answerability, video exceeds blind input by 21.7 points, oracle evidence exceeds video by 48.6 points and the reference description adds 13.1 points, against 16.3, 51.8 and 15.8 on all 71 cases; video and single-frame accuracy differ by 3.4 points (Table 16). Video-condition accuracy is the same in both halves (30.3% and 30.0%). The conclusions therefore hold where the record answers more of what is asked.

Acquisition of documented history. In the video condition, models acquire only 15–35% of the positive history findings the records document (Table 17). Every case documents 15–24 history features (median 18), of which 5–13 are recorded as present (median 8); a positive finding counts as acquired when a model question would reveal it, with a judge mapping questions to the symptom table. Video raises acquisition over blind input for every system, and multi-turn interaction raises it again; paired comparisons support both gains. The video gain appears in all 15 system–stratum cells when cases are split by the number of documented positive findings (≤ 7, 8, ≥ 9; $n = 29, 17, 25$). Across the 1,065 consultations, accuracy is 20.1% when fewer than a quarter of the positive findings are acquired, 33.9% at 25–50%, and 30.6% above half.

Scoring of τ. For case c, with decisive entries T_c and the chart entries $E_{\mathrm{acq},c}$ its consultation acquires, deduplicated within the case, $\tau_c = |T_c \cap E_{\mathrm{acq},c}|/|T_c|$. Entries are equally weighted and may support the diagnosis or exclude alternatives. Reported values pool over the 71 cases, $\tau_{\mathrm{pooled}} = \sum_c |T_c \cap E_{\mathrm{acq},c}| / \sum_c |T_c|$. A decisive entry counts as acquired when it, or another entry of the same chart that reports the same finding from the same kind of investigation, is returned; these equivalences are fixed per case before scoring, as described next. τ measures how much of the documented work-up a consultation reproduces; blinded ratings of the order lists are in Appendix C.3.

Equivalent chart entries. Scoring decisive entries by name alone undercounts evidence, so τ credits equivalent entries. Under name-only scoring, 22 of the 107 accurate video-condition consultations obtained no decisive entry, and a verifier review found that in 15 of them the decisive finding had been returned under a second entry for the same test (Table 18); for example, the hummingbird sign came back under "MRI brain with gadolinium" while the decisive entry is "MRI brain – midbrain". Before scoring, we therefore listed for every case the chart entries that report a decisive finding from the same kind of investigation. A judge proposed links from each full chart, and a

Table 15: **Availability of requested evidence by condition.** Per-case means over the usable consultations of the 71 clips; τ as in Table 2. History reports questions and unknown answers; investigations report orders, unavailable returns, distinct entries released, reported or derived (Table 13), and τ. Multi-turn uses up to ten rounds; reference uses audited descriptions.

History (questions asked per case; share answered "unknown", %)

system		blind	single frame	shuffled	video	multi-turn	reference
GPT-5.6-luna	questions	29.8	14.9	17.6	16.2	25.1	18.5
	unknown	92.1	86.5	84.8	85.7	84.8	83.3
Gemma-4-31B	questions	4.7	9.0	8.0	7.9	23.5	6.8
	unknown	82.5	81.1	77.7	77.5	84.4	66.6
MiMo-v2.5	questions	12.4	8.2	7.1	8.2	19.8	9.3
	unknown	89.3	88.7	84.2	84.2	86.4	70.5
MiniMax-M3	questions	19.6	15.9	11.6	11.3	26.3	12.1
	unknown	88.7	88.9	86.1	84.6	86.1	73.8
Qwen3.8-flash	questions	7.9	4.0	6.0	6.3	8.4	7.0
	unknown	91.9	88.3	86.0	87.9	88.4	77.8

Investigations (orders per case; share unavailable, %; entries released per case; τ, %)

system		blind	single frame	shuffled	video	multi-turn	reference
GPT-5.6-luna	orders	8.7	7.9	9.2	8.5	6.8	8.0
	unavailable	42.8	29.0	19.1	35.0	20.8	14.8
	entries	15.4	16.9	21.2	16.7	13.0	22.3
	τ	19.3	28.6	32.9	28.9	39.2	50.8
Gemma-4-31B	orders	3.9	4.1	3.7	3.5	4.9	2.6
	unavailable	13.6	23.3	16.2	17.8	27.1	6.7
	entries	17.6	10.5	8.0	8.2	6.2	6.9
	τ	22.3	17.3	15.0	20.3	17.6	19.9
MiMo-v2.5	orders	6.8	6.5	5.5	6.2	5.1	5.6
	unavailable	52.2	31.6	23.8	25.9	22.0	15.3
	entries	8.9	9.0	8.2	9.0	5.5	9.2
	τ	8.3	16.9	18.9	16.3	14.0	22.3
MiniMax-M3	orders	9.2	12.6	9.7	7.7	7.2	9.0
	unavailable	36.7	42.2	31.0	27.3	32.1	22.8
	entries	12.0	8.3	9.7	8.4	7.3	10.7
	τ	13.6	11.0	15.6	19.3	21.6	30.6
Qwen3.8-flash	orders	3.5	3.7	4.8	5.3	2.4	2.8
	unavailable	37.6	34.5	23.4	21.9	11.8	13.3
	entries	4.0	5.9	6.1	7.9	4.8	4.5
	τ	2.0	16.3	13.6	16.3	10.6	17.9

Table 16: **Main contrasts by record answerability.** Answerability is the share of a case's questions, pooled over five systems and six conditions, that its record settles (median 14.7%). Stage-2 accuracy (%) averaged over systems; differences in points with 95% paired case-level bootstrap intervals.

cases	n	video	video − blind	oracle − video	reference − video	video − single	video − shuffled
All cases	71	30.1	+16.3 [+9.3, +23.7]	+51.8 [+43.4, +60.0]	+15.8 [+8.7, +22.8]	+5.9 [−0.3, +12.4]	+5.4 [−0.3, +11.3]
Above-median answerability	35	30.3	+21.7 [+13.1, +30.9]	+48.6 [+36.6, +61.1]	+13.1 [+2.9, +23.4]	+3.4 [−5.1, +12.0]	+3.4 [−5.1, +12.0]
At or below median	36	30.0	+11.1 [+0.0, +22.2]	+55.0 [+43.9, +66.1]	+18.3 [+8.9, +27.8]	+8.3 [+0.0, +17.2]	+7.2 [−0.6, +14.4]
Top third	24	31.7	+22.5 [+11.7, +34.2]	+49.2 [+33.3, +65.0]	+15.8 [+3.3, +28.3]	+0.0 [−10.0, +10.0]	+3.3 [−7.5, +15.0]

Table 17: **Acquisition of documented history**: share (%) of the findings each record documents as present that the consultation's questions would have revealed, pooled over 71 cases, with 95% case-level bootstrap intervals; differences are paired over cases.

system	blind	video	multi-turn	video − blind	multi-turn − video
GPT-5.6-luna	19.0	34.8	43.6	+15.9 [+8.9, +23.2]	+8.8 [+1.4, +16.3]
Gemma-4-31B	7.9	19.5	31.6	+11.6 [+6.0, +17.4]	+12.1 [+4.7, +19.3]
MiMo-v2.5	14.7	23.8	38.8	+9.1 [+1.6, +16.7]	+15.0 [+6.3, +23.8]
MiniMax-M3	11.9	21.2	37.6	+9.3 [+4.2, +14.7]	+16.4 [+9.4, +23.2]
Qwen3.8-flash	7.4	14.8	20.9	+7.4 [+2.5, +12.4]	+6.0 [+0.3, +11.8]

second, pairwise pass kept 90 of 255 proposals, rejecting normal, less specific, merely compatible and cross-modality results. The links recover 12 of the 15 audited duplicates and none of the other audited consultations. Under this scoring 10 accurate consultations retain $\tau = 0$, τ rises by 0.4–7.0 points across the conditions of Table 2, and the main-text paired contrasts keep their direction.

Table 18: **Accurate video-condition consultations with name-only** $\tau = 0$ (22 of 107), classified by a verifier review of the orders, returned results and source report.

category	n	example
Decisive finding returned under a second entry for the same test	15	PSP: hummingbird sign under "MRI brain with gadolinium"
Different but reasonable diagnostic route	3	mucormycosis: sinonasal–orbital disease on MRI rather than CT
Decisive order lost to a matching miss	1	"MRI Brain" not matched to the pontine-infarct MRI
No supporting result; diagnosis from video and history	3	post-hypoxic myoclonus without the cardiac-arrest history

C.3 Case-Specific Investigation Selection

Model-selected work-ups beat a random list of the same length but not a fixed ten-item checklist, whose accuracy and decisive-evidence coverage are close to model selection. All arms use the same case records; accuracy, τ and blinded list ratings capture different properties of the work-up.

Experimental controls. We replay all 71 clips on GPT-5.6-luna, keeping the video, history questions and patient responses fixed. Only the investigation turn changes, followed by diagnosis using the returned evidence. The *budget* arm allows ten atomic tests, with one test per line and no bundled or generic requests such as "routine bloods". The *checklist* arms submit the same fixed list of ten or fifteen items for every case. A *random* arm submits ten entry names sampled uniformly from the pooled chart vocabulary. Atomic wording fixes the request, not the release: one free-text request can still match several chart entries (the released video arms return 1.1 to 2.4 entries per order, and the budget arm's ten atomic requests release 1.53 entries per answered order). Two *atomic checklist* arms therefore hold the ten-item checklist to one chart entry per order. In the *named* variant, each item is resolved to a single named chart entry of the case and only that entry is released; an item with no such entry returns "not performed / not available". In the *matched* variant, the checklist is submitted as written and the matcher releases at most one entry per order, the entry that most exactly

names the test. A third single-entry arm resubmits the budget arm's own ten orders under the same rule, so both sides use one release rule. All arms use the same matcher and case-specific chart.

The ten-item checklist comprises tone; power; eye movements; full blood count; electrolytes and renal function; liver function; glucose and HbA1c; thyroid function; vitamin B12; and contrast brain MRI. The fifteen-item list adds gait; copper and caeruloplasmin; EEG; nerve conduction studies with electromyography; and lumbar puncture.

Table 19: **Order-selection controls** on GPT-5.6-luna across 71 clips. Orders denotes mean orders per clip; unav. is the share of orders returning nothing. Entries denotes mean distinct chart entries released per clip, deduplicated across orders within each case (entries divided by orders gives the entries released per order). Brackets show 95% case-level bootstrap intervals. Δ is the paired difference in accuracy relative to the released free-text arm, which is the video arm of Table 2. τ as in Section 3.3. The two atomic checklists hold the ten-item checklist to one chart entry per order: *named* resolves each item to a single named chart entry of the case, *matched* submits the checklist as written and lets the matcher release at most one entry per order. *Budget, single entry* releases the budget arm's own orders under the matched rule. Blinded ratings of these lists are in Table 20.

Arm	Orders	Unav. (%)	Entries	ACCURATE (%)	Δ (pp)	τ (%)
Released (free)	8.5	35.0	16.7	46.5 [35.2, 57.7]	—	28.9 [22.1, 35.9]
Budget, 10 tests	10	25.1	10.0	43.7 [32.4, 54.9]	−2.8 [−12.7, 7.0]	25.9 [20.1, 31.9]
Budget, single entry	10	29.2	6.5	36.6 [25.4, 47.9]	−9.9 [−19.7, 0.0]	20.9 [16.2, 26.0]
Checklist, 10	10	3.8	12.4	42.3 [31.0, 53.5]	−4.2 [−14.1, 5.6]	22.9 [18.0, 28.0]
Checklist, 15	15	8.5	20.8	40.8 [29.6, 52.1]	−5.6 [−15.5, 4.2]	31.9 [26.4, 37.4]
Atomic, named	10	14.1	8.6	32.4 [21.1, 43.7]	−14.1 [−23.9, −4.2]	14.6 [10.3, 19.2]
Atomic, matched	10	5.2	9.5	35.2 [23.9, 46.5]	−11.3 [−21.1, −1.4]	15.9 [12.1, 19.9]
Random, 10	10	34.9	8.0	25.4 [15.5, 35.2]	−21.1 [−33.8, −8.5]	7.3 [4.3, 10.5]

Effect of an atomic-test budget. The itemised budget reduces the entries released from 16.7 per clip over 8.5 orders to 10.0 distinct entries over 10 orders (532 of 710 orders answered, 1.53 entries per answered order before deduplication), while τ is 25.9% against 28.9% in the released arm (Table 19). Accuracy is similar: 43.7% with the atomic budget and 46.5% in the released arm. The observed decisive-evidence yield persists when requests are itemised.

Comparison of ten-item strategies. The model-selected budget, fixed checklist and random list each issue ten items and reach 43.7%, 42.3% and 25.4% accuracy, with τ of 25.9%, 22.9% and 7.3% (Table 19). Model selection and the checklist exceed random selection by 18.3 and 16.9 points but differ from each other by only 1.4 points in accuracy. The fifteen-item checklist raises τ to 31.9% while accuracy is 40.8%. Availability differs across arms: 3.8% of checklist orders return nothing, versus 25.1% of model-selected and 34.9% of random orders. The checklist returns more results, and its decisive-evidence coverage is close to model selection at the ten-item budget (a 3.0-point difference whose interval includes zero).

Comparison at a matched ten-order budget. The ten-item arms above match list length but release 10.0, 12.4 and 8.0 entries per clip. We therefore limit each order to at most one chart entry for both the model-selected list and two checklist variants. The named checklist returns nothing for 14.1% of orders, partly because full blood count has no entry in 38 of 71 charts; the matched variant returns nothing for 5.2%. Under the symmetric release rule, the model-selected list reaches 36.6% accuracy and 20.9% τ, versus 35.2% and 15.9% for the matched checklist (Table 19). Neither accuracy nor decisive-evidence coverage separates the lists clearly (5.0 points of τ in favour of model

selection, interval including zero). Restricting the model's own orders to one entry reduces its accuracy by 7.0 points and τ by 5.0 points, showing that multi-entry release contributes to coverage. The original checklist's decisive hits came mostly from the eye-movement, contrast-MRI and glucose items. Blinded raters see the same ten orders in both model arms and give similar overall scores to the structured lists, while counting more unnecessary orders in the model's list (Table 20).

Blinded rating of the order lists. Two blinded raters scored the five ten-item arms on all 71 clips, seeing the video, visible sign, history exchange and one list at a time, without the chart, reference diagnosis or other arms (Table 20). Both place the random list last (1.35 and 1.85 overall) and the four structured lists together at 2.7–2.8. They count 3.4–3.6 unnecessary orders per ten in the model-selected list, versus 2.6–3.1 in the checklists and 5.8–7.7 in the random list. Across 355 rated lists, accuracy is higher when a decisive test is obtained (58.3% vs. 13.9%; n=175 and 180). Like the raters, τ separates random from structured lists, and a decisive test tracks accuracy.

Table 20: **Blinded rating of the ten-item order lists** on all 71 clips. Paired values report Rater 1 / Rater 2; scores and unnecessary-order counts are means; τ is computed on the same clips.

Strategy	Overall score (1–5)	Orders judged unnecessary	τ (%)
Random, 10	1.35 / 1.85	7.7 / 5.8	7.3
Checklist, 10	2.82 / 2.67	2.6 / 2.9	22.9
Atomic, named	2.80 / 2.79	2.6 / 3.1	14.6
Atomic, matched	2.76 / 2.77	2.8 / 3.1	15.9
Model-selected, 10	2.82 / 2.69	3.4 / 3.6	25.9

D Temporal-Window Training: A Paired Analysis

Post-training improves the small describer; where its window is placed matters little, but a short, densely sampled segment appears more useful than the same budget spread over the clip (Section 4.3). Below we give the teacher's windows, the adapter versus the untrained model, a blinded rating of the sign sentences, paired window-minus-random contrasts, and a sign-only adapter trained without windows. Each clip has three downstream runs per arm; comparisons use clip-level means and 10,000 paired bootstrap resamples.

Window statistics. The teacher was run on all 71 clips and returned a usable record for 68; it judged the 32-frame survey sufficient for 29 of these. The implied sampling rate within a window is 0.7, 1.4 and 2.9 frames per second at $K = 8, 16, 32$.

The untrained model on identical inputs. Table 23 evaluates unadapted Qwen3.5-4B on the same survey-plus-window input as the adapted student. Unadapted accuracy is 22–27% across budgets. The adapted student leads by 5.2, 8.0 and 6.6 points at $K = 8, 16, 32$, and its coverage is higher by 3.9, 5.2 and 5.1 points, with paired intervals excluding zero at every budget; accuracy point estimates also favour the adapter.

Independent recognition of the student's sentences. A blinded judge compared the sentences produced by the window-trained student and by the unadapted model against the reference phenomenology, component by component (Table 22). The adapted student named the body part correctly in 66% of clips (unadapted: 40%) and the movement character in 54% (unadapted: 10%), and its sentence was judged the closer one in 48 clips against 14, with 6 ties. Laterality is rarely correct in either model (9% versus 5% of the 57 clips whose reference states a side), and the activation condition is named correctly less often after training (18% versus 38%). Sign recognition therefore improves with training on the two components the training target emphasises, assessed independently of the consultation.

Paired contrasts for window position. The paired differences (Table 21) remain small across $K = 8, 16, 32$: accuracy changes by -0.5 to $+3.3$ points and source-workup coverage by at most 1.6 points. The data thus support the adapter effect more clearly than the choice of window.

A sign-only adapter without windows. To separate post-training from the window, we trained a second adapter with the same 68 clips, source-grouped folds, target sentence and prompt, for one epoch per fold, on $32{+}K$ frames spaced evenly over the whole clip, so that no window enters training or inference. Table 24 compares it with the window-trained adapter at the same frame budget. Its coverage matches the window-trained adapter at every budget (18.3–19.0% versus 17.3–19.4%) and exceeds the untrained model by 4.7–5.6 points. Its accuracy (26.8–27.2%) stays within 5.2 points of the untrained model and trails the window arm by 2.3, 2.8 and 6.6 points at $K = 8, 16, 32$; the random-window arm also leads it at every budget, by 2.8, 1.4 and 3.3 points. Learning to describe the sign therefore accounts for the coverage gain, whereas the accuracy lead of both windowed arms suggests that a short, densely sampled segment is more useful than the same budget spread over the clip; these differences remain within sampling error.

Table 21: **Paired window-minus-random differences** (percentage points; 95% paired case-level bootstrap intervals, 71 clips). W: window arm; R: random window of the same length. Both arms share the adapter and the 32-frame survey, as in Table 3.

K	Δ accuracy (pp)	$\Delta\tau$ (pp)
8	-0.5 $[-7.5, +6.6]$	$+0.9$ $[-2.2, +4.0]$
16	$+1.4$ $[-7.0, +9.4]$	$+0.7$ $[-1.9, +3.3]$
32	$+3.3$ $[-4.2, +10.8]$	-1.6 $[-5.0, +1.8]$

Table 22: **Blinded judgement of the student's sign sentences**, $K{=}32$ with the 32-frame survey. For each clip the judge saw the reference phenomenology and the two sentences and marked each component independently; unstated components are not scored, so denominators differ.

sentence	body part	character	side	condition	judged closer (clips)
no adapter	27/68 (40%)	7/68 (10%)	3/57 (5%)	13/34 (38%)	14
window-trained adapter	45/68 (66%)	37/68 (54%)	5/57 (9%)	6/34 (18%)	48

Table 23: **Untrained Qwen3.5-4B versus the window-trained student** on the same survey + window input (32-frame survey, same day, denominator 71, three runs per clip, clip-level bootstrap). Coverage is source-workup coverage (τ); differences are paired over clips.

	untrained		adapter		adapter − untrained	
K	acc. [95% CI]	cov.	acc. [95% CI]	cov.	Δacc. [95% CI]	Δcov. [95% CI]
8	23.9 [16.4, 31.9]	13.4	29.1 [20.7, 38.0]	17.3	$+5.2$ $[-3.8, +14.1]$	$+\mathbf{3.9}$ $[+0.1, +7.9]$
16	22.1 [14.6, 30.5]	13.4	30.0 [21.1, 39.4]	18.6	$+8.0$ $[-1.4, +17.4]$	$+\mathbf{5.2}$ $[+1.5, +9.2]$
32	27.2 [18.3, 36.6]	14.3	33.8 [24.4, 43.7]	19.4	$+6.6$ $[-1.9, +15.5]$	$+\mathbf{5.1}$ $[+0.6, +9.8]$

Table 24: **Sign-only adapter trained without windows** versus the window-trained adapter at matched frame budget (denominator 71, three runs per clip, clip-level paired bootstrap). No window: adapter trained and evaluated on $32+K$ evenly spaced frames. Window / random: the window-trained adapter on the 32-frame survey plus K frames from the student's predicted window or from a random window of the same length. Untrained: Qwen3.5-4B on the survey-plus-window input. Coverage is source-workup coverage (τ); differences are no window minus the listed arm.

K	arm	acc. [95% CI]	cov.	Δacc. [95% CI]	Δcov. [95% CI]
8	no window	26.8 [19.2, 34.3]	19.0	–	–
	window	29.1 [20.7, 38.0]	17.3	−2.3 [−10.8, +5.6]	+1.8 [−1.0, +4.6]
	random	29.6 [21.6, 38.5]	16.4	−2.8 [−10.8, +4.7]	+2.7 [−0.3, +5.8]
	untrained	23.9 [16.4, 31.9]	13.4	+2.8 [−5.2, +10.3]	**+5.6** [+2.1, +9.5]
16	no window	27.2 [19.2, 35.7]	18.3	–	–
	window	30.0 [21.1, 39.4]	18.6	−2.8 [−11.7, +6.1]	−0.3 [−3.1, +2.4]
	random	28.6 [20.2, 37.6]	17.9	−1.4 [−8.9, +6.1]	+0.3 [−2.6, +3.3]
	untrained	22.1 [14.6, 30.5]	13.4	+5.2 [−2.3, +12.7]	**+4.9** [+1.0, +8.9]
32	no window	27.2 [19.2, 35.7]	18.9	–	–
	window	33.8 [24.4, 43.7]	19.4	−6.6 [−13.6, +0.5]	−0.4 [−3.7, +2.9]
	random	30.5 [21.6, 39.9]	20.9	−3.3 [−10.3, +3.8]	−2.0 [−5.4, +1.4]
	untrained	27.2 [18.3, 36.6]	14.3	0.0 [−8.0, +8.5]	+4.7 [−0.5, +9.8]

E Recognition across Frame Budgets and Clinical Categories

Two analyses illustrate Section 4.1: recognizing a sign rarely places its cause in the differential, and in a worked example more frames do not make the differential more focused.

E.1 A Frame-Budget Worked Example

We trace how the frame budget changes GPT-5.6-luna's description and initial differential for PMC9815763, a woman with right-arm chorea from Graves' hyperthyroidism. The reference phenomenology is *continuous, rapid, irregular movements of one arm*; the reference aetiology is hyperthyroidism. Figure 9 shows the input at three budgets, and Table 25 compares responses at six.



Figure 9: **One clip at three frame budgets.** Uniformly sampled frames from the same 17.8-second clip at $K=1$, 4 and 8. Frames share a common scale; each row shows the full sampling budget.

Recognition across frame budgets. At $K=1$, the model interprets a static appearance as swelling; at $K=2$, it interprets the postural change as pain. At $K=4$, it notices irregular arm movement but emphasises posture, receiving partial credit. At $K=8$, it describes irregular, flowing, non-rhythmic movement. The $K=32$ answer also receives a correct grade; at $K=128$ the description misses the continuous, one-sided character and is graded partial.

Recognition does not necessarily narrow the differential. From $K=8$ onward, the model lists five to eight candidates spanning functional, drug-induced, seizure-related and movement-syndrome explanations. At higher budgets, functional movement disorder often leads the list despite correct or partial phenomenology grades, with variability cited in support. In this example, correct sign recognition can coexist with a broad differential that additional frames do not consistently narrow.

Table 25: **Responses at six frame budgets.** Phenomenology recognition and differential diagnosis, quoted from the responses. *Sign grade* is the grader's rating of phenomenology recognition.

K	Phenomenology recognition	Differential diagnosis	Sign grade
1	"lying on their side … a prominent dark/purplish area with swelling around the ear/temple and upper cheek"	acute trauma to the head; periorbital or auricular haematoma	wrong
2	"repeatedly bend forward and hold/rub the upper abdomen or lower chest"	upper abdominal pain; gastritis, ulcer or biliary disease	wrong
4	"markedly bent forward/stooped … intermittent, somewhat irregular/flinging or jerky movement of the opposite arm"	camptocormia from Parkinson disease or another parkinsonian syndrome	partial
8	"the movements look irregular, flowing, and somewhat twisting rather than rhythmic"	hyperkinetic movement disorder; chorea or choreoathetosis, possibly hemichorea	correct
32	"irregular, seemingly involuntary movements … jerky and variable, with prominent upper-limb and whole-body involvement"	functional movement disorder; "the marked variability … support[s] this over a fixed neurologic syndrome"	correct
128	"recurrent, forceful, involuntary movements of the trunk and upper limbs … episodic and somewhat irregular"	functional movement disorder, particularly functional truncal jerks or dystonia	partial

E.2 Recognition and Aetiological Coverage by Category

Figure 10 compares phenomenology recognition with aetiological coverage by clinical category, pooling five models and eight frame budgets; categories are ordered by the gap between them.

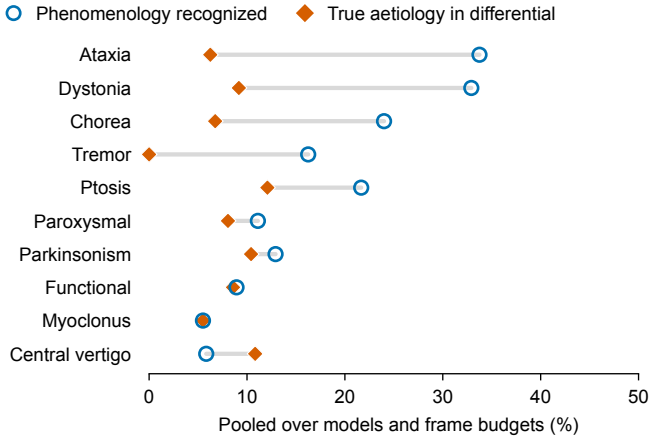


Figure 10: **Recognition and aetiological coverage by category.** Phenomenology recognition and aetiological coverage pooled over five models and eight frame budgets and ordered by their difference. Facial palsy is excluded because it contains only one clip.

Recognition exceeds coverage most clearly in ataxia (33.8% versus 6.2%), dystonia (32.9% versus 9.2%), chorea (24.0% versus 6.8%) and tremor (16.2% versus 0.0%). In these categories naming a visible sign does not translate into a compatible aetiological hypothesis, consistent with recognition and hypothesis formation being distinct hurdles (Section 4.1).

The two measures are close in paroxysmal events (11.1% versus 8.1%), parkinsonism (12.9% versus 10.4%), functional movement disorder (8.9% versus 8.6%) and myoclonus (5.5% for both). Only central vertigo reverses the pattern (5.8% recognition versus 10.8% coverage), where a broad differential can include a compatible candidate despite an inaccurate description. Neither measure tracks the other across categories, so the two are evaluated separately.

F RETRIEVAL: CANDIDATE COVERAGE AND A CONSULTATION EXAMPLE

This appendix supports Section 4.4. Retrieval supplies candidate causes the model misses, decontamination removes source articles without lowering mean accuracy, the work-up moves toward the documented one for four of five systems, and an equal-length list of unrelated causes reproduces much of the accuracy effect. A worked consultation closes the section. Figure 11 reports the questions asked and investigations ordered per case under each condition.

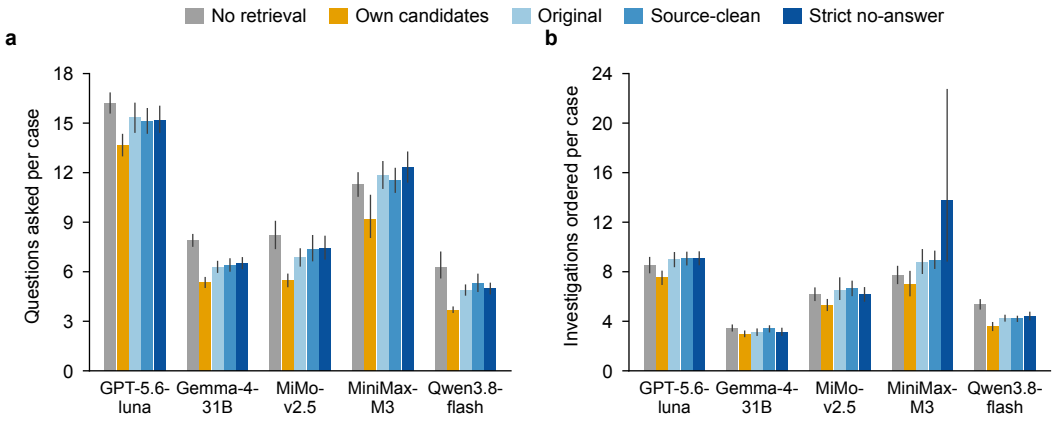


Figure 11: **Interaction counts under retrieval conditions** (71 cases): (**a**) questions asked and (**b**) investigations ordered per case. Colours denote no retrieval (grey), Own candidates (orange), Original (light blue), Source-clean (blue), and Strict no-answer (dark blue). Bars show case-level means with 95% bootstrap intervals.

Complementarity of candidate lists. The retrieved list recovers the true cause far more often than the model's video-only differential, and the two cover partly different cases. Across five models the retrieved list covers 57.7–62.0% of clips and the model alone 7.0–9.9%; their union covers 62.0–64.8% (Figure 12). The model's own hypotheses still add 1.4–5.6 points that retrieval misses, so we supply the combined list rather than replacing the model's hypotheses with retrieved candidates.

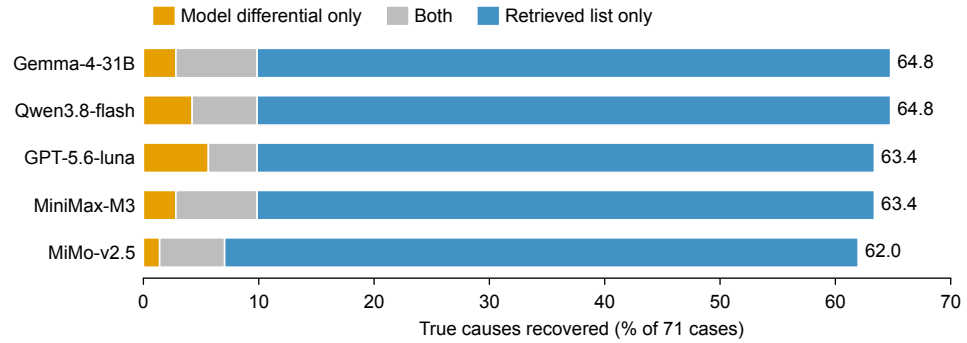


Figure 12: **The retrieved list covers causes the model misses.** Coverage of reference causes by the model's initial differential and the Source-clean retrieved list. Stacked segments distinguish causes covered only by the model, by both lists, or only by retrieval; the total is their union.

Decontamination. Source-clean filtering removed the source article for 24/71 cases; shared-DOI and near-duplicate-title rules removed no additional records. Strict no-answer filtering removed 999 records across 41 cases whose titles or excerpts name the confirmed diagnosis or an accepted equivalent. Table 27 gives candidate coverage and accuracy under the three filtering levels. Source-clean filtering lowers the coverage of the supplied list from 68.4 to 63.7% and changes accuracy by −4.2 to +7.0 points; Strict no-answer filtering lowers it by a further 13.3 points, to 50.4%.

Table 26: **Retrieval conditions minus the video condition** on the same 71 clips (percentage points, 95% case-level bootstrap intervals). Accuracy and coverage τ as in Table 2.

System	Condition	Δ accuracy [95% CI]	$\Delta\tau$ [95% CI]
GPT-5.6-luna	Source-clean	$+2.8$ $[-9.9, +15.5]$	$+8.3$ $[+0.9, +15.6]$
	Strict no-answer	$+1.4$ $[-11.3, +14.1]$	$+5.3$ $[-3.0, +13.4]$
Gemma-4-31B	Source-clean	-2.8 $[-14.1, +8.5]$	$+0.7$ $[-5.7, +6.5]$
	Strict no-answer	-4.2 $[-15.5, +7.0]$	-2.0 $[-8.3, +3.9]$
MiMo-v2.5	Source-clean	$+21.1$ $[+8.5, +33.8]$	$+13.3$ $[+5.0, +21.3]$
	Strict no-answer	$+15.5$ $[+2.8, +28.2]$	$+11.0$ $[+4.7, +17.3]$
MiniMax-M3	Source-clean	$+15.5$ $[+1.4, +28.2]$	$+8.0$ $[+2.0, +14.4]$
	Strict no-answer	$+23.9$ $[+9.9, +36.6]$	$+13.6$ $[+6.7, +20.7]$
Qwen3.8-flash	Source-clean	$+11.3$ $[0.0, +23.9]$	$+9.0$ $[+2.0, +15.6]$
	Strict no-answer	$+11.3$ $[0.0, +21.1]$	$+7.0$ $[+0.3, +13.7]$

We audited the 31,749 distinct records retained after source-clean filtering for other reports of the same patient. Of these, 144 records in 35 cases share at least one author with the source article; the 43 that also share two or more diagnosis words describe other patients or reviews from the same groups. An author-independent check found 7 records whose abstract describes a patient of the case's age and sex and shares a rare diagnosis term; each differs from the source patient in country, year or presentation, and no record meets both criteria. The audit found no same-patient report among the retained abstracts; reports identifiable only from full text remain possible.

Table 27: **Candidate coverage and accuracy under the three filtering levels** (71 cases). Coverage is the share of cases whose supplied list (the model's own Stage 1 differential followed by the retrieved causes) contains the true cause, graded with the Appendix H rank probe; accuracy is the ACCURATE rate of the consultation.

	candidate coverage (%)			accuracy (%)		
system	original	source-clean	strict	original	source-clean	strict
GPT-5.6-luna	70.4	63.4	52.1	50.7	49.3	47.9
Gemma-4-31B	70.4	64.8	52.1	29.6	31.0	29.6
MiMo-v2.5	64.8	62.0	47.9	36.6	43.7	38.0
MiniMax-M3	69.0	63.4	47.9	45.1	40.8	49.3
Qwen3.8-flash	67.6	64.8	52.1	33.8	33.8	33.8
Mean	68.4	63.7	50.4	39.2	39.7	39.7

Paired differences from the video condition. Source-clean retrieval raises τ with an interval above zero for four of five systems, and accuracy for MiMo-v2.5 and MiniMax-M3; the Qwen3.8-flash accuracy interval reaches zero (Table 26). Under the strict audit, τ still rises with an interval above zero for three systems, and the accuracy gains of MiMo-v2.5 and MiniMax-M3 remain.

Blinded order-list ratings. Table 28 adds the blinded ratings of Appendix C to the accuracy and source-workup outcomes. For MiMo-v2.5, MiniMax-M3, and Qwen3.8-flash, retrieval improves these measures together; unnecessary-order shares also fall for MiMo-v2.5 and Qwen3.8-flash. The matched-list control below tests how much of the gain depends on retrieved content.

Length- and form-matched control. The mismatched condition appends causes retrieved for a donor case from another category. A fixed per-clip seed and truncation to the source-clean list length match the form and number of candidates in all 71 cases. Own candidates alone lowers accuracy below the video condition for Gemma-4-31B, MiniMax-M3 and Qwen3.8-flash, and appending mismatched causes raises it again; for GPT-5.6-luna neither appended list changes accuracy reliably. Table 29 shows a content-specific accuracy gain for MiMo-v2.5, and higher τ with source-clean than with mismatched candidates for MiMo-v2.5 and Qwen3.8-flash, with paired intervals excluding zero; unrelated candidates reproduce part of the accuracy gains in MiniMax-M3 and Qwen3.8-flash.

Table 28: **Retrieval under the full set of measures** (71 cases). Accuracy and τ as in Table 2; orders is the mean per case. Quality (1–5) and unnecessary share (unnecessary orders over orders placed) are blinded ratings of each order list, reported as Rater 1 / Rater 2. Own: the model's own Stage 1 differential without retrieval; clean: the same list followed by source-clean retrieved causes.

system	arm	acc. (%)	τ (%)	orders	quality	unnecessary (%)
GPT-5.6-luna	video	46.5	28.9	8.5	3.38 / 3.10	22 / 25
	own	52.1	37.9	7.5	3.56 / 3.20	19 / 24
	clean	49.3	37.2	9.1	3.35 / 3.22	24 / 25
Gemma-4-31B	video	33.8	20.3	3.5	2.67 / 2.46	24 / 31
	own	18.3	12.0	3.0	2.43 / 2.46	32 / 38
	clean	31.0	20.9	3.4	2.76 / 2.52	20 / 36
MiMo-v2.5	video	22.5	16.3	6.2	2.63 / 2.46	37 / 42
	own	28.2	15.3	5.3	2.72 / 2.49	32 / 41
	clean	43.7	29.6	6.7	3.00 / 2.89	24 / 26
MiniMax-M3	video	25.4	19.3	7.7	2.51 / 2.41	38 / 44
	own	18.3	17.9	7.0	2.44 / 2.31	46 / 48
	clean	40.8	27.2	9.0	2.61 / 2.69	40 / 37
Qwen3.8-flash	video	22.5	16.3	5.3	2.58 / 2.32	32 / 42
	own	18.3	12.6	3.6	2.69 / 2.68	27 / 33
	clean	33.8	25.2	4.2	3.34 / 2.94	15 / 27

Table 29: **Length- and form-matched control for retrieval** (71 cases). Cells give accuracy / τ (%). Own: the model's own Stage 1 differential; mismatched: the same list followed by source-clean causes retrieved for an unrelated case, truncated to the length of the source-clean list; clean: source-clean retrieval. Differences are in accuracy percentage points with 95% paired case-level bootstrap intervals. The paired τ difference, clean − mismatched, is -0.7 $[-7.4, +6.2]$ for GPT-5.6-luna, $+3.7$ $[-3.9, +11.0]$ for Gemma-4-31B, $+13.0$ $[+6.7, +19.6]$ for MiMo-v2.5, $+4.3$ $[-1.6, +10.2]$ for MiniMax-M3 and $+9.3$ $[+4.3, +14.5]$ for Qwen3.8-flash.

system	video	own	mismatched	clean	clean − mismatched	mismatched − own
GPT-5.6-luna	46.5 / 28.9	52.1 / 37.9	53.5 / 37.9	49.3 / 37.2	-4.2 $[-15.5, +7.0]$	$+1.4$ $[-7.0, +9.9]$
Gemma-4-31B	33.8 / 20.3	18.3 / 12.0	22.5 / 17.3	31.0 / 20.9	$+8.5$ $[-2.8, +19.7]$	$+4.2$ $[-5.6, +14.1]$
MiMo-v2.5	22.5 / 16.3	28.2 / 15.3	22.5 / 16.6	43.7 / 29.6	$+21.1$ $[+9.9, +32.4]$	-5.6 $[-16.9, +5.6]$
MiniMax-M3	25.4 / 19.3	18.3 / 17.9	29.6 / 22.9	40.8 / 27.2	$+11.3$ $[-1.4, +23.9]$	$+11.3$ $[0.0, +22.5]$
Qwen3.8-flash	22.5 / 16.3	18.3 / 12.6	31.0 / 15.9	33.8 / 25.2	$+2.8$ $[-7.0, +12.7]$	$+12.7$ $[+2.8, +22.5]$

Consultation example. A shorter, more targeted order list can recover more decisive evidence. In the illustrative myoclonus case, both runs ask 18 history questions. With retrieval, the model orders 12 rather than 18 investigations, receives 53 rather than 36 chart entries, and obtains two of four decisive entries rather than none. It drops cultures, toxicology, serum ammonia, and head CT while adding nerve-conduction studies with electromyography and a formal oculomotor assessment.

In the retrieval condition, the returned decisive findings include rhythmic EMG bursts synchronised with the visible jerks and an EEG report documenting no EEG–EMG correlation. The final answer is *post-infectious subcortical myoclonus with cerebellar ataxia*, graded ACCURATE; without retrieval, it is *functional neurological symptom disorder*, graded NOT ACCURATE.

G Robustness to Corrupted History Responses

Histories can be wrong, including withheld or misreported information (Levy et al., 2018). We flip a fraction of the structured history responses, keep the video and questions fixed, and measure the change in investigation selection and final diagnosis. At 80% corruption accuracy falls for every model, and more orders do not compensate.

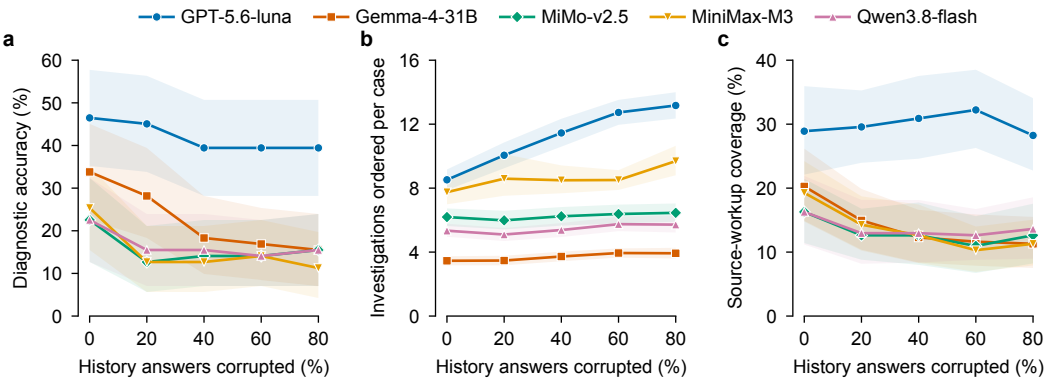


Figure 13: **Consultation outcomes across history-corruption levels**: (**a**) diagnostic accuracy, (**b**) investigation count, and (**c**) source-workup coverage. The video and history questions are held fixed; investigation orders and final diagnoses are generated using the modified responses. Shaded bands are marginal 95% bootstrap intervals over clips.

G.1 Experimental Protocol

For each consultation, we retain the video and history questions from the uncorrupted run and flip a specified fraction of the structured responses (`yes` to `no`; `no` or `unknown` to `yes`). The model then orders investigations and diagnoses from the modified history. Requested results come from the unchanged source-case chart.

Fixing the questions isolates the downstream response to altered history and precludes follow-up questions that might resolve inconsistencies. Each corruption level is compared with its uncorrupted baseline on accuracy, investigation count and τ.

G.2 Aggregate Results

At 80% corruption, diagnostic accuracy is lower than the uncorrupted baseline in all five models (Figure 13). Gemma-4-31B and MiniMax-M3 decline by 18.3 and 14.1 percentage points, respectively, alongside reductions in source-workup coverage.

Changes in workup size vary across models. GPT-5.6-luna increases its mean investigation count from 8.5 to 13.2 while maintaining approximately stable source-workup coverage. MiniMax-M3 shows a smaller increase in investigation count but lower diagnostic accuracy. Additional orders therefore do not consistently compensate for corrupted history.

G.3 Worked Example: Neuropathic Tremor in Charcot–Marie–Tooth Disease

In case `mild_neuropathic_CMT_PMC11600610`, the video shows postural tremor associated with CMT2A (Mohammed et al., 2024). Four chart entries are decisive: limb power and wasting, pes cavus, nerve conduction studies with electromyography (NCS/EMG), and targeted genetic testing. The source report describes one patient.

The model asks the same 18 questions in both runs. At nominal 40% corruption, seven replies change: one `yes` becomes `no`, one `no` becomes `yes`, and five `unknown` become `yes`. Tables 30 and 31 give the complete question-level responses and source evidence; a compound question returns `unknown` unless the documented evidence settles the whole statement.

The altered history points the model toward an acute cerebral work-up. It drops NCS/EMG, its only decisive hit without corruption, and requests stroke imaging, EEG, and cardiac monitoring. Both runs place 14 orders, but returned chart entries fall from 29 to 19 and decisive-entry recovery from $1/4$ to $0/4$. The diagnosis changes from hereditary axonal CMT2 (ACCURATE at the disease-entity level) to functional movement disorder (NOT ACCURATE).

Table 30: **History questions and source-aligned responses** in the worked example (a 34-year-old man, the sole patient of the source case report). *Basis* keys into Table 31, which quotes the supporting source evidence. A dash means no passage is quoted here, not verified absence. Question 8 is supported by S6. Starred rows are corrupted at the nominal 40% rate.

	Question	Basis	Uncorr.	Corr.
1*	Are you experiencing involuntary shaking, jerking, or flapping movements?	S1	yes	**no**
2*	Did these movements begin suddenly?	—	unknown	**yes**
3*	Are the movements present when your hands are at rest?	S2	no	**yes**
4	Do the movements worsen when you hold your arms out or use your hands?	S2	yes	yes
5	Do both hands move in the same way?	—	unknown	unknown
6*	Have you noticed weakness or clumsiness on one side of your body?	—	unknown	**yes**
7	Have you had numbness or loss of sensation?	S3	yes	yes
8	Have you had difficulty speaking, understanding speech, seeing, or walking?	S6	yes	yes
9	Have you had a severe or unusual headache?	—	unknown	unknown
10	Have you lost awareness or had episodes of staring or unresponsiveness?	—	unknown	unknown
11*	Have you recently had confusion, unusual sleepiness, or a change in your alertness?	—	unknown	**yes**
12	Have you had fever, vomiting, or significant dehydration recently?	—	unknown	unknown
13	Do you have known liver disease?	S4	no	no
14	Do you have known kidney disease?	S4	no	no
15	Do you have diabetes or thyroid disease?	S4†	unknown	unknown
16	Have you started, stopped, or changed any medications recently?	—	unknown	unknown
17*	Do you use alcohol or recreational drugs?	S5†	unknown	**yes**
18*	Have you had similar episodes before?	—	unknown	**yes**

† Settled in part only: S4 excludes diabetes but the report does not address thyroid disease, and S5 excludes illicit drug use but does not quantify alcohol use. Under the compound-question rule of Appendix H.2 the environment returns `unknown`.

Table 31: **Patient-level provenance** for every non-`unknown` response in the worked example. Sentences quote the single-patient source report (Mohammed et al., 2024) (`PMC11600610`, *Journal of Medical Case Reports*, CC BY-NC-ND 4.0); *Loc.* gives its section.

Key	Chart entry	Sentence in the source	Loc.
S1	`hand tremor when holding a posture: yes; the tremor started about six months ago: yes`	"[He] came to our hospital with a complaint of tremors of the hands of 6 months duration." — "The tremor was limited to the hands."	Case pres.
S2	`tremor at rest: no`	"We noted a postural tremor that attenuated when the patient assumed an anatomic position."	Case pres.
S3	`reduced sensation in the feet: yes`	"His sensory examination revealed a mildly decreased vibratory sensation; otherwise, he had no sensory level."	Clin. findings
S4	`known liver disease: no; known kidney disease: no; diabetes: no`	"[T]he patient had no history of diabetes, hypertension, psychiatric, cardiac, renal, liver diseases, or cancers."	Case pres.
S5	`illicit drug use: no`	"He had no history of illicit drug use and denied constitutional symptoms."	Case pres.
S6	`difficulty walking: yes`	Walking difficulty and bilateral distal weakness are documented.	Patient info.

H PROMPTS

The benchmark prompts are reproduced verbatim, with runtime substitutions shown as angle-bracketed placeholders. They are organised by function: model evaluation (§H.1), responses from the patient and chart environment (§H.2), retrieval (§H.3), and grading (§H.4). Prompts to the model under evaluation run on that model, and the teacher prompt of Appendix I on `openai/gpt-5.6-luna`; every other prompt, including the verifier used in Appendix C, runs on `deepseek/deepseek-v4.1-flash` through OpenRouter.

H.1 THE MODEL UNDER EVALUATION

Stage 1, from the video alone. Sent with K frames as interleaved images.

```
You are a doctor. Below are <K> frame(s) sampled in order across a ~<duration>
second video of a patient.

Describe what you see, then give your primary guess of the disease. You may list
several possible diagnoses.
```

Stage 2, turn 1: the history. The opening clause carries the role under test (*You are a neurologist seeing a new patient* or *You are a doctor seeing a new patient*); all reported runs use the latter. The candidate block is present only in the retrieval conditions and their list controls.

```
<role> Above are <K> frames sampled across a <duration>-second video of the
consultation, in order.

<candidate block, retrieval conditions only>You may now take a history, but under one
constraint: **Ask yes/no questions. The environment returns yes, no, or unknown
when the record does not establish the answer**,
and the patient will answer only yes, no, or unknown. Anything you do not ask, you do not
learn.

Ask as many or as few questions as you judge this patient needs. Stop when you
believe another question would not change what you think is wrong. Put them in the
order you would ask them - most informative first.

One question per line, numbered. Nothing else.
```

In the Blind condition the frames are omitted and the first paragraph is replaced by:

```
You are a doctor. A new patient has been referred to you and you have not yet seen
or examined them. You know nothing about them at all.
```

The candidate block, when present, precedes the constraint:

```
A literature search on the signs visible in this video returned the following
reported causes. The list is not guaranteed to contain this patient's cause, and
most entries in it are wrong for this patient:

<candidate causes>
```

Stage 2, turn 2: the investigations.

```
The patient answered:

<numbered questions, each with yes, no, or unknown>

You may now investigate. There is no fixed list to choose from.
Name whatever you would actually order, in your own words:
bedside examination manoeuvres, blood tests, imaging,
electrophysiology, invasive procedures, or a therapeutic trial.

A therapeutic trial is returned only if you name that specific
trial. Asking to "try treatment" returns nothing.

Order as many or as few investigations as you judge this patient
needs. Stop when another test would not change what you think is
wrong. Put the investigations most likely to settle the diagnosis
first.

One investigation per line, numbered. Name the test, not what you
expect it to show. Nothing else.
```

Stage 2, turn 3: the diagnosis.

```
The results are:

<one block per order: the order as written, then the chart entries it covered with
their values, or "not performed / not available">

For orders marked "not performed / not available", no documented
result is available in this environment. Do not infer a result.

Now give your diagnosis. State the single diagnosis you believe is correct - the
disease entity and its cause, as specifically as the evidence allows - then, on
separate lines, up to three alternatives you would still consider.

Begin with "DIAGNOSIS:" followed by the single answer on one line.
```

H.2 THE ENVIRONMENT

Both prompts below align free text with the case record. The patient prompt returns yes/no/unknown responses based only on documented features. The chart prompt returns matching entry names, and the harness retrieves their recorded results. Neither prompt is permitted to invent clinical findings.

The patient. The environment answers each question from the documented symptom table: `yes` for an established finding, `no` for an explicitly absent finding, and `unknown` when the record is insufficient. Omission from the table indicates unreported information, not a negative finding.

```
A doctor asked a patient these yes/no questions:

<numbered questions>

The patient's documented features are:

<the case's symptom table>

Answer each question using only the documented features above.

Use exactly one of the following responses:

- "yes" if the documented features explicitly establish that the complete
  statement is true;
- "no" if the documented features explicitly establish that the complete
  statement is false;
- "unknown" if the documented features do not provide enough information
  to determine whether the complete statement is true or false.

Do not infer unreported features. Absence from the documented feature table
does not imply absence of the clinical finding and must not by itself produce
a "no" response.

For compound questions, return yes or no only when the documented
evidence determines the truth of the complete statement.
Otherwise, return unknown.

Reply with ONLY a JSON object:

{"<question number>": "yes"|"no"|"unknown", ...}
```

The chart.

```
A doctor ordered these investigations. A numbered line may bundle several tests.

<numbered orders>

This chart holds results for exactly these entries:

<the case's investigation menu, therapeutic trials marked [NAMED-ONLY]>

For each numbered order, list the chart entries it covers - matching on what the
test is, not on wording. An order covers an entry only if it genuinely asks for
that test. Entries marked [NAMED-ONLY] are therapeutic trials: list one only if
that specific trial is named, never for a general request to try treatment.

Reply with ONLY a JSON object mapping each order number to a list of chart entry
names, exactly as written above: {"1": ["..."], "2": [], ...}
```

H.3 RETRIEVAL

Filtering. For Source-clean and Strict no-answer retrieval, before candidate-cause extraction, we remove records matching the source PMCID or DOI, or having title-token Jaccard similarity ≥ 0.85 with the source title. Source metadata are fetched from Europe PMC by PMCID. Titles are lower-cased, stripped of non-alphanumeric characters, and tokenised, dropping a fixed list of generic publication terms and function words.

Strict no-answer retrieval additionally checks titles and the first 260 abstract characters for the confirmed diagnosis or case-specific accepted equivalents. The diagnosis is truncated at the first dash to exclude supporting reasoning. Matching uses the same normalisation, discards equivalent strings shorter than four characters, and removes records containing any remaining equivalent as a substring. Because it uses ground truth, it serves only as a leakage control, not as a deployable filter.

Normalisation. The model's description is converted to canonical phenomenology terms before HPO matching. This vocabulary-normalisation step has no access to the diagnosis.

```
Rewrite the clinical description below as a list of standard neurological
phenomenology terms, the kind used as headings in a movement-disorder textbook.

Rules:
- one term per line, nothing else, no numbering, no explanation
- use the canonical noun form: "Gait ataxia", "Chorea", "Resting tremor",
  "Ptosis", "Nystagmus", "Bradykinesia", "Dystonia", "Myoclonus", "Facial palsy",
  "Spasticity"
- include the body region as a separate term when the description gives one
- when the description states a side or a distribution, add it on its own line
  using exactly one of: Unilateral, Bilateral, Left, Right, Generalized, Focal,
  Multifocal, Axial, Proximal, Distal, Alternating laterality
- include only what the description actually states; do not infer a diagnosis
- if the description states no abnormality, output the single line NONE

Description:
<the model's own description of the sign>
```

Extraction. Applied to the titles and abstract openings returned by the Europe PMC query.

```
Below are published papers, all retrieved because they describe the same clinical
sign. Each entry is a title followed by the start of its abstract.

For each entry, name the underlying CAUSE or DISEASE it was about, in four words
or fewer. If the paper is a review that enumerates several causes, list them
separated by " ; ". One entry per line, in the same order, nothing else. Write
SKIP if no cause is named.

<titles and abstract openings>
```

H.4 GRADING

Stage 1: the sign and the differential.

```
You are grading a vision model that was shown a short clinical video of one
patient and asked, with NO history and NO test results, to describe what it saw
and propose possible diagnoses.

Evaluate two separate outcomes:
(1) recognition of the visible phenomenology;
(2) coverage of a diagnostic hypothesis compatible with the reference case.

Diagnostic coverage does NOT require exact identification of the confirmed
disease. A clinically appropriate syndrome or broader aetiological category
may qualify under the case-specific acceptance criteria below.

GROUND TRUTH
  confirmed diagnosis : <true diagnosis>
  reference visible phenomenology : <reference sign>
  additional hypotheses accepted as COVERAGE, including qualifying syndromes
  and broader aetiological categories : <accept-as-coverage>
  related but insufficient hypotheses, NOT accepted as coverage :
    <related-but-not-covered>

THE MODEL'S ANSWER (free prose)
<the answer>
```

```
Read the whole answer. Grade SIGN and DIAGNOSIS independently.
Correct sign recognition does not by itself establish diagnostic coverage,
and an incorrect sign description does not rule out diagnostic coverage.

SIGN - did it identify the visible abnormality?
  correct : identifies the reference phenomenology or an equivalent physical
            description anywhere in the answer; clinical synonyms count.
  partial : describes a relevant abnormality in the correct body region,
            but the description is incomplete or too vague to establish
            the reference phenomenology.
  wrong   : identifies an incompatible phenomenology or the wrong body
            region, or does not describe a relevant visible abnormality.

DIAGNOSIS - does a proposed hypothesis cover the reference case?
  correct : identifies the confirmed disease or an equivalent diagnosis,
            OR identifies a syndrome or broader aetiological category
            included in the case-specific COVERAGE list.
            Equivalent clinical terminology is accepted.
            Exact disease identification and equal specificity to the
            confirmed diagnosis are NOT required for accepted hypotheses.
  partial : proposes a clinically related hypothesis that does not satisfy
            the COVERAGE criteria, including an entry in the
            related-but-not-covered list.
  wrong   : proposes an incompatible or unrelated hypothesis, or does not
            propose a diagnostic hypothesis.

Apply these rules:
- Do not reject an accepted hypothesis merely because it is broader than
  the confirmed disease.
- Do not automatically accept every broad syndrome or disease category.
  Broad hypotheses must match the case-specific COVERAGE list or an
  equivalent clinical expression.
- Merely repeating the visible sign does not establish coverage unless
  that expression also names an explicitly accepted diagnostic syndrome.
- Count hypotheses the model proposes as possibilities, even if they are
  not its leading diagnosis.
- Do not count a diagnosis mentioned only to deny or explicitly rule it out.
- Do not infer a diagnostic hypothesis that the model did not express.

Assign dx1 to the explicitly designated primary diagnosis, if present;
otherwise use the first proposed diagnosis. Assign dx2 and dx3 to the next
two distinct proposed diagnoses in presentation order.
Use "wrong" for missing slots.

Assign best using ALL distinct proposed diagnoses in the complete answer,
including diagnoses beyond dx3:
  correct : at least one hypothesis is graded correct.
  partial : none is correct, but at least one is partial.
  wrong   : all are wrong, or no diagnostic hypothesis is proposed.

The binary aetiological-coverage outcome is 1 if best is "correct",
and 0 otherwise. Partial hypotheses do not receive coverage credit.

Reply with ONLY this JSON:
{"sign":"correct|partial|wrong",
 "dx1":"correct|partial|wrong",
 "dx2":"correct|partial|wrong",
 "dx3":"correct|partial|wrong",
 "best":"correct|partial|wrong",
 "reason":"<= 12 words"}
```

Stage 1: the uncapped rank probe. This probe reads the complete answer and records the number of distinct proposed diagnoses and the rank of the first hypothesis satisfying the case-specific aetiological-coverage criteria in Section 3.3. Accepted syndromes and broader aetiological categories qualify; exact identification of the confirmed disease is not required.

```
A model was shown frames from a video of one patient and asked to describe
what it saw and propose possible diagnoses, with no history or test results.

TRUE DIAGNOSIS: <true diagnosis>
Additional hypotheses accepted as COVERAGE, including qualifying
syndromes and broader aetiological categories: <accept-as-coverage>
Related but insufficient hypotheses, NOT accepted as coverage:
<related-but-not-covered>

THE MODEL'S ANSWER:
<the answer>
```

```
Read the whole answer and list, in order of first appearance, every
distinct diagnosis proposed as a possibility. Merge synonymous entries
and exclude diagnoses mentioned only to rule them out.

Return the rank of the first hypothesis that covers the reference case:
either the confirmed disease or an explicitly accepted syndrome or
broader aetiological category. Equivalent clinical terminology counts.
Exact disease identification and equal specificity to the confirmed
diagnosis are NOT required. Do not accept other broad hypotheses unless
equivalent to an explicitly accepted hypothesis. Merely repeating the
visible sign does not qualify unless it names an accepted syndrome.

n_listed is the total number of distinct proposed diagnoses.
rank is the 1-based position of the first qualifying hypothesis.
Use rank = 0 if none qualifies. Consider the entire list, not only
the first three entries.

Reply with ONLY {"n_listed": <int>, "rank": <int>}
```

Stage 2: the committed diagnosis. Only the primary diagnosis is graded; the alternatives the model is allowed to append are recorded but not credited. A grade of `none` means NOT ACCURATE.

```
You are grading a doctor who watched a video of a patient, took a yes/no history,
ordered investigations, and then named a diagnosis.

GROUND TRUTH
  true diagnosis : <true diagnosis>
  count as accurate (the actual disease entity) : <accept-as-accurate>
  Accepted as partial: <accept-as-partial>

THE DOCTOR'S PRIMARY DIAGNOSIS
<the diagnosis>

Grade the primary diagnosis alone.
  accurate     : matches a case-specific accepted correct diagnosis,
                 including an accepted broader disease or syndrome
                 formulation; equivalent clinical terminology counts

partial      : matches the case-specific partial-credit criteria,
                 including a related syndrome, cause, or broad disease
                 family, but does not satisfy the correct-answer criteria

none         : satisfies neither the correct-answer criteria nor
                 the partial-credit criteria

Reply with ONLY {"grade":"accurate|partial|none","reason":"<= 12 words"}
```

I TEMPORAL-WINDOW TRAINING: REPRODUCIBILITY DETAILS

I.1 TEACHER: OFFLINE WINDOW SUPERVISION

The frozen teacher generates offline targets for temporal localisation. Given the video and two reference clinical fields, it selects at most one window or returns a no-window decision. Only the resulting text target is used to train the student; the teacher is absent at inference.

Model and decoding. We use `openai/gpt-5.6-luna` through OpenRouter, pinned to the OpenAI provider. The teacher runs once per clip (temperature 0, top-p 0.9, output limit 1100 tokens).

Inputs. Each silent clip is represented by 32 uniformly sampled frames, 380 px wide with preserved aspect ratio and JPEG quality 85. Each frame carries a top-left timestamp over an opaque patch, measured in seconds from clip start to two decimals. Images precede text in one user turn. The text supplies frame count, survey rate ($32/\text{duration}$), inter-frame interval, native recording rate, and shot boundaries for multi-shot clips.

The two clinical inputs are the verbatim reference phenomenology sentence and confirmed diagnosis. The prompt distinguishes windows located from visible frames, the reference sentence, or inference from the diagnosis. The teacher's explanations are not supplied to the student.

Target construction. We parse the last JSON object in the teacher's reply and convert a selected interval to its start time and span (end minus start), in seconds. The student target is text containing only a mode and, for a window request, one start/span pair. Frame rates, shot identifiers, confidence scores and explanations are excluded.

A `survey_sufficient=yes` response produces a `sufficient` target, ignoring any optional window. An `obtainable=no` response produces an `unobtainable` target. Contradictory flags are treated as invalid rather than as window requests.

Separation of supervision and input. Targets are generated offline before the five-fold split; each student is trained only on the targets in its training folds. Held-out teacher outputs are excluded. Neither student pass receives clinical context during training or inference. The reference phenomenology sentence serves only as the recognition target.

Full teacher prompt. The versioned template below is filled with video metadata, shot boundaries, and the two clinical fields. Doubled braces are literal-brace escapes for string substitution.

Teacher prompt (single-window protocol).

```
Above are {n_frames} frames covering the whole {duration}-second clip, in order, with the time
stamped on each. That is {sample_fps} frames per second - {sample_ms} milliseconds pass between one
frame and the next. The recording itself runs at {native_fps} frames per second.{shots_block}

You already know what this clip contains:
  what a clinician sees: {sign}
  what the patient has:  {diagnosis}

That sentence and that diagnosis are given to you and NOT to the person your window is for. Keep
track of which of the two you are using at each step: some of what you know is visible in these
frames, and some of it is not.

Your job is NOT to diagnose. It is to say which seconds of this recording, shown at what frame
rate, would let a viewer who knows nothing about it see the abnormality for themselves.

You choose only WHICH SECONDS and HOW MANY FRAMES PER SECOND. You cannot ask for a crop, a zoom, a
closer view, or any change of framing - the whole frame will be shown as it stands, at the size you
see above. If the thing is too small on screen to read at this framing, no choice of seconds or
rate will fix that, and the honest answer is that it cannot be obtained.

Nothing you ask for will be fetched. You are not being tested on whether you can then see it; you
are being asked to state the best request you can make, and to be explicit about how you arrived
at it.

Work in this order, and let the later steps follow from the earlier ones rather than being chosen
first.

1. What kind of thing is it.
     static_appearance   a feature you could read off a single still - a posture held, a lid
                         position, a pupil, an asymmetry that does not change
     continuous_movement something that goes on throughout, so any long enough stretch contains it
     episodic_event      something that happens at particular moments and not between them
     task_evoked         something that appears only while the patient is doing something, or while
                         being examined

2. Which part of the body, and how large it is on screen. If the part appears more than once in the
   picture - two hands, two eyes, two legs - say which one, by the side of the IMAGE it is on, not
   by the patient's own left and right. Give roughly what fraction of the frame width it occupies,
   and judge honestly whether a change in that part is readable at this size.

3. How long ONE occurrence lasts, in milliseconds - not how often it comes back. These are
   different numbers and only the first sets the frame rate: a movement lasting eighty milliseconds
   that returns every half second needs frames close enough together to catch the eighty, not the
   five hundred. Give both if the thing repeats.

   For a viewer to see an occurrence at all it must fall on at least two frames, so the frames must
   be closer together than HALF its duration:

     required frames per second  =  2000 / (duration of one occurrence in ms)

   Work that number out and compare it with the {sample_fps} frames per second above.

4. Propose at most ONE window: a stretch of seconds and a frame rate, entirely inside ONE shot.
   The windows array must contain zero or one entry, never more than one.

   For the window, say what it rests on:
     read_from_frames        you can point to the stamped frames where it is visible
     stated_in_the_sentence  the sentence you were given already names the moment or the task -
                             "on outstretching the arms", "while walking", "as the patient keeps
                             talking" - and your window follows from those words rather than from
                             anything you saw
     inferred_from_condition neither: you cannot see it here and the sentence does not say when,
                             so you are reasoning about when it is likely to be happening - while
                             the part is being used, while it is being tested, when it is at rest,
                             whatever the condition implies

   All three are legitimate. Answer stated_in_the_sentence whenever it is true, even if the frames
   happen to confirm it; we are counting how often the sentence, rather than the recording, is what
   located the sign.

   Be careful of one trap. A thing fast enough to need a higher rate is, by that very fact, a thing
   these frames are too slow to show, so you cannot expect to watch it happen and point there - and
   the movement that IS plainly visible above is the slow one, which is usually not the thing that
   matters. Do not simply name the seconds where the most conspicuous movement is.
```

```
   Give the window a confidence between 0 and 1, meaning how likely you think it is that a viewer
   shown that window, and told nothing, would describe the abnormality correctly.

   Two answers stand outside the window request, and you should give them when they are true:
     survey_sufficient   the frames above already show it; no window is needed
     obtainable = no     no choice of seconds or rate will show it at this framing

   You may still give one optional window when survey_sufficient is yes, but say so in the flag.
   If obtainable is no, return an empty windows array.

Separately, list the stamped times at which the thing can actually be SEEN in the frames above -
the ones you would point to as evidence. If it cannot be made out in any of them, give an empty
list. A run of consecutive frames is not evidence unless each of them shows it.

Your reasons will be read by a human reviewer and never shown to a model being trained. Even so,
write them so that they could be: name the part, the side, the size on screen, the timing, the
direction, and nothing that identifies a condition.

Reply as JSON:
{{"kind": "static_appearance or continuous_movement or episodic_event or task_evoked",
  "body_part": "...",
  "image_side": "left or right or either or not_applicable",
  "size_fraction_of_width": NUMBER between 0 and 1,
  "readable_at_this_size": "yes or no",
  "occurrence_duration_ms": NUMBER or null,
  "repetition_interval_ms": NUMBER or null,
  "required_fps": NUMBER or null,
  "sentence_names_when": "yes or no",
  "survey_sufficient": "yes or no",
  "obtainable": "yes or no",
  "not_obtainable_reason": "too_small or not_in_recording or other or null",
  "windows": [
    {{"shot": SHOT_NUMBER, "at_s": [START_SECONDS, END_SECONDS], "fps": NUMBER,
      "basis": "read_from_frames or stated_in_the_sentence or inferred_from_condition",
      "confidence": NUMBER between 0 and 1,
      "why": "one sentence"}}
  ],
  "evidence_s": [SECONDS, ...],
  "reason": "two or three sentences, following from steps 1 to 3"}}
Nothing else.
```

I.2 STUDENT: TRAINING AND TWO-PASS INFERENCE

The student shares one Qwen3.5-4B model and one LoRA adapter across two tasks: predicting a window from the survey and describing the visible phenomenology, both as plain text generation.

Two-pass inference. Pass 1 receives the timestamped 32-frame survey, without clinical fields, and predicts at most one start/span pair. Pass 2 generates one phenomenology sentence. A valid window adds an image–instruction turn containing K sampled frames; otherwise, recognition reuses the existing survey without presenting it again. The rules below determine which frames are used. Frame selection is discrete and does not propagate gradients.

Pass-1 response	Recognition evidence
Valid window	$K \in \{8, 16, 32\}$ evenly sampled frames within the window, without timestamps
`sufficient` or `unobtainable`	Original 32 survey frames; no new frames or repeated presentation
Invalid JSON	Skip window sampling; use the original survey without presenting it again

Text targets and instructions. Localisation predicts the teacher-derived mode and start/span, excluding reasoning and frame rate. For illustration, the interval $[2.0, 3.5]$ seconds can be represented as {`"mode":"request","start":2.0,"span":1.5`}. At $K = 16$, recognition receives 16 evenly sampled frames from that interval. The instruction requests at most one within-shot window, or a no-window response when the survey suffices or the sign cannot be captured at the available framing. It requests no diagnosis, explanation, crop, zoom, or frame rate.

Recognition uses this instruction ($N = K$ for a valid window, $N = 32$ for a fallback):

```
These {N} frames are sampled from a silent patient clip. In one sentence,
say what a clinician would see - the abnormality itself, in plain physical
words. Name no disease.
```

Its training target is the benchmark clinician's phenomenology sentence, verbatim. Neither pass receives that sentence, the diagnosis, or other clinical context as input.

Objective. Let $z_{1:L}$ be the tokenized teacher-derived window text and $y_{1:T}$ the reference phenomenology sentence. Both losses are mean autoregressive token cross-entropies:

$$\mathcal{L}_{\text{win}} = -\frac{1}{L}\sum_{\ell=1}^{L} \log p_\theta(z_\ell \mid z_{<\ell}, V_{32}, I_{\text{win}}), \qquad \mathcal{L}_{\text{phen}} = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, V_N, I_{\text{phen}}).$$

$$\mathcal{L} = 0.3\,\mathcal{L}_{\text{win}} + 0.7\,\mathcal{L}_{\text{phen}}.$$

Here V_{32} is the survey, V_N is the pass-2 input, and I is the instruction. Only target-text tokens contribute to loss; video, instruction, and template-prefix positions are masked. Only adapter parameters θ are trained, with no temporal-bin classifier, coordinate-regression head or frame-rate loss.

Training schedule and parameters. Recognition uses teacher-derived windows in epoch 1. During epoch 2, the probability of using the student's own prediction increases from 0 to 0.5. No-window responses and invalid JSON trigger the same survey fallback as at inference. Five-fold cross-validation groups cases by source article, with each student trained on four folds and evaluated on the held-out fold. The loss weight is fixed across folds. The table below lists all settings.

Base model	Qwen3.5-4B
Adapted	language side only; vision tower frozen
LoRA	$r = 16$, $\alpha = 32$, dropout 0.05, no bias
Target modules	`q,k,v,o,gate,up,down_proj`, `in_proj_qkv`, `out_proj`; `visual` excluded
Trainable	27.9M of 4.57B (0.61%)
Optimiser	AdamW, weight decay 0
Learning rate	1×10^{-4}, one-cycle, warmup 10% of steps
Batch	1 sample, gradient accumulation 4, gradient clipping 1.0
Precision	bfloat16, gradient checkpointing
Epochs (window adapter)	1 teacher-forced, 1 with scheduled sampling
Survey (pass 1)	32 frames, 380 px wide, JPEG q=85, timestamp burnt in
Recognition (pass 2)	valid window: $K \in \{8, 16, 32\}$, 384 px wide, no timestamp; otherwise reuse original 32-frame survey
Loss weight	$\lambda = 0.3$
Scheduled sampling	student window with probability ramped $0 \rightarrow 0.5$ over epoch 2
Folds	5, grouped by source article, fixed once and reused across budgets
Template	chat template rendered with the reasoning block closed
Teacher output limit	1100 tokens; parsing takes the last JSON object in the reply
Pass 1 decoding	greedy (`do_sample=false`), at most 1100 new tokens
Pass 2 decoding	greedy (`do_sample=false`), at most 64 new tokens
Python / PyTorch	3.10 / 2.8.0 (CUDA 12.8 runtime)
Transformers / PEFT	`5.13.0.dev0` / 0.19.1
GPU / driver	NVIDIA RTX A6000 (48 GB), driver `535.309.01`

Frame budget and comparison. A valid window adds K frames to the 32-frame survey; a fallback reuses the survey without further frames. Short windows may repeat source frames when K exceeds the number available at the native recording rate. Random preserves the predicted window duration but changes its position within the clip; both arms use the same adapter and fallback rule. The sign-only adapter uses the same clips, folds, target sentence and prompt, but is trained for one epoch per fold on $32+K$ evenly spaced frames. Paired comparisons appear in Appendix D.

J GRADING AND ORDER-MATCHING AUDITS

We check three aspects of reliability: agreement between automatic and human grading (§J.1), order-to-chart matching (§J.2), and the retrieval contrast under independent human grading (§J.3).

J.1 ALIGNMENT WITH HUMAN GRADING

An independent human reviewer re-graded sampled responses using the same rubric and blinding.

Sampling. For Stage 1, the sampling unit is one response per system, frame budget, and clip; we drew 100 responses uniformly at random from this population using a fixed seed. Four lacked usable paired grades and were excluded, leaving 96 for analysis. For Stage 2, no sampling was needed: the human reviewer graded the final diagnosis of every video-condition consultation of GPT-5.6-luna, one response per clip, giving 71 paired grades. Both lists are released with the benchmark.

Agreement. We report exact grade agreement and quadratically weighted Cohen's κ_w. Exact agreement means both graders assign the same category (not necessarily ACCURATE).

Table 32: **Agreement between the model grader and the independent human reviewer**; Stage 2 evaluation is restricted to the video condition (GPT-5.6-luna, all 71 consultations).

Stage	Outcome	n	Agreement	κ_w
1	Phenomenology recognition	96	89.6%	0.903
	Aetiology	96	96.9%	0.921
2	Final diagnosis	71	91.5%	0.947

For the primary Stage 2 outcome the graders agree on 65 of 71 responses (Table 33); all six disagreements involve adjacent grades, and none spans two levels.

Table 33: **Paired Stage 2 grades** in the video condition (71 consultations). Rows represent the model grader and columns the human reviewer.

Model / Human	Accurate	Partial	Not accurate
Accurate	32	1	0
Partial	1	10	2
Not accurate	0	2	23

J.2 HUMAN VERIFICATION OF ORDER MATCHING

We sampled 100 investigation orders from the video condition across five systems, uniformly at random with a fixed seed. The sample spans 52 clips. A human reviewer compared each order with its returned findings and the complete chart menu. Of these orders, 95 were matched correctly, 3 returned findings beyond those requested (*over-release*), and 2 omitted requested findings available in the chart (*missed matches*). The observed agreement with human review was therefore 95%.

J.3 RE-GRADING OF THE RETRIEVAL COMPARISON

Human re-grading confirms the retrieval gains: an independent reviewer gives source-clean consultations higher ACCURATE rates than video consultations for all three re-graded systems, with the largest gains for MiMo-v2.5 and MiniMax-M3 (Table 34). For each clip of GPT-5.6-luna, MiMo-v2.5 and MiniMax-M3, the reviewer saw the confirmed diagnosis, the accepted and partial-credit lists, and the video and source-clean final diagnoses in random order without condition labels, and assigned each the three-level grade of Appendix H. The baseline is the video arm of Table 2, as in Figure 7, and both graders of a system assessed the same consultation outputs.

Table 34: **Paired re-grading of video-condition and source-clean final diagnoses.** ACCURATE rates in percent; Δ is source-clean minus video (pp); paired clip counts follow each system.

system	grader	video	source-clean	Δ
GPT-5.6-luna (71)	human	46.5	53.5	$+7.0$
	automatic	46.5	49.3	$+2.8$
MiMo-v2.5 (71)	human	26.8	47.9	$+21.1$
	automatic	22.5	43.7	$+21.1$
MiniMax-M3 (71)	human	31.0	45.1	$+14.1$
	automatic	25.4	40.8	$+15.5$