TRACESCOPE: TRACING FAILURES FROM EVIDENCE TO DECISION IN UPPER-GI ENDOSCOPY

Anonymous authors
Paper under double-blind review

ABSTRACT

Upper-gastrointestinal (upper-GI) endoscopy spans clinical tasks from local visual assessment to examination-level interpretation and decision making. During an examination, clinicians localize anatomy, identify abnormalities, and characterize lesion attributes, while integrating distributed observations to formulate reports, diagnostic impressions, and guideline-guided decisions. These tasks require different scopes of visual evidence: local judgments rely on compact, target-relevant views, whereas examination-level interpretation requires integrating distributed and often redundant evidence across the examination. Recent endoscopic benchmarks have substantially broadened clinical task coverage, yet task definition and evidence scope are still commonly evaluated without systematic alignment. To bridge this gap, we introduce **TRACESCOPE**, a report-grounded multimodal benchmark for upper-GI endoscopy. Its main board contains 9,106 task-specific items from 4,893 examinations and 70,386 item–frame references. A shared clinical specification organizes six task families spanning site recognition, abnormal findings, finding attributes, report generation, diagnostic impression, and guideline reasoning. For each task, the delivered visual context is matched to its evidence requirements in clinical practice, ranging from compact, target-relevant views to redundant examination-level evidence. Across 22 multimodal models, the main board reveals a stage-selective capability profile: performance on local readouts does not extend uniformly to clinical composition, while reference-consistent reasoning remains separable from exact guideline decisions. Two complementary within-case analyses further show that selected multi-image presentation consistently outperforms equal-count uniformly sampled video input, while making guideline-sign polarity explicit yields the largest positive fleet-level change along the clinical-state cascade. Together, **TRACESCOPE** aligns stage-resolved clinical evaluation with task-appropriate evidence context, enabling unified analysis of model capability across the upper-GI workflow and of downstream decision sensitivity to visual and clinical information.

1 INTRODUCTION

Upper-gastrointestinal (upper-GI) endoscopy requires a sequence of linked clinical capabilities rather than a single visual prediction. During an examination, endoscopists localize observations across anatomical stations, identify and characterize mucosal abnormalities, and organize distributed observations into a findings report (Yadlapati et al., 2025; Areia et al., 2025). The resulting clinical evidence supports a diagnostic impression and, for guideline-codified conditions, a reproducible grade or clinical decision, as formalized by systems such as the Kyoto classification for *H. pylori*-associated gastritis, the Los Angeles classification for esophagitis, and the Paris classification for superficial neoplasia (Kato & Kamada, 2018; Wang et al., 2023; Lundell et al., 1999; Participants in the Paris Endoscopic Classification Workshop, 2003). This workflow spans local visual assessment, examination-level synthesis, and rule-guided decision making.

These clinical tasks require different scopes of visual evidence. Local findings and attributes can often be assessed from compact, target-relevant views, whereas examination-level reporting and diagnosis require integrating distributed and often redundant evidence across retained images or video. Recent endoscopic benchmarks have substantially broadened clinical coverage, from image-level question answering and disease assessment to anatomy, findings, diagnosis, recommendations, and

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: **TRACESCOPE overview.** *Top:* an upper-GI examination provides heterogeneous evidence contexts, from video and retained image series to selected views, together with its clinical report. *Center:* six task families span perception (T1–T3), examination-level synthesis (T4–T5), and rule-guided decision making (T6). *Bottom:* two same-case diagnostic axes provide a complementary form of tracing: Evidence Presentation varies how visual context reaches a fixed diagnostic target, whereas Diagnostic Cascade keeps the images fixed and progressively exposes clinical state and guideline information. Canonical clinical units and evidence-grounded construction link source records, task targets, and evaluation throughout the benchmark.

multi-stage reasoning (Gautam et al., 2024; 2025; Li et al., 2026b; Liu et al., 2025; Zhu et al., 2026; Chen et al., 2026), while long-context medical benchmarks increasingly study evidence integration over distributed or redundant procedural context (Du et al., 2026; Yao et al., 2026). Together, these directions highlight two complementary dimensions of evaluation: *what* clinical capability is being tested and *what scope of evidence* is appropriate for that capability. Their systematic alignment in upper-GI evaluation remains underexplored.

We introduce TRACESCOPE at this intersection. Its main board uses a shared upper-GI clinical specification to organize six task families spanning perception, examination-level synthesis, and rule-guided decision making, with task-specific evidence matched to the clinical unit being evaluated. This design provides a stage-resolved profile from local perception to examination-level synthesis and rule-guided decisions.

Two complementary within-case diagnostic axes extend this profile through controlled re-evaluation: *Evidence Presentation* varies how examination-level visual context reaches a fixed diagnostic target, whereas *Diagnostic Cascade* keeps the visual evidence fixed and progressively makes clinical state and guideline information explicit. Accordingly, *tracing* in TRACESCOPE has two complementary forms: localizing model strengths and failure patterns across clinically defined stages, and probing how a fixed downstream decision responds to controlled changes in its information interface.

Our contributions are fourfold:

• **A workflow-grounded upper-GI benchmark.** TRACESCOPE contains 9,106 items from 4,893 examinations across six task families spanning perception, examination-level synthesis, and guideline-guided decision making, with evidence scope matched to the clinical unit being evaluated.

Table 1: **Comparison with representative multimodal benchmarks.** *Scale* uses native units; pairs give images/questions, except MedHorizon (videos/questions; 759 h) and TRACESCOPE (examinations/items). *Task Fam.* denotes top-level tasks. I/MI/V = single image/multi-image/video. *Clin. Stages*, *Reason. Ref.*, *Guideline*, *Ctrl. Dx.*, and *Human Ref.* denote clinical reporting or decision stages, reasoning reference, guideline target, controlled diagnostic re-evaluation, and clinician reference. ✓/●/✗ = fully / partially / not supported. ♯ denotes an endoscopic subset.

Benchmark	Scale	Task Fam.	Evidence Context	Clin. Stages	Reason. Ref.	Guideline	Ctrl. Dx.	Human Ref.
OmniMedVQA (Hu et al., 2024) (Endo.♯)	1.88K samples	3	I	✗	✗	✗	✗	✗
GMAI-MMBench (Chen et al., 2024) (Endo.♯)	3.75K samples	7	I	✗	✗	✗	✗	✗
Kvasir-VQA (Gautam et al., 2024) (GS/CS)	6.5K/58.8K	6	I	✗	✗	✗	✗	✗
Kvasir-VQA-x1 (Gautam et al., 2025) (GS/CS)	6.5K/159.5K	18	I	✗	●	✗	●	✗
COLON-X (Ji et al., 2025) (CS)	212.7K/1.10M	18	I	✓	●	●	●	✗
EndoBench (Liu et al., 2025) (GS/CS/CE/SE)	6.83K VQA	12	I	✓	✗	●	●	✓
CE-Bench (Chen et al., 2026) (CE)	502.1K VQA	5	I/V	●	✓	✗	✗	✓
GI-Bench (Zhu et al., 2026) (GS/CS)	947/4.74K	5	I	✓	✗	●	✗	✓
MedHorizon (Du et al., 2026) (SE/CS/CE/US)	340/1.25K	13	V	●	✗	✗	●	✗
TRACESCOPE (GS, upper-GI)	**4.89K/9.11K**	**6**	**I/MI/V**	✓	✓	✓	✓	✓

• **Report- and evidence-grounded construction.** A shared clinical specification normalizes report-derived concepts, verifies visual support where required, and aligns the same canonical clinical units across item construction and evaluation.

• **Controlled diagnostic axes.** Two within-case analyses hold the diagnostic target fixed while varying either visual presentation or supplied clinical state, separating the effects of evidence selection, input quantity, and progressively explicit guideline information.

• **Model and physician evaluation.** Across 22 MLLMs and a blinded physician reference, TRACESCOPE reveals stage-selective performance, with local-task advantages not extending uniformly to clinical composition and reference-consistent reasoning remaining separable from exact guideline decisions. The paired analyses further show that selected visual presentation provides an advantage beyond input count, while explicit guideline-sign polarity produces the largest gain along the clinical-state cascade.

2 RELATED WORK

Medical and endoscopic benchmarks. Medical multimodal benchmarks range from broad cross-specialty evaluation, such as OmniMedVQA (Hu et al., 2024) and GMAI-MMBench (Chen et al., 2024), to endoscopy-specific test beds. Kvasir-VQA and Kvasir-VQA-x1 (Gautam et al., 2024; 2025) focus on image-level question answering and robustness, while Gastric-X (Li et al., 2026b) targets multimodal gastric-cancer assessment. More comprehensive endoscopic benchmarks increasingly broaden the clinical unit of evaluation: EndoBench (Liu et al., 2025) spans anatomy through recommendations, GI-Bench (Zhu et al., 2026) evaluates fine-grained endoscopic competence, and CE-Bench (Chen et al., 2026) studies multi-stage capsule-endoscopy reasoning. These efforts substantially expand clinical task coverage beyond isolated visual recognition. A complementary question, however, is what scope of visual evidence is appropriate for each clinical task, particularly when moving from local perception to examination-level reporting and diagnosis.

Distributed evidence and clinical process. Multi-image benchmarks such as MuirBench (Wang et al., 2025a) and MIOH (Min et al., 2026) study how models integrate information across views and changing image counts. Medical benchmarks extend this setting toward more distributed clinical evidence: MedThinkVQA (Yao et al., 2026) considers case-level integration, while MedHorizon (Du et al., 2026) evaluates long procedures in which relevant observations are embedded within broader visual context. A related line makes intermediate clinical reasoning itself part of the evaluation. DrVD-Bench (Zhou et al., 2026) and M3CoTBench (Jiang et al., 2026) expose reasoning stages, MedCalc-Bench (Khandekar et al., 2024) attributes errors to reasoning components, and MedChain (Liu et al., 2026) and MTBBench (Vasilev et al., 2026) evaluate decisions as clinical information progressively

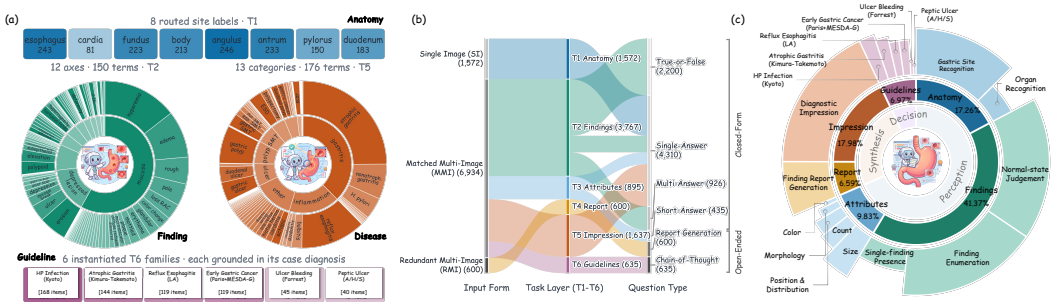


Figure 2: **Clinical specification and task organization.** (a) A shared taxonomy connects anatomy, finding, and disease vocabularies to six instantiated guideline families. (b) T1–T6 span task-appropriate input forms and structured or open-ended responses. (c) The resulting task families and subcategories compose the 9,106-item main board.

unfolds. Together, these studies highlight two complementary aspects of multimodal evaluation: how evidence is distributed and presented, and how clinical decisions change as relevant information becomes available.

Against this background, TRACESCOPE aligns an upper-GI-specific clinical specification with task-appropriate visual evidence. Its stage-resolved main board spans perception, examination-level synthesis, and guideline-guided decision making, while two within-case diagnostic axes separately probe sensitivity to visual presentation and supplied clinical state. Table 1 summarizes the closest comparisons.

3 THE TRACESCOPE BENCHMARK

TRACESCOPE organizes upper-GI evaluation along two complementary dimensions: the clinical capability being tested and the evidence context appropriate to that capability. Its six-task main board provides a stage-resolved view of upper-GI interpretation and decision making, grounded in retained visual evidence, contemporaneous reports, and normalized clinical concepts from the source examinations. A shared clinical specification aligns the target space, admissible evidence, response form, and scoring unit for each task.

Each source examination links retained endoscopic frames with its associated clinical report, preserving provenance from report statements and visual evidence to the derived benchmark targets. From 4,893 distinct examinations, the main board comprises 9,106 task-specific items and 70,386 item–frame references. Because the same frame may support multiple clinical questions, frame references are counted at the item level, while the visual input delivered to a model follows the serving configuration of the corresponding task. Each item retains the information needed for reproducible evaluation, including its examination, task, input form, normalized target, report provenance, and evidence links. The two diagnostic axes draw eligible examinations from the same cohort under separate paired protocols and are not included in the 9,106-item main-board count.

3.1 CLINICAL SPECIFICATION AND TASK FAMILIES

TRACESCOPE defines a shared clinical specification over anatomy, findings, attributes, diagnostic impressions, and guideline-governed decisions. Eight standardized upper-GI stations define T1 labels and route visual evidence. The finding vocabulary maps 1,079 surface aliases to 150 canonical descriptors across 12 attribute axes, supplying T2 and T3 targets and the clinical units recovered from T4 reports. The disease vocabulary distinguishes image-compatible impressions from pathology-only conclusions; T5 realizes 96 top-level labels, including 79 multiple-choice targets. These vocabularies operationalize the national endoscopy standard (Zhang & Linghu, 2020) and international Minimal Standard Terminology (Aabakken et al., 2009).

Clinicians validate the clinical vocabulary and its cross-walks between source expressions and canonical concepts. They also specify family-specific guideline cards covering applicability, observable

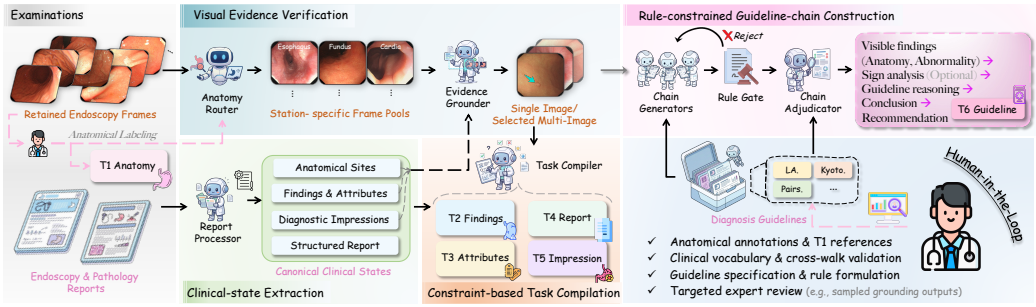


Figure 3: **The TRACESCOPE Multi-Agent Construction Engine.** Physician-annotated frames supply T1 items, and physician labels supervise the Anatomy Router. The Report Processor maps source reports into clinician-validated vocabularies; the Evidence Grounder checks visual support, and the Task Compiler produces T2–T5. Clinician-specified guideline cards and decision rules constrain the T6 branch, which combines Chain Generators, a deterministic Rule Gate, and a Chain Adjudicator. Clinical experts review sampled grounding outputs.

signs and their polarity, legal grades or decisions, quantitative criteria, and associated actions. T6 instantiates six guideline families. The cards constrain construction and deterministic checking.

The six task families span three broad forms of clinical capability—*Perception* (T1–T3), *Synthesis* (T4–T5), and *Decision* (T6):

- **T1: anatomy recognition** identifies a standardized anatomical station from one frame.
- **T2: abnormal findings** tests normal-state judgment, single-finding presence, and finding enumeration from matched frames. Finding enumeration uses single-choice and select-all questions.
- **T3: finding attributes** predicts categorical or quantitative properties such as number, size, distribution, morphology, and color.
- **T4: findings-report generation** organizes examination-level evidence into a site-indexed report.
- **T5: diagnostic impression** selects an image-compatible impression from the disease vocabulary.
- **T6: guideline-grounded reasoning** derives signs, grade or score, conclusion, and recommendation under a named guideline.

Appendix A defines the task forms, clinical vocabularies, and construction sources.

Evidence scope follows the clinical unit being evaluated. T1 uses a single image; T2 and T3 use compact matched evidence; T4 uses a redundant examination-level image set; and T5 and T6 use matched examination-level evidence. The benchmark therefore spans controlled local perception and more compositional examination-level tasks rather than imposing one visual context on every clinical stage. Outputs include binary and single-choice responses, select-all sets, open quantitative answers, reports, and reasoning chains. The same clinical specification defines the target space for construction and evaluation. Figure 2 connects the clinical taxonomy, task inputs and outputs, and benchmark composition.

3.2 EVIDENCE-GROUNDED BENCHMARK CONSTRUCTION

The clinical specification defines what each task can ask; the construction pipeline then binds each target to its source record and admissible evidence. T1 is selected directly from physician-annotated frames. For T2, T3 and T5, we normalize reports, route frames, and verify task-relevant visual claims before compiling eligible items. For T6, cited observations and family-specific clinical criteria form a separate reasoning branch (Figure 3).

Clinical-state extraction. The *Report Processor* segments findings by anatomical site, preserves negation and quantitative attributes, and maps context-dependent findings and impressions into constrained schemas. Closed vocabularies and validated cross-walks bind each mention to a canonical concept and its source field. Endoscopic impressions and pathology conclusions retain their distinct source identities. The output is a canonical clinical state anchored to the report.

Visual evidence verification. Physician annotations provide T1 reference labels directly and supervise the anatomical classifier used by the *Anatomy Router* to assign retained frames to eight station-specific pools. The *Evidence Grounder* compares a report-derived candidate with routed frames, records visual support, and ranks supporting evidence. A report-derived concept enters a visually grounded task only when the referenced frames satisfy its support policy: T2 requires support for the finding claim, T3 requires the queried attribute to be visually determinable, and T5 requires evidence compatible with the endoscopic impression. Appendix A.3 details T6 chain generation and adjudication.

Constraint-based task compilation. The *Task Compiler* instantiates T2–T5 question stems, option order, response forms, and clinically adjacent distractors through seeded deterministic templates. It checks record integrity, evidence linkage, answer-space membership, and quantitative semantics before an item enters the benchmark. Models assist clinical-state extraction and evidence verification; final T2–T5 item instantiation follows the controlled specification and templates.

Guideline-chain construction. For each T6 case, independent *Chain Generators* receive the same routed frames, clinical state, executable guideline card, answer space, and requested reasoning layers. They connect cited observations to sign states and interpret them under the card, with the recorded grade or pathology outcome anchoring the conclusion.

The deterministic *Rule Gate* checks applicability, cited-frame existence, anatomical compatibility, legal outputs, arithmetic where defined, and coherence between sign analysis and conclusion. It rejects invalid derivations and identifies disagreements among surviving chains. The *Chain Adjudicator* revisits disputed signs in the cited frames and retains, demotes, or drops each claim. A case is instantiated as a T6 item when generation, rule checking, and evidence-level adjudication produce an admissible reference chain. Each item retains a canonical target, evidence links, and a reasoning path from observation to clinical conclusion. Appendix A details construction sources, prompts, and validation.

3.3 Specification-aligned evaluation

Construction and evaluation share the same clinical specification. A report concept maps to a canonical identifier, the Task Compiler instantiates a target in that answer space, and evaluation recovers the same clinical unit from a model output. Closed-form tasks use exact or set-based agreement, while quantitative attributes use unit-aware numeric matching.

For the two compositional generative tasks, evaluation separates clinical content from surface wording. T4 measures anatomical coverage, canonical finding recovery, and unsupported content, while T6 measures agreement with the reference grade or diagnosis and reasoning consistency. Fixed semantic rubrics provide complementary analysis of report quality and guideline reasoning, with lexical and embedding similarities retained as wording-sensitive controls. Appendix B specifies the populations, parsers, rubrics, and scoring procedures.

4 Complementary Diagnostic Axes

The six-task main board traces where model strengths and failure signatures appear across clinically defined stages under task-appropriate evidence. The two diagnostic axes provide a complementary within-case view: they hold the examination, downstream target, answer space, and scoring rule fixed while varying one information interface at a time. *Evidence Presentation* changes how examination-level visual context is selected and delivered, whereas *Diagnostic Cascade* keeps the images fixed and progressively makes clinical state and guideline information explicit. These repeated conditions are analyzed separately from the 9,106 main-board items.

Evidence Presentation. Among the benchmark examinations, 193 retain video-level evidence suitable for controlled multi-presentation evaluation. For each examination, one fixed four-choice T5 diagnostic-impression question is evaluated under five primary presentations: *Selected Multi-Image*, *Retained Multi-Image*, and uniformly sampled video-frame presentations at 128, 256, and 512 frames (*Video-128*, *Video-256*, and *Video-512*). The question, options, reference impression, prompt

Table 2: **Stage-resolved main-board results on TRACESCOPE.** Header counts give task items; metric definitions and eligible populations appear in Appendix B.4. Darker shading and bold mark stronger and best deterministic scores. *Random* is matched chance, and the physician row is a separate reader-study reference.

Model	Anatomy (1,572)		Findings (3,767)				Attributes (895)					Report (600)					Impression (1,637)	Guidelines (635)	
	Gastric Site	*Organ*	*Normal Presence*		*Finding Enumeration*		*Size*	*Count*	*Pos./Dist.*	*Morph.*	*Color*	*Findings Report*					*Impr.*	*Reasoning*	
	Acc	Acc	Acc	Acc	MCQ-S Acc	MCQ-M Ex	MRA	MRA	Acc	Acc	Acc	Site Cov.	Find. Rec.	R-L	BS_{bio}	$SEFR^{J}$	Acc	Grade Acc.	$ELRC^{J}$
random	0.250	0.250	0.500	0.500	0.252	0.067	-	-	0.252	0.250	0.250	-	-	-	-	-	0.251	-	-
physician (subset)	1.000	1.000	0.943	0.947	1.000	0.600	0.650	0.836	0.800	0.800	1.000	0.988	0.540	0.484	0.846	0.702	0.867	0.600	0.659
Proprietary MLLMs																			
GPT-5.4	0.449	0.690	0.882	0.923	0.576	0.152	0.556	0.473	0.573	0.432	0.508	0.804	0.301	0.148	0.715	0.484	0.487	0.318	0.619
Claude Opus 4.7	0.661	**0.857**	**0.919**	0.925	0.573	0.150	0.527	0.263	**0.670**	**0.551**	0.646	0.888	0.397	0.153	0.715	0.532	0.399	**0.378**	0.655
Gemini 3.8 Flash	**0.765**	0.833	0.897	0.908	0.599	**0.200**	**0.569**	0.540	**0.670**	0.503	0.538	**0.945**	0.339	0.151	0.720	0.489	0.605	0.362	0.618
General-purpose MLLMs																			
InternVL3.5-8B	0.401	0.312	0.642	0.917	0.534	0.138	0.234	0.294	0.519	0.362	0.600	0.640	0.347	0.171	0.723	0.458	0.470	0.311	0.469
InternVL3.5-14B	0.475	0.192	0.874	0.911	0.604	0.179	0.504	0.227	0.553	0.395	0.631	0.455	0.156	0.143	0.707	0.451	0.469	0.327	0.485
MiniCPM-V-4.5 (8B)	0.456	0.185	0.782	0.935	0.432	0.140	0.521	0.517	0.453	0.346	0.435	0.746	0.409	0.149	0.709	0.435	0.509	0.299	0.438
Qwen3-VL-4B	0.346	0.181	0.802	0.834	0.543	0.112	0.393	0.443	0.524	0.427	0.677	0.518	0.294	0.142	0.712	0.401	0.503	0.289	0.413
Qwen3-VL-8B	0.322	0.235	0.746	0.927	0.537	0.132	0.225	0.412	0.539	0.427	0.677	0.567	0.286	0.126	0.721	0.423	0.603	0.278	0.416
Qwen3-VL-32B	0.339	0.477	0.893	0.941	0.602	0.131	0.440	0.498	0.597	0.503	0.692	0.815	0.453	0.144	0.723	0.425	0.602	0.291	0.517
Qwen3.5-4B	0.349	0.669	0.807	0.899	0.437	0.058	0.404	0.340	0.476	0.395	0.585	0.539	0.343	0.179	0.730	0.401	0.307	0.329	0.458
Qwen3.5-9B	0.426	0.521	0.821	0.937	0.559	0.120	0.403	0.422	0.519	0.411	0.708	0.578	**0.516**	0.174	0.740	0.417	0.396	0.348	0.498
Qwen3.5-27B	0.608	0.585	0.916	0.935	**0.655**	0.185	0.526	0.305	0.544	0.492	0.708	0.740	0.430	**0.184**	**0.749**	0.456	**0.631**	0.354	0.561
Video-native MLLMs																			
LLaVA-NeXT-Video-7B	0.244	0.326	0.556	0.150	0.268	0.027	0.009	0.000	0.374	0.254	0.508	0.197	0.078	0.133	0.678	0.371	0.364	0.155	0.296
VideoChat-Flash-7B	0.338	0.310	0.593	0.730	0.248	0.093	0.019	0.519	0.447	0.335	0.308	0.228	0.023	0.139	0.699	0.426	0.461	0.250	0.416
Medical MLLMs																			
ColonR1 (3B)	0.163	0.683	0.519	0.197	0.287	0.026	0.292	0.128	0.408	0.270	0.400	0.249	0.179	0.144	0.693	0.439	0.523	0.252	0.421
CE-R1 (7B)	0.373	0.188	0.380	0.591	0.498	0.108	0.039	0.120	0.325	0.465	**0.723**	0.048	0.015	0.133	0.673	0.317	0.586	0.051	0.224
HealthGPT-Pro-8B	0.384	0.556	0.701	0.942	0.627	0.143	0.428	0.626	0.481	0.422	**0.723**	0.362	0.084	0.070	0.680	0.405	0.547	0.248	0.412
Hulu-Med-4B	0.409	0.103	0.766	**0.945**	0.490	0.072	0.399	**0.809**	0.485	0.286	0.431	0.646	0.164	0.157	0.716	0.445	0.516	0.258	0.394
Hulu-Med-7B	0.545	0.467	0.785	0.918	0.518	0.125	0.403	0.235	0.495	0.416	0.523	0.261	0.091	0.147	0.740	0.465	0.395	0.256	0.377
Lingshu-I-8B	0.439	0.603	0.762	0.901	0.582	0.107	0.472	0.282	0.388	0.395	0.646	0.434	0.129	0.145	0.714	0.429	0.409	0.291	0.419
Lingshu-32B	0.437	0.592	0.712	0.937	0.607	0.112	0.468	0.387	0.495	0.389	0.462	0.440	0.171	0.182	0.727	0.459	0.602	0.267	0.439
MedGRPO-7B	0.388	0.268	0.542	0.797	0.466	0.102	0.200	0.321	0.529	0.422	0.492	0.242	0.071	0.153	0.729	0.409	0.524	0.246	0.393

structure, and scoring rule remain unchanged; only the delivered visual context varies. Diagnostic-impression accuracy is the primary endpoint. Two auxiliary count controls separate presentation from visual-input quantity: *Video-Match-N* matches the number of selected images for each examination, whereas *Video-Dense-N* matches the retained-image count. This axis therefore probes how stable the same diagnosis remains as visual context becomes broader and less selected.

Diagnostic Cascade. Among the main-board examinations, the second axis uses 1,220 with usable endoscopic images and pathology-verified *H. pylori* labels (459 positive and 761 negative). Each case is evaluated under five cumulative states: *Images*; + *Anatomy*; + *Findings*; + *Sign Polarity*; and + *Decision Threshold*. Anatomical locations come from the physician-supervised Anatomy Router. Findings and sign states are assembled from report extraction with targeted clinician annotations and checks during guideline-chain construction. The later states expose sign polarity and a clinician-specified decision rule: infection is predicted when infection-associated signs outnumber infection-negative signs by at least two. Images, question, pathology-backed target, answer space, prompt structure, and scoring rule remain fixed. The four adjacent transitions therefore probe how the same decision responds as information moves from anatomical and finding observations to guideline semantics and an explicit rule boundary. Balanced accuracy summarizes each state, while Cohen's κ is the primary chance-corrected endpoint for paired transitions. Panel composition, input delivery, and paired statistics are detailed in Appendices A and B.

5 EXPERIMENTS AND RESULTS

5.1 EXPERIMENTAL SETUP

Main-board and clinical-reference evaluation. We evaluate 22 MLLMs (19 open-source and three proprietary) spanning general-purpose, video-native, and medical-domain-specific families on the 9,106-item, six-task main board. Deterministic tasks use structured metrics; T4 report generation and T6 guideline reasoning additionally use fixed semantic rubrics. For clinical reference, two endoscopists independently answered a stratified 295-question panel spanning all tasks with the same evidence, answer spaces, and scoring rules, blinded to identifiers and reference answers.

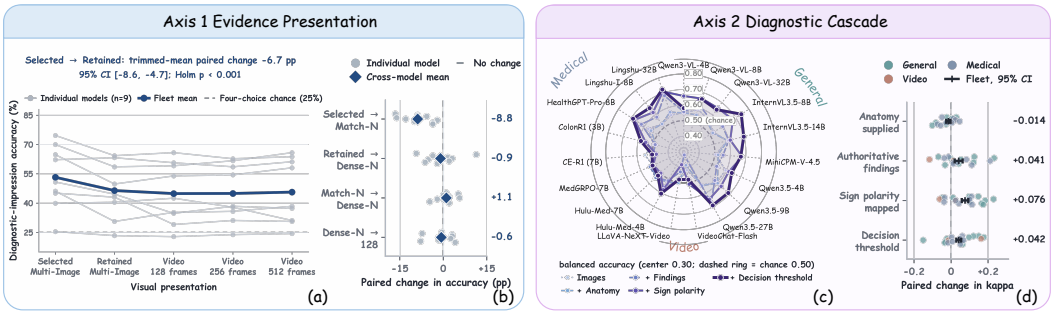


Figure 4: **Controlled responses along two diagnostic axes.** (a) T5 accuracy across five visual presentations; (b) count-controlled contrasts; (c) balanced accuracy across five clinical states; and (d) adjacent changes in Cohen's κ. Small points denote models, large markers the fleet, and intervals 95% bootstrap confidence intervals.

Diagnostic axes and statistical analysis. The diagnostic axes measure within-case decision sensitivity to controlled changes in the information interface. Evidence Presentation evaluates nine local models on 193 fixed diagnostic-impression targets; Diagnostic Cascade evaluates 19 local models on 1,220 fixed *H. pylori* targets. Their repeated conditions are reported separately from the main-board item count. Evidence Presentation uses diagnostic accuracy and paired percentage-point changes; Diagnostic Cascade uses balanced accuracy for state summaries and Cohen's κ for chance-corrected paired transitions. Bootstrap confidence intervals and multiplicity correction follow the prespecified paired protocols. Model and serving configurations, prompts, metrics, rubrics, and reader-study details appear in Appendix B; expanded results appear in Appendix C.

5.2 Stage-resolved capability and failure patterns

The main board resolves perception, examination-level synthesis, and rule-guided decision making into six task-specific capability profiles.

Performance is stage-selective. Across the 17 deterministic columns in Table 2, proprietary models lead nine, general-purpose open-source models five, and medical models three. Medical MLLMs lead selected finding and attribute readouts, whereas general-purpose or proprietary systems lead report, diagnostic-impression, and exact guideline scores. Leadership shifts across clinical outputs rather than following a single model ordering.

Local recognition and clinical composition remain distinct. Single-finding presence reaches an absolute accuracy of 0.945, whereas exact-set accuracy for multi-finding enumeration peaks at 0.200. In report generation, site coverage reaches 0.945, canonical finding recall 0.516, and judge-based SEFR 0.532; diagnostic-impression accuracy reaches 0.631. These readouts use different task populations and metrics, capturing local decisions, multi-finding composition, report content, and diagnostic impressions as distinct clinical outputs. Physician accuracy on the corresponding reader-study subsets is 0.947 for single-finding presence, 0.600 for exact multi-finding enumeration, and 0.867 for diagnostic impression.

Reasoning quality and exact guideline decisions remain separable. Exact guideline-grade agreement peaks at 0.378, whereas judge-based ELRC reaches 0.655; physicians obtain 0.600 and 0.659, respectively, on the reader-study subset. Reference-consistent reasoning structure can therefore be recovered without consistently yielding the exact guideline decision. Within Qwen3.5, scaling from 4B to 27B raises diagnostic-impression accuracy from 0.307 to 0.631, while exact guideline-grade agreement changes from 0.329 to 0.354. The paired axes next hold the downstream target fixed to test sensitivity to visual presentation and supplied clinical state.

5.3 Responses to Evidence Presentation

Evidence Presentation tests a fixed diagnostic decision under visual contexts ranging from selected images to broader examination evidence. Figure 4a–b shows a clear asymmetry: the dominant fleet-level change follows removal of the selected presentation; denser uniform sampling produces less consistent responses.

The dominant shift occurs when selection is removed. Replacing Selected Multi-Image with Retained Multi-Image changes accuracy by -6.7 percentage points (95% CI $[-8.6, -4.7]$; Holm-adjusted $p < 0.001$). Intervals cross zero for each subsequent transition from Retained Multi-Image through Video-512, and model trajectories improve, flatten, or decline. The fleet-wide effect is thus concentrated in evidence selection, not the nominal increase from 128 to 512 sampled frames.

Selected presentation is not explained by input count alone. Replacing Selected Multi-Image with an equal-count uniformly sampled video-frame presentation reduces accuracy for all nine models by 8.8 percentage points on average. The other count-controlled contrasts are smaller and mixed, averaging -0.9, $+1.1$, and -0.6 points (Table 14). For these fixed targets, presentation and selection determine model response more consistently than input quantity.

5.4 Responses along the Diagnostic Cascade

Diagnostic Cascade progressively operationalizes clinical observations, guideline semantics, and the decision boundary. Figure 4c–d localizes the largest, most consistent gain to mapping clinical findings onto guideline-sign polarity.

Clinical information has stage-selective value. Anatomy produces $\Delta\kappa = -0.014$; gains emerge with findings ($+0.041$), peak when sign polarity becomes explicit ($+0.076$), and continue with the decision threshold ($+0.042$). All four bootstrap intervals exclude zero and remain significant after Holm correction (Table 17). The cascade therefore elicits stage-selective rather than uniform responses.

Findings and rule information reshape different error components. The class-wise profiles also differ (Table 19). Findings raise mean sensitivity from 0.633 to 0.791 as specificity changes from 0.427 to 0.313. Sign polarity preserves sensitivity at 0.786 while raising specificity to 0.408; the threshold raises specificity further to 0.490. Balanced accuracy consequently rises from 0.530 after anatomy to 0.619 at the final state. Observations, guideline semantics, and the decision boundary therefore affect different components of the same binary decision.

Rule access remains distinct from rule application. After sign polarity, 15 of 19 models improve in κ; after the threshold, 13 improve, with varying effect sizes. The final model state averages $\kappa = 0.230$ and balanced accuracy 0.619. Directly applying the clinician-specified net-count rule to the same examinations yields $\kappa = 0.512$ and balanced accuracy 0.762, characterizing rule execution after the boundary is supplied. Appendix C.4 reports full state, family, and robustness results.

6 Conclusion

We introduced TraceScope, a six-task upper-GI benchmark that aligns report-derived targets, visual evidence, and evaluation through a shared clinical specification. Across 22 models, the main board reveals a non-uniform capability profile: performance on local readouts does not extend uniformly to clinical composition, and reference-consistent reasoning remains distinct from exact guideline decisions. Two within-case axes sharpen this picture, showing that selected visual evidence retains an advantage beyond input count, while explicit sign polarity produces the largest gain along the Diagnostic Cascade. Together, these results position TraceScope as a framework for tracing where capability breaks down from evidence to decision, and how those decisions shift with the visual and clinical information made available.

AI USE STATEMENT

Generative AI supported benchmark design and experimental protocols, clinical-report processing, result interpretation, manuscript editing, figure creation, and plotting-code development. Language models assisted with translation-memory completion, closed-vocabulary extraction, report-to-frame grounding, T6 reference-chain generation and adjudication, and rubric-based scoring of T4 and T6 responses. AI-assisted construction and evaluation followed predefined clinical vocabularies, visual-support policies, deterministic compilation rules, executable guideline checks, and semantic rubrics. The authors reviewed this work and take responsibility for the manuscript, benchmark, analyses, claims, and artifacts. Generative AI was not used to create synthetic datasets, develop theoretical models, formulate or prove mathematical claims, or propose or refine research hypotheses.

ETHICS STATEMENT

This study received Institutional Review Board (IRB) approval. Records and images were fully de-identified before construction; there was no prospective recruitment or intervention and no foreseeable additional risk. Data handling complied with applicable privacy and legal requirements. Clinical annotations underwent independent review and adjudication, and standardized evaluation procedures were used to limit avoidable bias. The authors declare no conflicts of interest or sponsorship. Details on data handling and annotation appear in Appendix A.1.

REPRODUCIBILITY STATEMENT

The paper and appendices describe all models, algorithms, configurations, benchmarks, prompts, and scoring procedures. After acceptance, we will release the complete benchmark, including images and annotations, all code and configurations for the reported results, and the anatomy-recognition training code. The accepted version will provide the public repository URL and versioned release.

REFERENCES

Lars Aabakken, Bjorn Rembacken, Olivier LeMoine, Konstantin Kuznetsov, Jean-Francois Rey, Thomas Rösch, Glen Eisen, Peter Cotton, and Masayuki Fujino. Minimal standard terminology for gastrointestinal endoscopy–mst 3.0. *Endoscopy*, 41(08):727–728, 2009.

Anthropic. Introducing Claude Opus 4.7. `https://www.anthropic.com/news/claude-opus-4-7`, 2026a. API model claude-opus-4-7, accessed Sep. 25, 2026.

Anthropic. Introducing Claude Opus 4.8. `https://www.anthropic.com/news/claude-opus-4-8`, 2026b. API model claude-opus-4-8, accessed Sep. 25, 2026.

Miguel Areia, Gianluca Esposito, Philippe Leclercq, Marcin Romańczyk, Jasmin Zessner-Spitzenberg, Pedro G Delgado Guillena, Ashraf Monged, Raul Honrubia López, Hugo Uchima, Eduardo J Ruiz Ballesteros, et al. Performance measures for upper gastrointestinal endoscopy: a european society of gastrointestinal endoscopy (esge) quality improvement initiative–update 2025. *Endoscopy*, 57 (11):1268–1297, 2025.

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

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37: 94327–94427, 2024.

Wenting Chen, Shengyuan Liu, Boyun Zheng, et al. An adaptive foundation model with evidence-based clinical reasoning for gastroenterology. *Research Square preprint rs-8459685*, 2026. doi: 10.21203/rs.3.rs-8459685/v1.

Bodong Du, Bowen Liu, Yang Yu, Xinpeng Ding, Zhiheng Wu, Shuning Wang, Shuo Nie, Naiming Liu, Qifeng Chen, Yangqiu Song, et al. Medhorizon: Towards long-context medical video understanding in the wild. *arXiv preprint arXiv:2605.06537*, 2026.

JohnA H Forrest, NDC Finlayson, and DJC Shearman. Endoscopy in gastrointestinal bleeding. *The Lancet*, 304(7877):394–397, 1974.

Sushant Gautam, Andrea M Storås, Cise Midoglu, Steven A Hicks, Vajira Thambawita, Pål Halvorsen, and Michael A Riegler. Kvasir-vqa: A text-image pair gi tract dataset. In *Proceedings of the first international workshop on vision-language models for biomedical applications*, pp. 3–12, 2024.

Sushant Gautam, Michael Riegler, and Pål Halvorsen. Kvasir-vqa-x1: a multimodal dataset for medical reasoning and robust medvqa in gastrointestinal endoscopy. In *MICCAI Workshop on Data Engineering in Medical Imaging*, pp. 53–63. Springer, 2025.

Google DeepMind. Gemini 3.1 Pro. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, 2026a. Official model card, accessed Sep. 25, 2026.

Google DeepMind. Gemini 3.8 Flash. `https://deepmind.google/models/model-cards/gemini-3-8-flash/`, 2026b. Official model card, accessed Sep. 25, 2026.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183. IEEE, 2024.

Ge-Peng Ji, Jingyi Liu, Deng-Ping Fan, Huazhu Fu, and Nick Barnes. Colon-x: Advancing intelligent colonoscopy toward clinical reasoning. *arXiv preprint arXiv:2512.03667*, 2025.

Juntao Jiang, Jiangning Zhang, BAI Jinsheng, Weixuan Liu, Weiwei Jin, Zhucun Xue, Yong Liu, Xiaobin Hu, Shuicheng YAN, et al. M3cotbench: Benchmark chain-of-thought of mllms in medical image understanding. In *International Conference on Learning Representations*, volume 2026, pp. 89996–90036, 2026.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

M Kato and T Kamada. Endoscopic findings for risk stratification of gastric cancer. *Kyoto classification of gastritis. 2nd ed. Tokyo: Nihon Medical Center*, pp. 120–126, 2018.

Nikhil Khandekar, Qiao Jin, Guangzhi Xiong, Soren Dunn, Serina S Applebaum, Zain Anwar, Maame Sarfo-Gyamfi, Conrad W Safranek, Abid A Anwar, Andrew Zhang, et al. Medcalc-bench: Evaluating large language models for medical calculations. *Advances in Neural Information Processing Systems*, 37:84730–84745, 2024.

Ken Kimura and T Takemoto. An endoscopic recognition of the atrophic border and its significance in chronic gastritis. *Endoscopy*, 1(03):87–97, 1969.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026a.

Yuanzhe Li, Hao Chen, Rui Yin, Juyan Ba, Yu Zhang, and Sheng Lu. Gastric-x: A multimodal multi-phase benchmark dataset for advancing vision-language models in gastric cancer analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2490–2501, 2026b.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, Yihang Su, Kao-Jung Chang, Haoliang Li, Linlin Shen, Michael R Lyu, and Wenting Chen. Medchain: Bridging the gap between llm agents and clinical practice with interactive sequence. *Advances in Neural Information Processing Systems*, 38, 2026.

Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. Endobench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. *arXiv preprint arXiv:2505.23601*, 2025.

LR Lundell, J Dent, JR Bennett, AL Blum, D Armstrong, JP Galmiche, F Johnson, M Hongo, JE Richter, SJ Spechler, et al. Endoscopic assessment of oesophagitis: clinical and functional correlates and further validation of the los angeles classification. *Gut*, 45(2):172–180, 1999.

Joonki Min, Chaeyun Kim, Hyungwook Choi, Yejin Kim, Kihyun Kim, Yohan Jo, and Joonseok Lee. Fine-grained multi image object hallucination benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18295–18305, 2026.

Manabu Muto, Kenshi Yao, Mitsuru Kaise, Mototsugu Kato, Noriya Uedo, Kazuyoshi Yagi, and Hisao Tajiri. Magnifying endoscopy simple diagnostic algorithm for early gastric cancer (mesda-g). *Digestive endoscopy*, 28(4):379–393, 2016.

OpenAI. Introducing Structured Outputs in the API. `https://openai.com/index/introducing-structured-outputs-in-the-api/`, 2024. API model gpt-4o-2024-08-06, accessed Sep. 25, 2026.

OpenAI. Introducing GPT-5.4. `https://openai.com/index/introducing-gpt-5-4/`, 2026a. Product release, accessed Sep. 25, 2026.

OpenAI. Introducing GPT-5.5. `https://openai.com/index/introducing-gpt-5-5/`, 2026b. API model gpt-5.5-0424-global, accessed Sep. 25, 2026.

Participants in the Paris Endoscopic Classification Workshop. The Paris endoscopic classification of superficial neoplastic lesions: esophagus, stomach, and colon. *Gastrointestinal Endoscopy*, 58(6 Suppl):S3–S43, 2003. doi: 10.1016/s0016-5107(03)02159-x.

Pedro Pimentel-Nunes, Diogo Libânio, Ricardo Marcos-Pinto, Miguel Areia, Marcis Leja, Gianluca Esposito, Monica Garrido, Ilze Kikuste, Francis Megraud, Tamara Matysiak-Budnik, et al. Management of epithelial precancerous conditions and lesions in the stomach (maps ii): European society of gastrointestinal endoscopy (esge), european helicobacter and microbiota study group (ehmsg), european society of pathology (esp), and sociedade portuguesa de endoscopia digestiva (sped) guideline update 2019. *Endoscopy*, 51(04):365–388, 2019.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026. URL `https://qwen.ai/blog?id=qwen3.8`.

Tatsuzo Sakita and Hiroshi Fukutomi. Endoscopic diagnosis. In Yasuo Yoshitoshi (ed.), *Ulcer of Stomach and Duodenum*, pp. 198–208. Nankodo, Tokyo, 1971.

Prateek Sharma, John Dent, David Armstrong, Jacques JGHM Bergman, Liebwin Gossner, Yoshio Hoshihara, Janusz A Jankowski, Ola Junghard, Lars Lundell, Guido NJ Tytgat, et al. The development and validation of an endoscopic grading system for barrett's esophagus: the prague c & m criteria. *Gastroenterology*, 131(5):1392–1399, 2006.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Yuhao Su, Anwesa Choudhuri, Zhongpai Gao, Benjamin Planche, Van Nguyen Nguyen, Meng Zheng, Yuhan Shen, Arun Innanje, Terrence Chen, Ehsan Elhamifar, et al. Medgrpo: Multi-task reinforcement learning for heterogeneous medical video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2788–2798, 2026.

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

Kiril Vasilev, Alexandre Misrahi, Eeshaan Jain, Phil F Cheng, Petros Liakopoulos, Olivier Michielin, Michael Moor, and Charlotte Bunne. Mtbbench: A multimodal sequential clinical decision-making benchmark in oncology. *Advances in Neural Information Processing Systems*, 38, 2026.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, volume 2025, pp. 62624–62650, 2025a.

Kaijie Wang, Jing Zhao, Haifeng Jin, Lina Meng, Yihong Fan, Yanlin Zhou, Cheng Ye, Meng Li, Panpan Ma, Linwensi Zhu, et al. Establishment of a modified kyoto classification scoring model and its significance in the diagnosis of helicobacter pylori current infection. *Gastrointestinal endoscopy*, 97(4):684–693, 2023.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Rena Yadlapati, Dayna Early, Prasad G Iyer, Douglas R Morgan, Neil Sengupta, Prateek Sharma, and Nicholas J Shaheen. Quality indicators for upper gi endoscopy. *Gastrointestinal endoscopy*, 101 (2):236–260, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zonghai Yao, Benlu Wang, Yifan Zhang, Junda Wang, Iris Xia, Zhipeng Tang, Shuo Han, Zhichao Yang, Arman Cohan, and Hong Yu. Medical thinking with multiple images. In *International Conference on Learning Representations*, volume 2026, pp. 101924–101988, 2026.

Zhenyu Yi, Jianwei Xu, Yue Hu, Zhongwei Qiu, Sijing Li, Liang Huang, Bin Lv, Ling Zhang, and Yingda Xia. Endovlm: An endoscopy vision-language pre-training model via anatomy-guided sparsity and progressive alignment. *arXiv preprint arXiv:2608.04472*, 2026.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026.

Shutian Zhang and Enqiang Linghu (eds.). *Standard Terminology for Diagnosis and Treatment in Digestive Endoscopy*. People's Medical Publishing House, Beijing, 2020. ISBN 978-7-117-30584-6. In Chinese.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`.

Tianhong Zhou, Yingtao Zhu, Chuxi Xiao, Haiyang Bian, Lei Wei, Xuegong Zhang, et al. Drvd-bench: Do vision-language models reason like human doctors in medical image diagnosis? *Advances in Neural Information Processing Systems*, 38, 2026.

Yan Zhu, Te Luo, Pei-Yao Fu, Zhen Zhang, Zi-Long Wang, Yi-Fan Qu, Zi-Han Geng, Jia-Qi Xu, Lu Yao, Li-Yun Ma, et al. Gi-bench: A panoramic benchmark revealing the knowledge-experience dissociation of multimodal large language models in gastrointestinal endoscopy against clinical standards. *arXiv preprint arXiv:2601.08183*, 2026.

APPENDIX INDEX

- **A — Benchmark construction and reference validation:** A.1 source cohort and physician annotation; A.2 clinical specification and guideline cards; A.3 construction engine and prompts; A.4 benchmark composition and diagnostic panels; A.5 reference standards and evidence validation.
- **B — Evaluation protocol:** B.1 models and serving; B.2 input delivery and diagnostic axes; B.3 prompts and response forms; B.4 metrics and statistics; B.5 semantic rubrics; B.6 physician reader study.
- **C — Additional experimental results:** C.1 expanded task results; C.2 T4 and T6 judge results; C.3 Evidence Presentation; C.4 Diagnostic Cascade.
- **D — Representative task examples:** D.1 evaluated response forms; D.2 visual examples across tasks.
- **E — Scope and interpretation.**
- **F — Discussion and outlook.**

A BENCHMARK CONSTRUCTION AND REFERENCE VALIDATION

A.1 SOURCE COHORT AND PHYSICIAN ANNOTATION

Source scope and data governance. The benchmark comprises 4,893 distinct examination identifiers sourced across multiple centers. All clinical records and endoscopic images were fully de-identified prior to benchmark construction, with each examination assigned a stable, non-identifying benchmark identifier and an optional hashed key to support internal patient-level grouping where available. Personal names and identifier-bearing source paths are entirely excluded. Within each case record, findings, impressions, pathology annotations, examination metadata, and image-set pointers remain linked through the non-identifying examination identifier. The benchmark is structured at the item level: multiple tasks can be derived from a shared examination, which yields the final set of 9,106 evaluation items across the 4,893 distinct examinations.

Physician annotation. Physician station annotations supply the T1 reference labels and supervise the *Anatomy Router*; T1 gold labels remain physician-assigned. For the station-annotation corpus, two independent endoscopists, each with more than five years of experience, initially labeled every frame. A senior endoscopist with more than 15 years of experience then reviewed the annotations; any remaining disagreements were referred to a fourth senior expert for final adjudication and a unified ground truth.

The station-annotation corpus contains 6,589 frames from 334 examinations. The Anatomy Router uses 4,132 frames from 216 examinations for training and 833 frames from 38 examinations for held-out testing. A separate pool of 1,624 frames from 80 examinations supplies the T1 references. These partitions keep the T1 evaluation images outside router training. T1 items are drawn directly from physician annotations, retaining clear, high-quality views while reducing near-duplicate frames within an examination and balancing anatomical classes. This procedure yields 1,572 single-frame T1 items. References for the remaining tasks are anchored to the contemporaneous endoscopy report.

Data release. The complete de-identified benchmark, including images and annotations, will be made publicly available after acceptance under the applicable institutional data-sharing policy.

A.2 CONTROLLED CLINICAL SPECIFICATION AND GUIDELINE CARDS

The clinical specification connects anatomy, findings, disease impressions, and guideline-governed decisions. Figure 5 expands the term hierarchy, while Figure 6 shows the six guideline families instantiated by T6. Figure 7 shows the same vocabularies as the benchmark items exercise them. The specification preserves separate roles for a canonical concept, its surface aliases, its source field, and the task in which it is admissible.

Anatomy. Eight standardized upper-GI stations define the T1 label space and the frame-routing interface.

Findings and attributes. The abnormal-finding map contains 150 canonical terms across 12 attribute axes, normalized from 1,079 surface aliases. Normal-polarity entries represent explicit normal statements and negation rather than inferring them from absence.

Diagnostic impressions. The disease vocabulary contains 176 standardized terms across four organs and 13 categories, with 128 active for the upper-GI scope. T5 realizes 96 distinct top-level labels; its multiple-choice variants use 79 distinct target strings. Distractors draw from a larger union of target and option strings, while the answer space remains defined by top-level labels.

The terminology package cross-maps the national endoscopy standard (Zhang & Linghu, 2020) to the international Minimal Standard Terminology (Aabakken et al., 2009). Clinical experts validate the vocabulary and cross-walks from source expressions to canonical concepts; fail-closed checks ensure that benchmark targets use mapped source phrases.

Guideline cards and benchmark families. Clinical experts specify the guideline cards and formulate their clinical criteria, including applicability conditions, visual signs, legal conclusions,

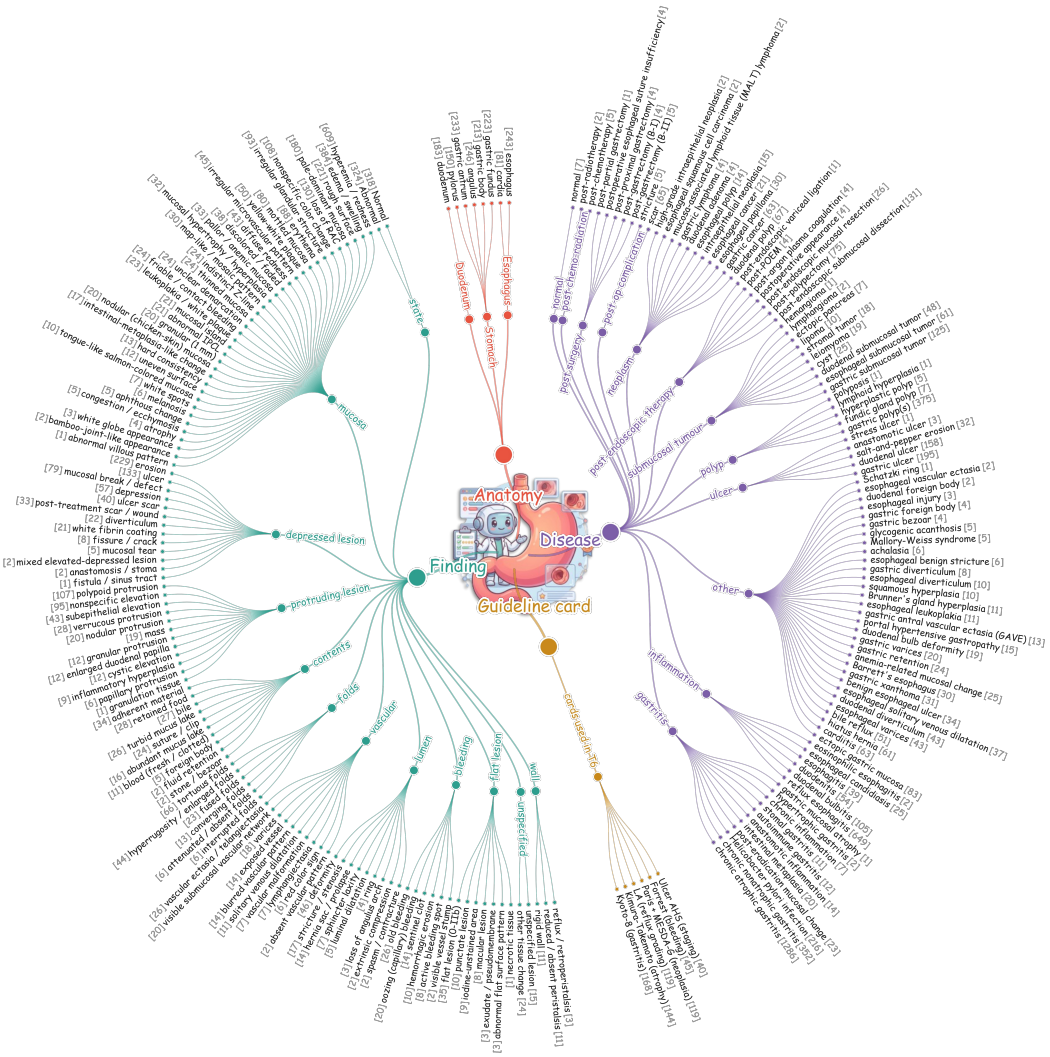


Figure 5: The controlled vocabularies as a term hierarchy: anatomy, the twelve attribute axes of standardized finding terms, disease impressions, and guideline cards.

thresholds where applicable, and associated actions. The registry contains 14 serialized YAML cards representing 12 clinical systems, with three cards encoding the *H. pylori* system. Multiple cards may encode one clinical family. The benchmark instantiates six families: an eight-sign *H. pylori* assessment drawing on Kyoto-associated endoscopic findings (Kato & Kamada, 2018; Wang et al., 2023); the Los Angeles classification (Lundell et al., 1999); a gastric-lesion family combining Paris morphology (Participants in the Paris Endoscopic Classification Workshop, 2003), MESDA-G magnifying-endoscopy criteria (Muto et al., 2016), and MAPS II management context (Pimentel-Nunes et al., 2019); the Forrest bleeding classification (Forrest et al., 1974); the Kimura–Takemoto atrophy classification (Kimura & Takemoto, 1969); and the Sakita active/healing/scarring ulcer stages (Sakita & Fukutomi, 1971). Prague C&M (Sharma et al., 2006) is reserved for video-based evaluation because its extent measurement depends on a dynamic landmark.

The *H. pylori* assessment uses six infection-associated and two infection-negative endoscopic signs. In main-board T6, their visual interpretation supports a binary conclusion anchored to pathology-confirmed infection status.

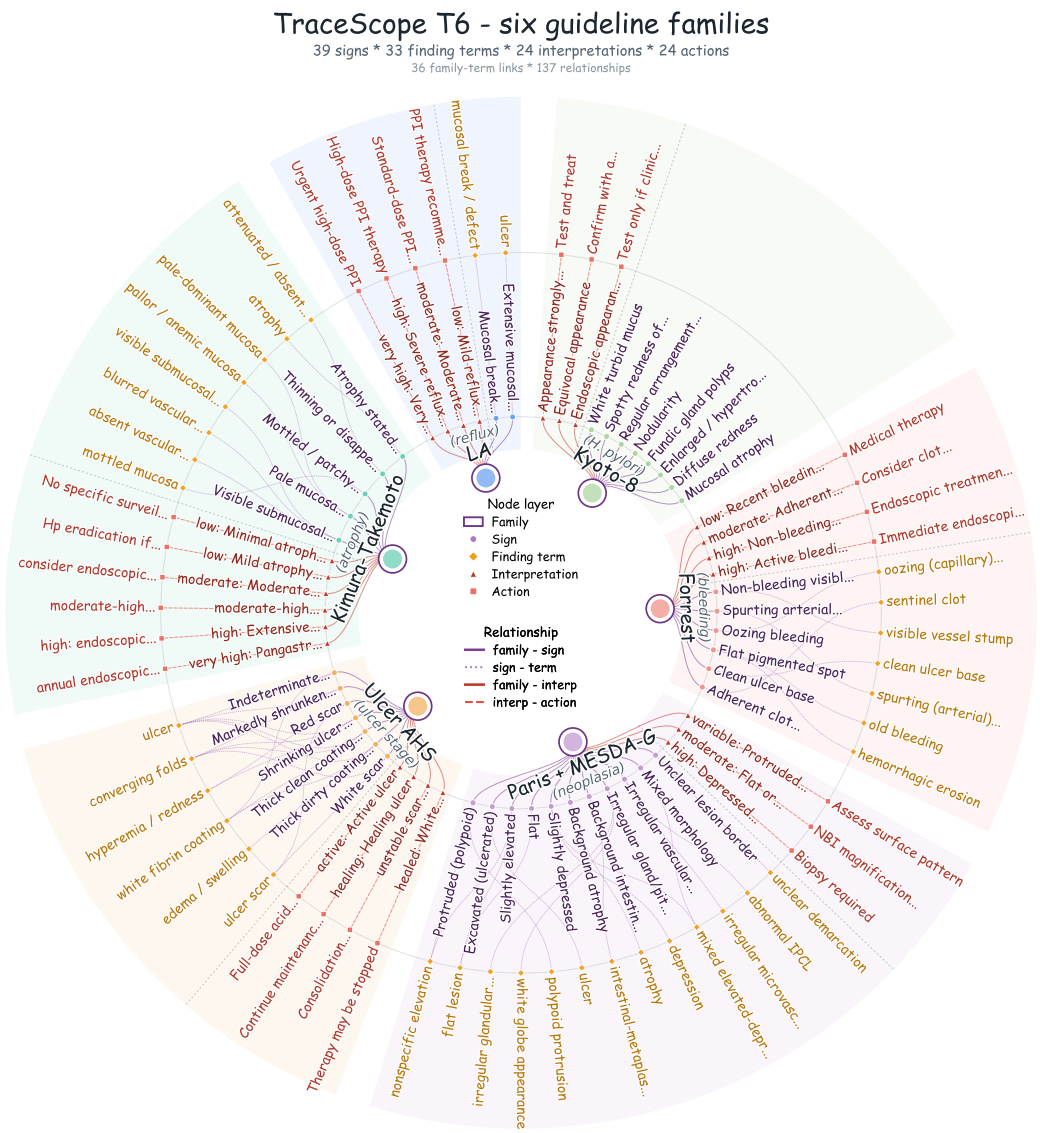


Figure 6: The six T6 guideline families instantiated in the benchmark and their links to signs, standardized findings, interpretation bands, and actions. The graph contains 39 signs, 33 finding terms, 24 interpretations, 24 actions, and 137 relationships.

A.2.1 Task families and clinical roles

The six task families represent distinct clinical stages of upper-GI reporting and decision-making. Anatomy locates an observation; findings establish what is visible; attributes describe clinically relevant properties; a findings report organizes observations across sites; a diagnostic impression summarizes the examination; and guideline reasoning maps observations and clinical signs to a rule-governed decision. This organization defines the clinical representations evaluated by each task. Main-board items are scored within their own tasks, while the diagnostic axes use paired cases to test how a fixed decision responds to changes in visual evidence or supplied clinical state.

T1: Anatomy Recognition (Anatomy). *Organ Recognition* identifies the upper-GI organ shown in a single frame. *Gastric Site Recognition* distinguishes anatomical stations within the stomach. These two levels of localization establish where a visual observation belongs, so later finding descriptions

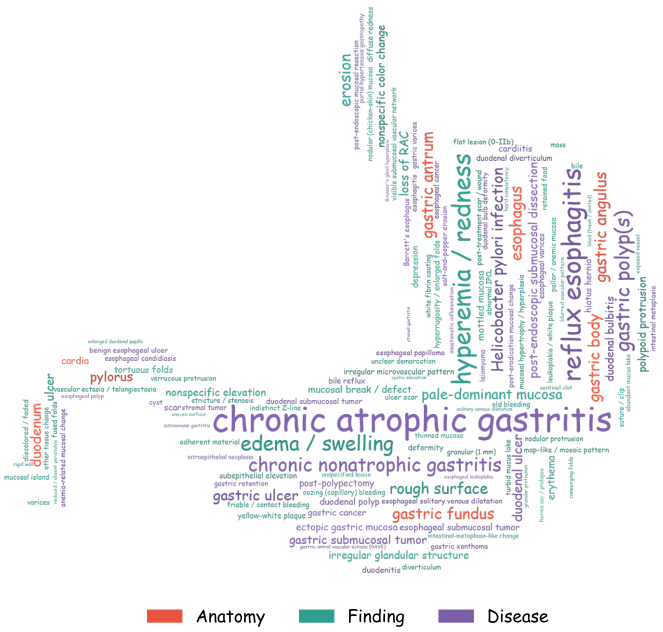


Figure 7: The vocabularies as the benchmark items exercise them: one word per anatomical site, standardized finding, and disease impression, sized by how many items reference it. Frequencies count T1 site labels, the findings listed in T2 finding answers, and the labels listed in T5 impression answers; terms referenced only once are omitted, leaving 8 sites, 111 findings, and 88 disease terms.

and report statements can be associated with the appropriate site. Both are single-choice questions grounded in physician station annotations.

T2: Abnormal Finding Assessment (Findings). *Normal-State Judgment* determines whether the represented site is normal or abnormal. *Single-Finding Presence* asks whether a named abnormal finding is visible. *Finding Enumeration* identifies the abnormal findings supported by the delivered images, using either a single-choice or a select-all question. The first two subtasks test state and named-cue recognition; enumeration tests recovery of the finding identity or set needed to describe the examination. The two enumeration question forms belong to the same subtask and are scored separately because a single answer and an exact set make different demands on the response.

T3: Finding Attribute Specification (Attributes). *Size* and *Count* specify the extent and number of a finding through quantitative answers. *Position and Distribution* describes where a finding lies and how it is spread; *Morphology* identifies its shape or structural appearance; and *Color* records its visual coloration. These properties turn a finding label into a clinically informative description and supply detail used in site-level reporting and, where applicable, guideline interpretation. The categorical attributes use single-choice responses, while the quantitative attributes use open answers.

T4: Findings Report Generation (Report). The model produces an examination-level findings report organized by anatomical site. This task integrates observations from a larger, partly redundant image set into a coherent account of what was seen and where. Its clinical content is evaluated through site coverage and canonical finding recovery, alongside complementary wording-sensitive and semantic measures. Report generation is one task; its scoring dimensions are not separate subtasks.

T5: Diagnostic Impression Prediction (Impression). The model selects an image-compatible endoscopic impression from the controlled disease vocabulary. This step summarizes the examination at the diagnosis level, beyond listing findings or composing descriptive text. The target is anchored

Table 3: **Construction roles and constraints.** Each role has a bounded input, output, and admissibility rule.

Role	Input → output; admissibility rule	Implementation
Report Processor	Source reports → canonical clinical state; closed vocabulary and source-field preservation	Qwen3-235B-A22B (Yang et al., 2025)
Anatomy Router	Frame → upper-GI station; physician-supervised eight-way labels	Anatomy recognition model (DINOv3 ViT-B/16 (Siméoni et al., 2025))
Evidence Grounder	Candidate and routed frames → support decision and ranked frames; task-specific visual-support policy	Qwen3-VL-235B-A22B-Instruct-FP8 (Bai et al., 2025)
Task Compiler	Eligible record and evidence → T2–T5 item; seeded templates and valid answer spaces	Deterministic code
Chain Generators	Frames, state, and card → candidate T6 chains; shared schema and cited evidence	gemini-3.1-pro (Google DeepMind, 2026a); qwen3.8-max (Qwen Team, 2026); conditional gpt-4o-0806 (OpenAI, 2024)
Rule Gate	Candidate chain → pass/reject and disagreements; executable clinical and structural checks	Deterministic code
Chain Adjudicator	Surviving chains and cited frames → resolved chain; keep, demote, or drop disputed claims	claude-opus-4-8 (Anthropic, 2026b)

to the endoscopy-sourced impression; pathology-only conclusions remain distinct from what the delivered visual evidence can support.

T6: Guideline-Grounded Clinical Reasoning (Guidelines). The model interprets matched examination images under a named guideline family and produces a clinically grounded decision. Its sectioned response links visible findings and cited frames to sign analysis, guideline reasoning, a committed grade or conclusion, and a recommendation where the family requests one. These are reasoning components of a single task, rather than independent benchmark subtasks. T6 tests how visual observations support the relevant grade or diagnosis, with reference agreement and semantic chain quality read separately.

A.3 Construction engine implementation and prompt contracts

Anatomy recognition model. The Anatomy Router uses an eight-class DINOv3 ViT-B/16 classifier initialized from EndoVLM pretraining weights (Yi et al., 2026), with a linear classification head. The head concatenates the backbone's normalized class token with the mean of its normalized patch tokens, producing one logit for each standardized upper-GI station. It is trained on 4,132 physician-labeled frames from 216 examinations. On the held-out 833-frame test set from 38 examinations, the router achieves 0.8872 overall accuracy and 0.8971 macro-averaged class accuracy. Class accuracy ranges from 0.7969 for the gastric antrum to 0.9885 for the duodenum.

Role implementations and admissibility contracts. Table 3 maps each construction role to its input, output, admissibility rule, and implementation. Models perform bounded extraction, grounding, or chain-generation operations; the controlled specification and deterministic gates determine which outputs enter the benchmark.

The fallback generator gpt-4o-0806 is invoked when a primary chain is absent, fails a gate, or disagrees on a structured sign state. The Chain Adjudicator resolves visual claims while retaining the recorded reference conclusion. The following prompt excerpts preserve model-facing instructions and runtime slots.

A.3.1 Clinical-state extraction

Qwen3-235B-A22B receives the findings-report translation instruction and two constrained extraction interfaces:

Findings-report translation prompt

Translate the Chinese endoscopy findings report into concise clinical English.

Requirements:

- Preserve anatomical sites, abnormal findings, procedures, and negated normal findings.
- Do not add diagnosis or pathology not present in the findings report.
- Keep the output as a report paragraph or site-wise report.

Chinese findings:

`{findings}`

Figure 8: Qwen3-235B-A22B report-translation prompt. The structured findings field is rendered without added diagnostic content.

Site-finding annotation prompt

[English rendering of a Chinese prompt]

You are a senior GI-endoscopy data-annotation expert. The input is the "examination findings" of a gastroscopy report, already segmented by anatomical site. For each site segment extract the ABNORMAL findings; normal descriptions are handled by rules and are not your concern.

1. **Closed term set.** The `term` field must be one of the standard terms below; text with no match goes to `unmapped`.

   `{vocabulary}`

2. **Fields per finding.**
   - `report_index`, `segment_index`: align with the input numbering (required).
   - `site`: the segment's anatomical site.
   - `term`: a closed-set term.
   - `status`: present, or negated when covered by "not seen / not reached / no / denied". Negation scope spans "and / comma": "no ulcer or bleeding" yields two negated objects.
   - `certainty`: certain by default; uncertain when "suspected / consider / possible / cannot exclude / ?" appears.
   - `attributes` (fill whenever the source states them): distribution, count, size, shape, color, border, surface, attachment, surrounding, `sub_position` (lesser/greater curvature, anterior/posterior wall, upper/mid/lower third). "40 cm from the incisors" is a distance, not a size.
   - `treatment_state`: any of [electrocautery, forceps removal, EMR, ESD, APC, clipping, injection, banding] when the lesion was treated in this exam.
   - `biopsy`: true, or omit.
   - `modality`: any of [NBI, magnification, iodine staining, indigo carmine] when stated.
3. Do not output normal descriptions ("mucosa smooth", "vascular pattern clear", "red-and-white coloring", "normal peristalsis", "clear Z-line", "no abnormality", "unremarkable"). A wholly normal segment produces no objects.
4. **Center-specific conventions:** `{style_hint}` `{fewshot}`

5. **Output.** Pure JSON only, two flat top-level lists `findings` and `unmapped` (no nesting). No explanation; when the whole batch is normal return `{"findings":[],"unmapped":[]}`.

Figure 9: Qwen3-235B-A22B site-finding extraction prompt. Each site segment is mapped onto the closed term set, with negation, certainty, and stated attributes retained.

Diagnosis and conclusion annotation prompt

[English rendering of a Chinese prompt]

You are a senior GI-endoscopy and pathology data-annotation expert. The input is one exam's ENDOSCOPIC diagnosis conclusion and PATHOLOGY diagnosis conclusion. Extract every disease entity with a definite diagnostic meaning and structure it.

1. **Closed term set.** `term` must come from the standard diagnosis names below; unmatched text goes to `unmapped`.
   `{vocabulary}`

2. **Fields per entity.**
   - `source`: endoscopy or pathology (required; the same name on both sides is output twice).
   - `term`: the closed-set standard (parent) diagnosis.
   - `modifier`: e.g. "with erosion / bleeding / intestinal metaplasia / bile reflux / enlarged folds". Parent and modifier are split: "chronic atrophic gastritis with intestinal metaplasia" → `term=chronic atrophic gastritis`, `modifier=with intestinal metaplasia`. Omit if none.
   - `location`; `severity` (mild/moderate/severe/focal); `grade`; `stage` (A1/H1/S1, C-1/C-2, active, post-op, recurrent); `certainty` (certain, or uncertain on "consider/possible/suspected/?"); `hp_status` (positive/negative only when the entity itself states Hp status). Omit empty ones.

3. Do NOT extract: procedural acts (biopsy, sampling, staining, resection); bare recommendations (suggest follow-up / EUS / surgery); structured pathology gradings of the form "chronic inflammation (++) activity (+) atrophy (-) metaplasia (+) HP (-)" (parsed programmatically) — but natural-language pathology such as "severe atrophy" IS output; negative results ("HP(-)", "no tumor seen"); bare site names.

4. **Key rules:** the endoscopic conclusion is extracted from endoscopy text only and must NOT be upgraded using pathology (endoscopy "polyp" → gastric polyp, not hyperplastic polyp). Pathology "chronic superficial gastritis" → chronic non-atrophic gastritis; "chronic active inflammation" → chronic active gastritis. "HP(+)" → `term=Hp infection`, `source=pathology`, `hp_status=positive`. Entities sharing a site in one sentence inherit it. Post-op states are entities too.

5. **Output.** Pure JSON only, two flat top-level lists `entities` and `unmapped`, every entity carrying `report_index`. No explanation, no nesting.

Figure 10: Qwen3-235B-A22B diagnosis-extraction prompt. Endoscopic and pathological entities are extracted separately and never upgraded across the two sources.

A.3.2 Evidence verification

The Evidence Grounder judges the supplied claim or answerability from displayed frames, emits schema-constrained JSON, and treats missing or uncertain evidence as unsupported. Qwen3-VL-235B-A22B-Instruct-FP8 receives the compare-only grounding interface. The single-site form

extends to multi-site cases by attaching the predicted station to each frame while preserving the support decision and ranked-frame output:

Compare-only grounding prompt

[English rendering of the Chinese compare-only prompt]

You will see `{n}` images from the [`{site}`] station of one gastroscopy, numbered 1-`{n}`.

Report description for this station:
"`{text}`"

Examination impression (context only):
"`{impression}`"

1. Decide whether these images provide sufficient visual support for the report description.
2. If supported, rank the supporting frames from highest to lowest confidence and give the corresponding per-frame confidence (high, medium, or low).

Rules:

- Judge the report description from the images. The impression must not make a description supported.
- Evaluate diffuse changes as well as discrete lesions. For diffuse changes, rank readability rather than severity and set `` `uniform` `` true.
- If unsupported, return empty frame and confidence lists and set `` `best` `` to null.
- `` `best` `` is the first frame in `` `supported_frames` ``; omit uncertain frames.

Return JSON only:

```
{"supported": true,
 "supported_frames": [<numbers>...],
 "per_frame_confidence": ["high|medium|low"...],
 "best": <number>, "uniform": false, "note": "<optional>"}
```

Figure 11: Qwen3-VL-235B-A22B-Instruct-FP8 compare-only grounding prompt. The report-derived description is judged against anatomically routed frames, with the examination impression admitted as context only.

A.3.3 GUIDELINE-CHAIN CONSTRUCTION

The two primary T6 generators and the conditional fallback receive the same case-specific generation interface. The established conclusion fixes the requested output label but cannot serve as evidence. The derivation must cite supplied frames and follow the card's clinical criteria:

T6 graded chain-generation prompt

[Dynamically assembled English generator prompt]

You are a senior gastrointestinal endoscopist analyzing upper-GI endoscopy frames under the endoscopic diagnostic guideline for `{disease}`.

[Guideline]

```
{guideline_text}
```

[Frame map]

```
{frame_ids_and_supplied_stations}
```

[Reasoning policy]

```
{shared_visual_evidence_rules}
```

`{family_specific_rules}`

[Established fact] The final conclusion for this case is: `{answer}`. This is given. Do not use it as a premise or as evidence. Point out the specific findings in the images that support it and show how the guideline rules make it hold. Never reason "because the conclusion is X, therefore ...". In [Conclusion], write exactly `` `{answer}` `` and put all explanation in [Guideline reasoning].

`{optional_confirmed_or_report_derived_signs_and_criteria}`

[Answer space] In [Conclusion], give exactly one of `{answer_space}`.

Answer in the requested sections, in order:

`{layer_instructions}`

Each visible finding must cite its supporting frame IDs and supplied station. Keep raw appearance, sign analysis, guideline reasoning, conclusion, and recommendation non-redundant. The retained class must rest on its own guideline-defining visual criterion; report context, the supplied conclusion, and exclusion of alternatives cannot fill an evidence gap. State exact measurements only when a calibrated reference or reliable guideline landmark is visible.

Figure 12: T6 generation prompt shared by gemini-3.1-pro, qwen3.8-max, and the conditional gpt-4o-0806 fallback. The established conclusion fixes the output label but is barred from serving as evidence.

Algorithm 1: Rule Gate for a T6 candidate chain

Input: candidate chain, case evidence, executable guideline card, and legal answer space.
Verify: (1) card applicability and defining visual support; (2) output schema and cited-frame existence; (3) anatomical compatibility and legal grades; (4) score arithmetic where applicable and requested-step linkage; (5) coherence between sign analysis and conclusion.
Output: reject invalid chains; retain admissible chains and a structured agenda of disagreements for evidence-level adjudication.

claude-opus-4-8 receives the surviving readings, cited frames, card, conclusion anchor, and generated disagreement agenda. Its interface requires one resolved chain and a keep, demote, or drop ruling for every disputed claim:

T6 adjudication prompt

[Dynamically assembled English adjudication prompt]

You are a senior gastrointestinal endoscopist adjudicating `{count}` independent readings of the same upper-GI examination.

[Established fact] The conclusion is: `{answer}`.

`{confirmed_or_report_derived_reference}`

Do not re-judge that conclusion or score the readings. Check each finding against its cited frame and rule on the generated disagreement agenda. Every disputed claim receives one disposition:

- **keep:** the frames support it; name the settling frame.
- **demote:** the frames are equivocal; mark the sign indeterminate.
- **drop:** the cited frames do not show the claim.

`{shared_and_family_specific_policy}`

Do not cite an unavailable frame or reason backwards from the conclusion. In [Conclusion], write exactly `{answer}`; put the derivation in [Guideline reasoning].

```
{frame_map}
{guideline}
```

Candidate readings

```
{readings}
```

What to rule on

```
{agenda}
```

What to write

Return one chain in the same requested sections as the generators, written as your own reading rather than as a summary, followed by [Adjudication]. First record:

```
applicability: <supported frame(s)|insufficient> --- <visual prerequisite>
defining_support: <supported frame(s)|insufficient> --- <retained class criterion>
```

Then use only the supplied sign IDs:

```
sign <sign_id>: <keep|demote|drop>
  (<present|absent|not_demonstrated|indeterminate>) <frame> --- <reason>
discarded: <unsupported claim> --- <reason>
```

Figure 13: claude-opus-4-8 adjudication prompt. The surviving readings are ruled on one generated disagreement at a time, with keep, demote, and drop as the only dispositions.

A.4 BENCHMARK COMPOSITION AND PAIRED ANALYSIS PANELS

Benchmark composition. Table 4 summarizes the 9,106 main-board items from 4,893 distinct examination identifiers and their 70,386 item–frame references. The main board samples six clinical stages for clinical breadth; task-specific examination counts overlap and are not additive. The diagnostic axes reuse eligible examinations from this cohort under separate paired protocols, and their condition variants are not additional main-board items.

The task counts arise from different reference and eligibility paths. T1 is a curated subset of physician-annotated frames: within-examination near-duplicate views are reduced, clear images are retained, and sampling balances the anatomical classes. T2 and T3 instantiate questions from normalized report findings and attributes that satisfy their respective visual-support requirements. T4 has a candidate report for every examination. Its 600 examination-level items are sampled after separately dividing examinations into four image-count strata and four finding-count strata, so the selected set spans both visual-context length and finding burden. T5 retains endoscopy-sourced impressions with compatible visual evidence and a legal answer-space mapping. T6 retains cases with a recorded grade or pathology conclusion, applicable guideline card, cited visual support, and an admissible reference chain after generation, rule checking, and evidence-level adjudication. Appendix A.5 identifies the reference field and evidence requirement for each task.

Table 4: **TRACESCOPE composition.** The main board contains 9,106 items from 4,893 distinct examination identifiers and 70,386 item–frame references. Referenced-frame counts precede model-specific serving limits. The diagnostic axes reuse eligible examinations under paired conditions outside this total. T2 contains 2,200 binary, 641 single-choice, and 926 select-all items; T3 contains 460 single-choice and 435 open quantitative items. Of 635 T6 chains, 633 enter the main grade population. The main board has no train/validation/test partition.

Layer	Input tier	n	Referenced frames per item		Reference source
			Median	Max	
T1 anatomy recognition	single image	1,572	1	1	clinician station annotation
T2 abnormal findings	matched multi-image	3,767	3	3	grounded report finding
T3 finding attributes	matched multi-image	895	3	3	grounded report attribute
T4 findings report gen.	redundant multi-image	600	32	396	structured findings report
T5 diagnostic impression	matched multi-image	1,637	16	24	endoscopy impression
T6 guideline reasoning	matched multi-image	635	8	10	grade/pathology and card
Total	three still-image tiers	**9,106**	—	—	item-level provenance

Evidence Presentation panel. The panel contains 193 examinations with an image-compatible diagnostic impression, a selected multi-image set, a retained examination image set, and video-level evidence for all three sampled tiers. One fixed four-choice T5 diagnostic-impression question is constructed per examination and evaluated under Selected Multi-Image, Retained Multi-Image, and uniformly sampled video-frame presentations at 128, 256, and 512 frames (Video-128, Video-256, and Video-512). The question, options, reference impression, prompt structure, and scoring rule are fixed within each five-condition group. The panel therefore contains 193 unique targets and 965 case-condition evaluations per model. Selected Multi-Image is a presentation selected from the examination, not a frame-level diagnostic-evidence reference. Two auxiliary video-frame presentations match, case by case, the selected and retained image counts; they are analyzed as count controls rather than additional primary conditions.

Diagnostic Cascade panel. Within the benchmark cohort, all 1,220 examinations with usable endoscopic visual evidence and a pathology-verified binary *H. pylori* infection label form the eligible Diagnostic Cascade panel (459 positive, 761 negative). The 19 locally served models evaluate each case under five cumulative states: Images, + Anatomy, + Findings, + Sign Polarity, and + Decision Threshold. The physician-supervised Anatomy Router supplies anatomical locations; report extraction provides the finding state, supplemented and checked through targeted clinician annotations during guideline-chain construction. The later states add the sign-polarity mapping and its decision boundary. The images, question, pathology-backed target, answer space, prompt structure, and scoring rule remain fixed. Every examination remains in every intention-to-treat denominator; empty, invalid, or off-vocabulary outputs score as incorrect.

Table 5: **Reference and evidence matrix.** The reference field, delivered evidence, and scoring unit remain linked at item level.

Task	Reference source	Delivered evidence / support policy	Evaluation unit
T1	Physician station annotation	Single annotated frame	Anatomical site
T2	Normalized report finding	Matched frames; visible finding support	Finding state
T3	Normalized report attribute	Matched frames; queried attribute determinable	Attribute value
T4	Structured findings report	Examination-level image set	Site–finding units
T5	Endoscopy-sourced impression	Matched frames; image-compatible diagnosis	Diagnostic impression
T6	Recorded grade or pathology and executable card	Matched frames; cited support and rule checks	Grade and reasoning units

The *H. pylori* component of the main-board T6 task and the Diagnostic Cascade both draw from the source examination cohort. T6 candidates are sampled for coverage across phenotype and guideline-sign patterns, including positive and negative states. A candidate is instantiated as a T6 item when generation, rule checking, and evidence-level adjudication produce an admissible guideline-reasoning reference. All 168 reference-valid, coverage-oriented *H. pylori* T6 items also occur in the 1,220-case Diagnostic Cascade panel.

A.5 Reference standards and evidence validation

Reference sources and evidence requirements differ by task (Table 5). Source validity asks whether the target is recoverable from an identified record field; evidence validity asks whether the delivered images support a visual target; scoring validity asks whether an equivalent model response is recognized in the same canonical unit. These checks connect report provenance to task construction and evaluation.

Clinical experts reviewed sampled grounding outputs. Their decisions were incorporated into visual-support labels through support, rejection, or correction. Construction decisions remain linked to their source records and evidence references. The multi-center cohort is described in Appendix E; translation and frame-answerability procedures are detailed in Appendix A.3.

B Evaluation protocol

This appendix specifies model identities and serving, input delivery, task prompts, deterministic metrics, paired statistics, semantic rubrics, and the physician reader study. Per-column metric definitions appear in Section B.4.

B.1 Evaluated models and serving configuration

The evaluated fleet contains 19 locally served open-source checkpoints and three proprietary API models. Table 6 gives each model's backend, context window, image cap, and generation budget. The construction models (Appendix A.3) and the text-only evaluation judge, GPT-5.5 (OpenAI, 2026b), are separate from the evaluated fleet.

GPT-5.4. GPT-5.4 (OpenAI, 2026a) is evaluated through API build `gpt-5.4-0305-global` with interleaved image–text input.

Claude Opus 4.7. Claude Opus 4.7 (Anthropic, 2026a) is evaluated through API build `claude-opus-4-7` with interleaved image–text input.

Gemini 3.8 Flash. Gemini 3.8 Flash (Google DeepMind, 2026b) is evaluated through API id `gemini-3.8-flash` with interleaved image–text input.

Qwen3-VL (4B/8B/32B). We evaluate the dense Qwen3-VL Instruct checkpoints at 4B, 8B, and 32B (Bai et al., 2025) using their multi-image input path.

Qwen3.5 (4B/9B/27B). We evaluate the dense Qwen3.5 checkpoints at 4B, 9B, and 27B (Team, 2026) on the multi-frame image-list path, with thinking mode disabled. Table 6 gives their context budgets.

InternVL3.5 (8B/14B). InternVL3.5-8B and -14B (Wang et al., 2025b) use one dynamic visual tile per frame within the declared context budget.

MiniCPM-V-4.5 (8B). MiniCPM-V 4.5 (8B) (Yu et al., 2026) uses its supported multi-image path and fast-thinking chat-template setting.

VideoChat-Flash-7B. VideoChat-Flash-7B (Li et al., 2026a) uses its native video entry point rather than the vLLM image-list path.

LLaVA-NeXT-Video-7B. LLaVA-NeXT-Video-7B (Zhang et al., 2024) is evaluated with its DPO checkpoint on the Vicuna-7B-v1.5 base and native video processor.

Hulu-Med (4B/7B). Hulu-Med-4B and -7B (Jiang et al., 2025) use their supported Transformers backend; their respective language backbones are Qwen3-4B and Qwen2.5-7B.

MedGRPO-7B. MedGRPO-7B (Su et al., 2026) is evaluated as `uAI-NEXUS-MedVLM-7B-RL` on a Qwen2.5-VL-7B base.

CE-R1 (7B). CE-R1 (Chen et al., 2026) uses the Deep image checkpoint, a rank-8 LoRA (α=16) on a Qwen2-VL-7B base.

ColonR1 (3B). ColonR1 (Ji et al., 2025) is evaluated as a full-parameter Qwen2.5-VL-3B fine-tune.

HealthGPT-Pro-8B. HealthGPT-Pro-8B (Lin et al., 2025) is evaluated on its Qwen3-VL-8B backbone.

Lingshu (I-8B/32B). Lingshu-I-8B and -32B (Xu et al., 2026) use InternVL3-8B and Qwen2.5-VL-32B backbones, respectively; only I-8B uses the InternVL tile setting.

B.2 Input delivery and diagnostic-axis protocols

Table 6 reports the serving backend, context length, per-prompt image cap, and generation budget for each model. Decoding uses temperature 0 and top-p 1.0. The following rules define how visual inputs reach each backend.

Image delivery. Each frame is downscaled to a 525×490 (257,250-pixel) budget before serving. If an examination exceeds a model's per-prompt frame cap, the harness selects frames evenly across the sequence with both endpoints retained. Table 6 records the caps and model-specific context budgets. T4 items reference a median of 32 and up to 396 retained frames. On the 64-frame image-list paths, endpoint-preserving uniform sampling yields a median of 32 and a maximum of 64 served frames; the served count for an item with n referenced frames is $\min(n, 64)$.

Diagnostic-axis input delivery. Evidence Presentation serves the selected multi-image set, the retained examination image set, and uniformly sampled video-frame presentations at 128, 256, and 512 frames for each of its 193 eligible examinations. Two additional conditions sample the same number of video frames as the selected and retained still-image sets for each case (Video-Match-N and Video-Dense-N). All nine models are evaluated under these seven presentation conditions, with delivery adapted to their supported visual interface. The T5 diagnostic question, options, target, prompt structure, scoring rule, and greedy 16-token decoding protocol are identical across them; all 12,159 model responses are decoded. Diagnostic Cascade serves the same images to 19 locally served models across five cumulative clinical-state conditions: Images, + Anatomy, + Findings, + Sign Polarity, and + Decision Threshold. Its 1,220-case eligible *H. pylori* cohort and pathology-backed reference criterion are defined in Appendix A.4.

Video-derived presentation delivery. Each video-derived condition is constructed by uniformly sampling the source examination video. Eight models receive the sampled frames as ordered image lists; MedGRPO receives the same sampled sequence through its native-video processor. This preserves the presentation content and prescribed frame count while using each model's supported visual interface.

Attention concentration. The recorded last-token attention weights are normalized over the visual units delivered to each model and presentation. The share90 measure is the smallest fraction of these units whose cumulative attention reaches 90%; normalized entropy summarizes how evenly attention is distributed over them. MedGRPO merges frame pairs, so one pair is its visual unit, whereas the other models use frames. Both measures describe concentration over the delivered input. They

Table 6: **Per-model evaluation configuration.** Serving backend and per-model budgets for the 19-model fleet: *Ctx* the image-tier `max_model_len`, *Img* the `max_images` cap, and *Gen* the `max_tokens` budget. The final column gives the Axis 1 delivery interface, maximum sampled-frame tier, and raised context after "@" for the nine-model panel (L: ordered image list; V: native-video processor). Every frame is downscaled to a 525×490 pixel budget and decoding is uniform (temperature 0, top-p 1.0). Notes: [a]Hulu-Med served via HF transformers (a vLLM assert); [b]InternVL3.5 pinned to `max_dynamic_patch:1`; [c]Lingshu-I-8B inherits the InternVL processor, pinned `crop_to_patches:false`—Lingshu-32B is a Qwen2.5-VL checkpoint, routes to a different processor and needs no such pin; [d]CE-R1 is a rank-8 LoRA (α=16) on Qwen2-VL-7B; [e]VideoChat-Flash served via its own entry point, video-native; [f]MiniCPM-V-4.5 pinned to `enable_thinking:false`, its fast-thinking branch; [g]Qwen3.5-4B/9B/27B likewise run with thinking disabled. An em dash denotes a model outside the Axis 1 panel.

Model	Serving	Ctx	Img	Gen	Axis 1 input
Qwen3-VL-4B	vLLM	32,768	64	1024	L/512 @ 131,072
Qwen3-VL-8B	vLLM	32,768	64	1024	L/512 @ 131,072
Qwen3-VL-32B	vLLM	32,768	64	1024	—
Qwen3.5-4B[g]	vLLM	32,768	64	1024	L/512 @ 163,840
Qwen3.5-9B[g]	vLLM	32,768	64	1024	L/512 @ 163,840
Qwen3.5-27B[g]	vLLM	32,768	64	1024	L/512 @ 163,840
InternVL3.5-8B[b]	vLLM	32,768	64	1024	—
InternVL3.5-14B[b]	vLLM	32,768	64	1024	—
MiniCPM-V-4.5 (8B)[f]	vLLM	32,768	64	1024	—
Hulu-Med-4B[a]	HF transf.	—[a]	64	1024	L/512
Hulu-Med-7B[a]	HF transf.	—[a]	64	1024	L/512
HealthGPT-Pro-8B	vLLM	32,768	64	1024	L/512 @ 131,072
Lingshu-I-8B[c]	vLLM	32,768	64	1024	—
Lingshu-32B	vLLM	32,768	64	1024	—
MedGRPO-7B	vLLM	32,768	64	1024	V/512 @ 32,768
ColonR1 (3B)	vLLM	32,768	64	1024	—
CE-R1 (7B)[d]	vLLM+LoRA	32,768	64	2048	—
LLaVA-NeXT-Video-7B	vLLM	8,192	64	512	—
VideoChat-Flash-7B[e]	custom HF	—	—	—	—

do not identify which units contain clinically decisive evidence, and we compare changes within a model rather than rank attention levels across architectures. Table 15 reports these descriptors for the primary presentations.

Hulu-Med uses its Transformers backend with the common frame and pixel budgets. InternVL3.5 and Lingshu-I-8B use one visual tile per frame; Lingshu-32B uses its Qwen2.5-VL processor. MiniCPM-V-4.5 uses its fast-thinking template. Exact settings are listed in Table 6.

B.3 Task prompts and response forms

Each task is posed under its canonical question form; the answering instruction and the closed/open option layout per form are given verbatim below, and the question column of Table 21 illustrates the corresponding item forms. For Evidence Presentation, the T5 diagnostic-impression question, answer space, target, and answer-format instruction remain fixed while only the delivered visual context changes. For Diagnostic Cascade, the binary *H. pylori* question, pathology-backed target, visual evidence, and answer space remain fixed while designated upstream clinical states and guideline information are added cumulatively.

T2 question forms. T2 comprises normal-state judgment, single-finding presence, and finding enumeration. Its 3,767 items include 642 normal-state and 1,558 presence true-or-false (*TF*) questions, 641 single-choice finding-enumeration questions (*MCQ-S*), and 926 select-all finding-enumeration questions (*MCQ-M*). The two TF forms and MCQ-S use single-option accuracy; MCQ-M uses exact-set agreement. The three single-answer blocks are also scored separately as *TF-Norm*, *TF-Pres*, and *MCQ-S-Find* before an item-count-weighted combination. The two TF blocks have two-option answer spaces, for which a constant-answer floor is reported alongside chance.

In the expanded results table, *SA* denotes open short-answer responses, *RG* denotes findings-report generation, and *CoT* denotes the sectioned reasoning-chain response format used for T6.

Evaluation and LLM-judge scoring. At evaluation the item stem is sent with a task-specific answer-format instruction. Closed forms append one of two fixed instructions; T6 appends a sectioned answer format whose sections are selected per guideline (not every family uses all five), with the frame vocabulary fixed to the frames actually supplied. The main-board format names the guideline and enumerates its signs and permitted answer tokens where specified by the family card, without supplying its decision criteria (four of the six T6 families enumerate signs, and all six enumerate answer tokens). Axis 2 preserves the question while adding clinical information in layers: + Sign Polarity supplies the diagnostic direction of each sign, and + Decision Threshold adds the net-count decision boundary.

Answer-format instruction

 # Closed-form answer instructions

- **MCQ-S/TF:** Reply with a single option letter only. Do not explain.
- **MCQ-M:** Reply with the letters of every option that applies, comma-separated, e.g. "A, C". Do not explain.

 # T6 main-board answer format (closed-book; sections and per-family exposure set by the card)
Guideline: `{guideline_name}`

This case has `{count}` frames, labeled F1-F`{count}`; cite only these. The rule text and its quantitative thresholds are never supplied; the sign inventory and the permitted grade tokens are enumerated where the guideline's card fixes them, and left to the model's own clinical knowledge otherwise.

Answer in the following sections, in order:

- **[Visible findings]** Describe what you see, citing frame numbers (e.g. `{example}`).
- **[Sign analysis]** Go through the signs below, one line each, "sign name: present/absent (frames)", where the guideline enumerates them; otherwise assess the relevant signs and cite the supporting frames.
- **[Guideline reasoning]** State and apply the named guideline's rules step by step.
- **[Conclusion]** Give one committed final grade or diagnosis, written verbatim as exactly one of the guideline's permitted tokens where it enumerates them.
- **[Recommendation]** State the next clinical step, one per line.

Figure 14: Answer-format instruction. Closed forms append one of two fixed instructions; T6 appends the sectioned format, whose sections and frame vocabulary are selected per guideline.

The text-only judge interface receives one fixed input per item: the task prompt exactly as the model received it, the reference, and one candidate output—plus, for T6, the family's guideline card, adapter row and answer space. It compares the candidate with the reference, uses the card to interpret family-specific criteria, and credits equivalent meaning rather than lexical identity. Dimensions, N/A conditions, failure flags and the output schema are supplied by the rubric in Appendix B.5; the shared call skeleton is reproduced below.

Judge call skeleton

```
{rubric: dimensions, family adapter, prompt-side protocol, output schema}
```

 # Case
(## Guideline card, ## Family adapter and ## Answer space appear for T6 only.)
Task prompt — exactly as the model received it

```
{the closed-book task prompt exactly as the model received it: option list, option
 letters, T6 answer format and frame vocabulary included}
```

Reference

```
{the reference: the recorded grade or pathology-backed conclusion (authoritative),
 plus the advisory per-sign reference and the lesion ledger where recorded}
```

Guideline card

```
{the T6 family's card: sign inventory, finding-to-sign maps, applicable boundaries,
 interpretation bands, recommended actions}
```

Family adapter

```
{the family's adapter row: entry criteria - decisive evidence - confirmation gate}
```

Answer space
Conclude with exactly one of: `{the permitted grade or diagnosis tokens}`
Candidate — `{candidate label}`

```
{one candidate output, blinded to the model's identity}
```

 # What to return
Return only the JSON object specified in the rubric — no prose, no markdown fences.

Figure 15: Judge call skeleton. The task prompt, the reference and one blinded candidate are assembled per item; images are never shown to the judge.

B.4 Deterministic metrics and statistical analysis

Clinical scoring units. Closed-form tasks are scored by exact or set match against each item's randomized option list. Unless a column is explicitly set-valued or numeric, *Acc* denotes exact option-match accuracy over its stated population. Quantitative T3 responses are parsed with unit constraints and a declared tolerance. T4 reports are scored through the terminology layer (§3.1) for recovered sites, findings, and unsupported content. T6 is scored by extracting whether a legal guideline label is stated and whether it matches the reference. Text-overlap metrics are retained as wording-sensitive controls for T3, T4, and T6. Reference-defined metrics use declared item populations, with any model-specific exclusions reflected in the corresponding denominators. Prediction-conditioned diagnostics, including hallucination density and agreement over extractable calls, carry their per-model denominators and are not treated as fully paired rankings. The main-board T6 exact-grade population contains 633 items, with 628 evaluable replies for ColonR1; these denominators are reported with the corresponding cells. T3 contains 460 categorical single-choice items and 435 quantitative items, for 895 total. Four of the single-choice items belong to two categories with fewer than ten items each; they remain in the benchmark count but are omitted from the three grouped categorical scores, whose denominators are 206, 185, and 65 (456 items in total).

Prior-only diagnostic controls. Each closed-form column has a matched-chance reference computed from its item-level option counts. The prior-only control uses a fixed response independent of image content: the plurality category for categorical columns and fixed values for quantitative columns. Its scores characterize answer-frequency structure under each metric.

Multi-answer exact match and F1. For item i, let P_i and G_i be the predicted and gold option sets. The reported exact-set rate (Ex) and macro-averaged set F1 are

$$\mathrm{Ex} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[P_i = G_i], \qquad \mathrm{F1} = \frac{1}{N}\sum_{i=1}^{N} \frac{2|P_i \cap G_i|}{2|P_i \cap G_i| + |P_i \setminus G_i| + |G_i \setminus P_i|}.$$

Option order is ignored. A parse failure is represented by an empty prediction set and receives zero on both measures while remaining in the denominator.

Open numeric accuracy. Numeric answers are converted to common units and represented as intervals. For a predicted interval $p = [p_l, p_u]$ and gold interval $g = [g_l, g_u]$, define their non-overlap distance and the item-level relative error as

$$d(p, g) = \max\{0,\, g_l - p_u,\, p_l - g_u\}, \qquad e_i = \max_{g \in G_i} \min_{p \in P_i} \frac{d(p, g)}{(g_l + g_u)/2}.$$

Thus overlapping intervals have zero error, every gold measurement must be matched, and surplus predicted measurements are ignored. A missing prediction has $e_i = \infty$; a zero-valued gold is correct only when the prediction also contains zero. Accuracy at tolerance τ and mean relative accuracy (MRA) are

$$\mathrm{Acc}_\tau = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[e_i \le \tau], \qquad \mathrm{MRA} = \frac{1}{|\mathcal{T}|}\sum_{\tau \in \mathcal{T}} \mathrm{Acc}_\tau, \quad \mathcal{T} = \{0.1, 0.2, 0.3, 0.5\}.$$

Size and count are evaluated on separate populations with the same rule.
For T1, Site and Organ are separate single-choice accuracy columns on their own populations ($n = 1{,}146$ and $n = 426$, respectively). For T3, Position&Dist., Morphology, and Color are categorical single-choice accuracy columns with $n = 206$, 185, and 65, respectively; Size and Count use the unit-aware numeric rule above.

Additional readouts in the expanded board. Table 9 separates T2 Normal-state judgment, finding presence, and single-choice finding enumeration; multi-answer finding F1 and exact-set accuracy remain separate. For T3, *MRA* averages Acc_τ over the four thresholds above. T4 *Site* and *Find* are per-item macro-averaged fractions of reference stations and findings recovered. *Hal* is the fraction of asserted findings absent from the reference, averaged over reports with at least one asserted finding; *Fab* is the rate of reports asserting an abnormal finding on the 45-item normal-reference subset. BLEU-4, R-L (ROUGE-L), and METEOR measure lexical overlap; BS_{rbt} is BERTScore-F1 with RoBERTa-large embeddings and BS_{bio} is BERTScore-F1 with BioBERT embeddings. These text metrics compare the generated text with the reference and complement the clinical-content measures. For T6, *Stated* is the share of replies yielding a recoverable grade, *Ex|st* is exact-grade accuracy conditional on stating one, and *Exact* is exact-grade accuracy over all items, so $\mathrm{Exact} = \mathrm{Stated} \times \mathrm{Ex|st}$. The SEFR and ELRC columns are defined in Appendix B.5.

Paired comparisons. Each diagnostic-axis contrast compares the same examinations under different input conditions, retaining an examination's paired responses together when resampled. Evidence Presentation reports T5 diagnostic-impression accuracy and four prespecified adjacent changes across its five primary presentations on the 193-case panel. Model-wise intervals use 10,000 paired bootstrap draws over examinations. For each fleet replicate, the 193 examinations are resampled within each of the nine models and the model-wise changes are summarized by a trimmed mean that drops the minimum and maximum; Holm correction covers the four fleet transitions. The two count-matched video conditions supply four additional descriptive paired contrasts, with model-wise exact McNemar tests available in the frozen analysis.
Diagnostic Cascade reports Cohen's κ as its primary chance-corrected endpoint and balanced accuracy, sensitivity, and specificity as complementary binary-decision readouts. All 1,220 cases remain in the intention-to-treat denominator. Its 10,000 bootstrap replicates resample examinations within pathology class, using the same draw across states and models. Holm correction covers the four prespecified fleet-level adjacent $\Delta\kappa$ tests. The deterministic comparison applies the clinician-specified net-count rule to the consolidated sign states, providing a reference operating point on the same cases. We report 95% intervals for paired effects and keep model-wise and family-wise responses visible alongside fleet summaries.

B.5 SEMANTIC RUBRICS FOR GENERATIVE TASKS

The construction Chain Adjudicator resolves T6 reference chains before evaluation. The text-only evaluation judge instead scores blinded model outputs against those fixed references and does not alter benchmark targets. Its scores complement the deterministic metrics.

The two generative layers carry prespecified five-band rubrics—SEFR (structured endoscopic finding report) for T4 and ELRC (endoscopic long reasoning chain) for T6. Every judge call uses a fixed triple: the task prompt exactly as the model received it, the reference, and one candidate output—plus, for T6, the family's guideline card, adapter row and answer space. The judge compares each candidate with the fixed reference using the card to interpret family-specific criteria. The images are withheld, so its output measures *reference/card consistency*. Three design rules follow from that constraint:

- **Division of scoring.** Deterministic extraction computes measurement accuracy, repetition, and grade arithmetic. The text-only judge scores semantic dimensions not captured by those rules; image-dependent judgments such as artefact versus lesion and frame obscuration are outside its input.
- **Named bands.** The judge assigns one of five named bands per dimension without returning weights or totals. Code-side aggregation maps the bands to $100/80/60/40/10$ (Table 7); per-dimension means accompany composite scores in Appendix C.2.
- **Separate failure flags.** Table 8 defines safety, alignment, reference-quality, and validity flags outside dimension means. The safety group supplies a non-compensatory per-item endpoint; its model-level rates and recommendation coverage appear in Appendix C.2.3.

SEFR, the report rubric (T4). SEFR scores what a candidate findings report recovers of the reference report at four dimensions (Table 7): station coverage and spatial localization; lesion entities with the reference's mucosal abnormalities, pertinent negatives and unsupported assertions; grounded specificity; and report structure with containment of diagnostic inference. The entity dimension is anchored on an explicit primary/secondary lesion ledger the judge records per item, and omission of a high-risk lesion identified in the reference is its failing case. The anti-elaboration rule aligns specificity credit with reference-supported content: unsupported qualifiers earn no specificity credit and are charged under the entity dimension. Physician-written reports and model outputs are interleaved blind on the same study items under this rubric.

ELRC, the chain rubric (T6). ELRC scores a candidate chain at five dimensions (Table 7): factual and landmark fidelity, including applicable quantitative boundaries; agreement with the recorded grade or pathology-backed conclusion; recommendation policy consistency, judged conditionally on the grade the candidate itself committed—a wrong grade followed by the action the card prescribes for that wrong grade is not double-penalised; deductive rigour, which requires the retained class to be positively proved from defining signs rather than asserted because alternatives were not seen or because supplied text states it; and uncertainty calibration, which asks whether claim strength matches the candidate's own stated evidence. Patient-level safety is read separately, per item, as the `clinical_action_safety` endpoint—safe, unsafe escalation, unsafe reassurance, or not applicable—against the reference condition and the card's confirmation gates. This endpoint remains outside dimension means. Two support rules move known ambiguities out of the band decision rather than into it. First, an explicitly indeterminate conclusion is credited only where the reference grade is itself indeterminate; the commitment status is recorded separately in the `grade_extraction` field described under the output schema below. Second, the engine-annotated per-sign reference is an advisory cue, while the recorded grade or pathology-backed conclusion remains authoritative. Where a candidate coherently disagrees with the cue while matching the reference conclusion, the judge prefers the higher band and raises `reference_conflict`.

The scoring protocol is fixed in the prompt. Both rubrics run under the same prompt-side rules, making each band decision reproducible from the fixed instrument. The judge assigns exactly one of the five named bands per dimension, or sets `na` where that dimension's stated N/A condition holds, and emits no number, weight, total or cap. It must not default to *strong*: a dimension the candidate does not address takes the corresponding lower band. It judges meaning rather than phrasing, so a clinically equivalent wording has recovered the content; it does not re-decide the case, recording `reference_conflict` instead of adjusting a band when it believes the reference is wrong; and it scores nothing the rubric excludes by design—measurement accuracy, repetition, intervention performance, or image-dependent judgments. Each dimension carries one evidence span of at most 25 words, quoted or closely paraphrased from the candidate, which is what makes a contested

Table 7: **The two text-only judge rubrics for the generative layers.** Every judge call takes the triple of the task prompt exactly as the model received it, the reference, and one candidate output (for ELRC, also the family's guideline card, adapter row and answer space); the judge never sees the images, so everything it emits measures reference/card consistency, not visual truth. The judge assigns one of five named bands per dimension—*exemplary*, *strong*, *acceptable*, *deficient*, *failing*—mapped to the percentile scale $100/80/60/40/10$ in code; the prompt emits no numbers, weights or caps. Weights renormalise over the dimensions whose N/A condition does not hold: SEFR's D1–D2 form its objective-alignment track (0.65) and D3–D4 its quality track (0.35); ELRC splits $0.60/0.40$. Full band criteria and the family adapter table are released with the rubric.

Rubric	Dimension (weight)	What the five bands score	N/A when
SEFR (T4 report)	D1 — station coverage & spatial localization (0.30)	Whether the candidate documents the stations the reference documents and anchors each focal lesion to the same sub-anatomical landmark or wall.	never
	D2 — lesion entity, pertinent negatives & reference-unsupported assertion (0.35)	Recovery of the reference's focal lesions and mucosal abnormalities, affirmation of stations the reference calls normal, and absence of asserted content the reference does not support; the failing anchor is the omission of a high-risk lesion, read from the reference's own wording.	never
	D3 — grounded specificity (0.20)	Recovery of the reference-stated qualifiers of each abnormality (color, texture, margin, extent); unsupported qualifiers earn no credit.	the reference states no attribute-level qualifier anywhere
	D4 — report structure & inference containment (0.15)	Station-wise reporting order and containment of diagnostic inference to what the reference states at that station.	never
ELRC (T6 chain)	D1 — factual, landmark & threshold fidelity (0.30)	Consistency of asserted sites, lesion identities and defining signs with the reference; correct use of applicable quantitative boundaries.	never
	D2a — grade equivalence (0.15)	Agreement of the committed grade or diagnosis with the recorded grade or pathology-backed label, within the item's answer space.	never
	D2b — recommendation policy consistency (0.15)	Whether the recommendation matches the card's action for the candidate's own grade, and is proportionate to its expressed certainty. Policy consistency, not patient-level safety; safety is the separate endpoint.	the candidate commits no grade and offers no recommendation, or the card specifies no action
	D3 — deductive rigour & fallacy resistance (0.25)	Whether the retained class is positively proved from defining signs, entry criteria are checked, and the chain is free of exclusion-only reasoning and of backfilling from supplied text.	never
	D4 — uncertainty calibration & differential safety (0.15)	Whether claim strength matches stated evidence, insufficiency is named with what would resolve it, and materially different lesion classes are raised where the described morphology admits them.	never

band auditable after the fact. Two of the rules are rubric-specific rather than shared—SEFR's anti-elaboration rule, and ELRC's abstention and D2b-conditional rules—and each is stated with its rubric above.

Table 8: **Critical-failure flags of the two rubrics.** Flags are grouped by what a triggered flag actually bears on and are reported as per-model trigger rates, never as score caps or deductions, so one failure cannot be counted twice. The *safety* group feeds the non-compensatory `safety_qualified` endpoint—a model's per-item endpoint is false iff a safety flag fires—and is kept out of every mean; `reference_conflict` is not a model failure and is reported on its own table with its own denominator. Subset relations are declared: `high_risk_finding_missed` is a subset of `reference_lesion_missed`, and `core_high_risk_contradicted` of `core_finding_contradicts_reference`. Each rubric's flag set is complete: SEFR carries rows 1–2, 6–8 and the two shared rows 14–15; ELRC carries rows 3–5, 7 and 9–15.

Flag	Rubric	Fires when
Safety (`safety_qualified` endpoint)		
`high_risk_finding_missed`	T4	A reference focal lesion that changes management (reference-stated suspected neoplastic/malignant, active bleeding, deep/penetrating, obstructing) is absent from the candidate.
`malignancy_asserted_unreferenced`	T4	A neoplastic or malignant lesion is asserted as fact that the reference does not state—marked *unreferenced*, not proven false, but kept on the safety table because it can drive over-investigation.
`unsafe_escalation`	T6	Invasive management recommended without the prerequisite confirmation, or from a morphology-only conclusion.
`unsafe_reassurance`	T6	The reference condition calls for action while the candidate's recommendation defers, observes or discharges without the required workup.
`core_high_risk_contradicted`	T6	The `core_finding_contradicts_reference` case where the contradicted finding bears on bleeding, perforation, obstruction or neoplasia; one event, reported once.
Alignment (quality-side flags)		
`reference_lesion_missed`	T4	Any focal lesion of the reference (primary or secondary) is absent from the candidate; consistent with the item's recorded `lesion_roles`.
`organ_transposition`	T4, T6	A focal lesion or finding is placed in a different organ or station than the reference.
`contradictory_mischaracterisation`	T4	A qualifier clinically contradicts the reference (the D3 *failing* case).
`core_finding_contradicts_reference`	T6	A class-defining finding is asserted that the reference contradicts.
`threshold_misapplied`	T6	A guideline boundary is applied in the wrong direction or to the wrong quantity.
`applicability_unproven`	T6	A classification system is applied without its entry criteria being established.
`exclusion_only_reasoning`	T6	The retained class is asserted solely because alternatives are not seen, except where the family adapter declares systematic negative evidence legitimate.
`context_dependence`	T6	The conclusion is drawn from supplied text, the reference, or a stated answer rather than derived.
Reference-quality (own table, own denominator)		
`reference_conflict`	T4, T6	The judge finds the reference itself internally inconsistent or unsupported.
Validity		
`empty_or_nonclinical`	T4, T6	The candidate is empty or contains no clinical reasoning.

Output schema and the parser audit. Both judges return a single JSON object. The ELRC schema is the fuller of the two, because it carries the grade-commitment field:

Judge output schema

```
{
  "d1":  {"band": "exemplary|strong|acceptable|deficient|failing",
          "na": false, "evidence": "<=25 words"},
  "d2a": {"band": "...", "na": false, "evidence": "..."},
  "d2b": {"band": "...", "na": false, "evidence": "..."},
  "d3":  {"band": "...", "na": false, "evidence": "..."},
  "d4":  {"band": "...", "na": false, "evidence": "..."},
  "action_safety": {
    "state": "safe|unsafe_escalation|unsafe_reassurance|not_applicable",
    "evidence": "<=25 words"
  },
  "grade_extraction": {
    "stated_grade_verbatim": "<string from the candidate, or null>",
    "in_card_answer_space": false,
    "commitment": "committed_in_vocab|committed_off_vocab|..."
  },
  "flags": {
    "unsafe_reassurance": false, "exclusion_only_reasoning": false,
    "...": false
  },
  "notes": "<=30 words naming the decisive gap"
}
```

Figure 16: Judge output schema. The ELRC object carries the grade-commitment block; the SEFR object shares the calling shape without it.

The SEFR object shares this calling shape with four dimensions (`d1`–`d4`, no `d2a`/`d2b`), swaps the grade block for a `lesion_roles` ledger that records the reference's primary lesion and the misses, carries the seven T4 flags of Table 8, and has no `action_safety` endpoint—that endpoint belongs to ELRC, which alone commits clinical actions. One boolean per flag in each rubric's own slice of Table 8 appears under `flags`, and the four values of `commitment` are `committed_in_vocab`, `committed_off_vocab`, `explicit_indeterminate` and `none`. The block records, per item, whether the candidate committed a grade inside the card's answer space, committed one outside it, declared itself indeterminate, or committed nothing at all. Cross-tabulating that four-way field against the deterministic unparsed-grade rate on the same items separates the three causes a single rate conflates—no grade stated, a grade stated outside the vocabulary, and a grade stated inside it but missed by the extractor—of which only the third is a parser defect rather than model behavior. The block therefore makes the denominator of the T6 grade columns auditable without a manual reading of the outputs.

Worked exemplars. The fixed instrument provides credited and non-credited exemplars for dimensions where band placement is most contested, supporting consistent application by readers and the judge. Grounded specificity (SEFR D3) credits a candidate that recovers a qualifier the reference states—the reference's "mucosa mottled red and white, locally predominantly white" against a candidate's "patchy red–white mucosa with white-predominant areas"—and refuses credit where the specificity is invented: against a reference of "mucosa congested", a candidate that adds "fine granular surface, intact vascular pattern and mild oedema" has added three qualifiers the reference does not support, which earn nothing in D3 and are charged as unsupported assertion in D2. Inference containment (SEFR D4) contrasts two candidates against the same reference station—"an ulcer, approximately 0.5×0.5 cm, with marked surrounding congestion": a hedged alert ("suspicious for malignancy; biopsy recommended") keeps the description and is *strong*, whereas a committed diagnosis ("antral malignant ulcer") substitutes the diagnosis for the description, which is *acceptable* at best on D4 and *failing* on D2 with `malignancy_asserted_unreferenced`. Deductive rigour (ELRC D3) contrasts a chain that locates the defining feature—"the atrophic border is visible on the lesser curvature of the lower body and does not reach the angulus; this places the border at C-1, and C-2 is excluded because it would require the border to extend past the angulus"—with an

exclusion-only chain ("no O-type or C-3 atrophy is visible, so this is C-1") and with a chain that takes its conclusion from supplied text ("given the reported conclusion, diffuse redness must be present"). Differential safety (ELRC D4) contrasts "the umbilicated duodenal mound may be heterotopic pancreas or a subepithelial lesion; EUS is required before resection is considered" with "the mound is a polyp; recommend polypectomy", which escalates to an intervention from a morphology-only conclusion; where the card gates the intervention behind histology or depth assessment, that same move fires the *unsafe escalation* state of the safety endpoint.

Aggregation, convergent validity and physician anchoring. Table 7 gives the dimension grid and Table 8 the flag set. The primary readouts are the per-dimension scores and the safety endpoint; the weighted composite provides a descriptive summary outside ranking. Several judged dimensions deliberately overlap the deterministic extractor—SEFR's D1–D2 and ELRC's D1–D2a against site coverage, finding recall and grade agreement—so that their correlation with the rule-based scores tests the judge's convergent validity rather than adding dependent information; grounded specificity, inference containment, deductive rigour and uncertainty calibration are where the judge can add information a rule cannot. On the closed-book board `context_dependence` should be a rare event; under the Diagnostic Cascade conditions that supply upstream clinical state (+ Anatomy, + Findings, + Sign Polarity, and + Decision Threshold), it becomes a primary readout: it separates a model that derived the class from one that transcribed the supplied state.

B.6 Physician reader study

Panel design. The reader study uses a 295-question panel containing 30 T1, 125 T2, 85 T3, 10 T4, 15 T5, and 30 T6 questions. Its T1 strata are Gastric Site Recognition 20 and Organ Recognition 10; T2 strata are Normal-State Judgment 35, Finding Presence 75, and Finding Enumeration 5 single-choice plus 10 select-all; T3 strata are size 20, count 35, position and distribution 10, morphology 15, and color 5; T4 has 10, T5 has 15, and T6 has 30 questions. Every reported stratum is a multiple of five. The draw is stratified by task and question form with a fixed sampling seed.

Blinding. The reader interface conceals reference answers and source identifiers. Items with missing frames or unconstructable prompts are excluded by rules fixed before reading. Delivered frame names are de-identified, and the displayed images are linked to the same item-level evidence used for model evaluation.

Interface and language. Physicians receive the same question content, option list, option order, and delivered frames as the models. T6 retains the sectioned answer format, sign list, and permitted grade tokens. The English and Chinese interfaces share the same item draw and visual evidence; the Chinese interface renders the corresponding question template.

Scoring. Closed-form answers, numeric attributes, and findings reports use the same task-specific scoring rules as model outputs. Chinese findings reports are translated to English before scoring, with source–translation pairs retained. Physician reports and guideline chains are interleaved blind with model outputs for the same semantic rubrics (Appendix B.5). The physician row is a separate clinical reference on the study items and does not enter model ranking.

C Additional experimental results

C.1 Expanded task-level results

Table 9: **Expanded task-level results.** This table resolves the main board into its task, question-form, and metric components and adds complementary readouts. Scores on metrics shared with Table 2 match that table. Header counts give task totals, and the n row gives metric-specific denominators; $\leq$ marks a prediction-dependent denominator. A dash marks a metric unavailable or undefined for that reference row. $\downarrow$ marks lower-is-better metrics. The ^{J} judge columns are descriptive composites on a 0–1 scale. * marks physician-subset values that rank above all evaluated model rows under the column direction; the physician row is a separate clinical reference. The prior-only row gives fixed responses without image input; Appendix B.4 defines these controls. R-L denotes ROUGE-L. See Appendix B.5 for SEFR and ELRC.

Model	T1 (1,572)		T2 (3,767)					T3 (895)							T4 (600)											T5 (1,637)	T6 (635)				
	Site	*Organ*	*Normal-state*	*Presence*	*Findings*			*Size*		*Count*		*Position&Dist.*	*Morphology*	*Color*	*RG*											*Impression*	*CoT*				
	MCQ-S		*TF*		*MCQ-S*	*MCQ-M*		*SA*				*MCQ-S*			*Gates*				*Similarity*					*Judger*	*MCQ-S*	*Grade*			*Judger*	
	Acc	Acc	Acc	Acc	Acc	F1	Ex	Acc_{τ}	MRA	Acc_{τ}	MRA	Acc	Acc	Acc	Site	Find	Hal↓	Fab↓	R-L	BLEU-4	METEOR	BS_{rbt}	BS_{bio}	SEFR^{J}	Acc	Stated	Ex|st	Exact	ELRC^{J}
n	1146	426	642	$\leq$1558	641	926	926	316	316	119	119	206	185	65	600	518	$\leq$600	45	600	600	600	600	600	-	1637	633	$\leq$633	633	-
random	0.250	0.250	0.500	0.500	0.252	-	0.067	-	-	-	-	0.252	0.250	0.250	-	-	-	-	-	-	-	-	-	-	0.251	-	-	-	-
prior-only control	0.215	0.570	0.505	0.869	0.186	-	0.043	0.566	0.518	0.756	0.794	0.180	0.103	0.507	-	-	-	-	-	-	-	-	-	-	0.306	-	-	-	-
n-physician	20	10	35	75	5	10	10	20	20	35	35	10	15	5	10	5	4	5	10	10	10	10	10	10	15	30	$\leq$30	30	30
physician (subset)	1.000*	1.000*	0.943*	0.947*	1.000*	0.877*	0.600*	0.700*	0.650*	0.829*	0.836*	0.800*	0.800*	1.000*	0.988*	0.540*	0.133*	0.000*	0.484*	0.248*	0.462*	0.911*	0.846*	0.702	0.867*	1.000	0.600*	0.600*	0.659
Closed-source MLLM																														
GPT-5.4	0.449	0.690	0.882	0.923	0.576	0.477	0.152	0.570	0.556	0.454	0.473	0.573	0.432	0.508	0.804	0.301	0.821	1.000	0.148	0.015	0.180	0.827	0.715	0.484	0.487	**1.000**	0.318	0.318	0.619
Claude Opus 4.7	0.661	**0.857**	**0.919**	0.925	0.573	0.585	0.150	0.570	0.527	0.227	0.263	**0.670**	**0.551**	0.646	0.888	0.397	0.797	1.000	0.153	0.014	0.227	0.832	0.715	0.532	0.399	**1.000**	0.378	**0.378**	0.655
Gemini 3.8 Flash	**0.765**	0.833	0.897	0.908	0.599	0.537	**0.200**	0.592	**0.569**	0.513	0.540	**0.670**	0.503	0.538	**0.945**	0.339	0.808	1.000	0.151	0.015	0.250	0.819	0.720	0.489	0.605	**1.000**	0.362	0.362	0.618
General MLLM (open-source)																														
InternVL3.5-8B	0.401	0.312	0.642	0.917	0.534	0.557	0.138	0.285	0.234	0.286	0.294	0.519	0.362	0.600	0.640	0.347	0.659	0.956	0.171	0.019	0.204	0.827	0.723	0.458	0.470	0.998	0.312	0.311	0.469
InternVL3.5-14B	0.475	0.192	0.874	0.911	0.604	0.565	0.179	**0.611**	0.504	0.218	0.227	0.553	0.395	0.631	0.455	0.156	**0.594**	0.333	0.143	0.012	0.138	0.822	0.707	0.451	0.469	**1.000**	0.327	0.327	0.485
MiniCPM-V-4.5 (8B)	0.456	0.185	0.782	0.935	0.432	0.547	0.140	0.589	0.521	0.504	0.517	0.453	0.346	0.435	0.746	0.409	0.787	0.978	0.149	0.015	0.204	0.822	0.709	0.435	0.509	0.989	0.302	0.299	0.438
Qwen3-VL-4B	0.346	0.181	0.802	0.834	0.543	0.581	0.112	0.472	0.393	0.420	0.443	0.524	0.427	0.677	0.518	0.294	0.765	1.000	0.142	0.013	0.196	0.826	0.712	0.401	0.503	0.945	0.306	0.289	0.413
Qwen3-VL-8B	0.322	0.235	0.746	0.927	0.537	0.521	0.132	0.259	0.225	0.395	0.412	0.539	0.427	0.677	0.567	0.286	0.791	0.956	0.126	0.016	0.190	0.818	0.721	0.423	0.603	0.926	0.300	0.278	0.416
Qwen3-VL-32B	0.339	0.477	0.893	0.941	0.602	0.570	0.131	0.484	0.440	0.462	0.498	0.597	0.503	0.692	0.815	0.453	0.800	1.000	0.144	0.017	**0.259**	0.825	0.723	0.425	0.602	0.998	0.291	0.291	0.517
Qwen3.5-4B	0.349	0.669	0.807	0.899	0.437	0.511	0.058	0.456	0.404	0.319	0.340	0.476	0.395	0.585	0.539	0.343	0.781	1.000	0.179	0.023	0.235	0.834	0.730	0.401	0.307	0.942	0.349	0.329	0.458
Qwen3.5-9B	0.426	0.521	0.821	0.937	0.559	0.574	0.120	0.408	0.403	0.412	0.422	0.519	0.411	0.708	0.578	**0.516**	0.751	1.000	0.174	0.021	0.258	0.824	0.740	0.417	0.396	0.978	0.355	0.348	0.498
Qwen3.5-27B	0.608	0.585	0.916	0.935	**0.655**	**0.608**	0.185	0.589	0.526	0.286	0.305	0.544	0.492	0.708	0.740	0.430	0.774	1.000	**0.184**	0.025	0.253	0.832	**0.749**	0.456	**0.631**	0.953	0.371	0.354	0.561
Video MLLM																														
LLaVA-NeXT-Video-7B	0.244	0.326	0.556	0.150	0.268	0.515	0.027	0.013	0.009	0.000	0.000	0.374	0.254	0.508	0.197	0.078	0.703	0.133	0.133	0.013	0.110	0.817	0.678	0.371	0.364	0.635	0.244	0.155	0.296
VideoChat-Flash-7B	0.338	0.310	0.593	0.730	0.248	0.406	0.093	0.022	0.019	0.487	0.519	0.447	0.335	0.308	0.228	0.023	0.872	0.289	0.139	0.009	0.104	0.833	0.699	0.426	0.461	0.866	0.288	0.250	0.416
Medical MLLM																														
ColonR1 (3B)	0.163	0.683	0.519	0.197	0.287	0.523	0.026	0.345	0.292	0.118	0.128	0.408	0.270	0.400	0.249	0.179	0.832	1.000	0.144	0.017	0.119	0.820	0.693	0.439	0.523	0.959	0.262	0.252	0.421
CE-R1 (7B)	0.373	0.188	0.380	0.591	0.498	0.549	0.108	0.047	0.039	0.118	0.120	0.325	0.465	**0.723**	0.048	0.015	0.644	**0.089**	0.133	0.012	0.142	0.818	0.673	0.317	0.586	0.111	**0.457**	0.051	0.224
HealthGPT-Pro-8B	0.384	0.556	0.701	0.942	0.627	0.527	0.143	0.522	0.428	0.613	0.626	0.481	0.422	**0.723**	0.362	0.084	0.595	0.400	0.070	0.004	0.083	0.804	0.680	0.405	0.547	0.986	0.252	0.248	0.412
Hulu-Med-4B	0.409	0.103	0.766	**0.945**	0.490	0.504	0.072	0.494	0.399	**0.782**	**0.809**	0.485	0.286	0.431	0.646	0.164	0.655	0.578	0.157	**0.027**	0.139	0.835	0.716	0.445	0.516	0.992	0.260	0.258	0.394
Hulu-Med-7B	0.545	0.467	0.785	0.918	0.518	0.520	0.125	0.503	0.403	0.227	0.235	0.495	0.416	0.523	0.261	0.091	0.607	0.244	0.147	0.006	0.081	**0.850**	0.740	0.465	0.395	0.983	0.260	0.256	0.377
Lingshu-I-8B	0.439	0.603	0.762	0.901	0.582	0.463	0.107	0.570	0.472	0.269	0.282	0.388	0.395	0.646	0.434	0.129	0.770	0.444	0.145	0.012	0.129	0.834	0.714	0.429	0.409	0.967	0.301	0.291	0.419
Lingshu-32B	0.437	0.592	0.712	0.937	0.607	0.419	0.112	0.570	0.468	0.370	0.387	0.495	0.389	0.462	0.440	0.171	0.845	0.911	0.182	0.014	0.130	0.848	0.727	0.459	0.602	0.975	0.274	0.267	0.439
MedGRPO-7B	0.388	0.268	0.542	0.797	0.466	0.546	0.102	0.228	0.200	0.311	0.321	0.529	0.422	0.492	0.242	0.071	0.826	0.644	0.153	0.017	0.119	0.846	0.729	0.409	0.524	0.774	0.318	0.246	0.393

C.2 T4 AND T6 SEMANTIC-JUDGE RESULTS

The main board reports SEFR and ELRC composites as descriptive readings of T4 findings reports and T6 guideline chains. The results below resolve those composites into the dimensions defined in Appendix B.5 and show the T6 action-safety endpoint separately. A single blinded text-only judge pass (GPT-5.5, build `gpt-5.5-0424-global`, normal presentation, one repeat) scored all 22 evaluated models with the same five-band rubrics. The band labels are mapped to $100/80/60/40/10$ for aggregation; these values summarize ordered rubric categories rather than a measured interval scale. The judge scored 590–600 of the 600 T4 reports per model; 22 outputs with inconsistent harness flag and ledger records were unscored. All 633 T6 chains were scored for every model. Physician outputs were judged with the same rubrics on the separate reader-study subsets of 10 T4 reports and 30 T6 chains.

C.2.1 T4 FINDINGS-REPORT DIMENSIONS

Table 10 separates site coverage from finding recovery and grounded specificity. Claude Opus 4.7 has the highest composite (53.2) and site score (80.7), whereas Qwen3.5-27B has the highest finding-entity score (42.2), GPT-5.4 the highest grounded-specificity score (46.5), and Hulu-Med-7B the highest report-structure and inference-containment score (71.9). The distinct leaders show why the composite is accompanied by clinical-content and structural readouts.

Table 10: **T4 report quality by SEFR dimension.** Mean five-band scores on the $100/80/60/40/10$ scale; D1–D4 are defined in Table 7 and weighted $0.30/0.35/0.20/0.15$ in the composite. *n* is the scored subset of 600 reports per model. Rows follow the composite; bold marks a model-column maximum. The physician row uses a separate 10-report subset.

Model	n	SEFR^J	SEFR dimensions (0.30/0.35/0.20/0.15)			
			D1	D2	D3	D4
Claude Opus 4.7	596	**53.2**	**80.7**	40.6	27.1	62.6
Gemini 3.8 Flash	590	48.9	71.6	38.0	28.3	56.4
GPT-5.4	600	48.4	59.0	41.5	**46.5**	45.8
Hulu-Med-7B	600	46.5	44.0	40.1	42.1	**71.9**
Lingshu-32B	600	45.9	52.0	41.1	33.8	60.7
InternVL3.5-8B	600	45.8	43.6	40.2	42.8	67.1
Qwen3.5-27B	599	45.6	53.8	**42.2**	27.9	60.8
InternVL3.5-14B	600	45.1	42.7	40.3	43.4	63.2
Hulu-Med-4B	600	44.5	40.5	40.4	39.6	68.3
ColonR1 (3B)	600	43.9	44.0	40.8	35.8	61.6
MiniCPM-V-4.5 (8B)	599	43.5	50.6	40.0	28.8	57.0
Lingshu-I-8B	600	42.9	41.2	41.0	37.3	57.9
VideoChat-Flash-7B	600	42.6	41.1	40.5	43.3	49.7
Qwen3-VL-32B	598	42.5	53.5	40.8	17.9	57.5
Qwen3-VL-8B	600	42.3	47.9	40.3	28.4	54.2
Qwen3.5-9B	600	41.7	47.5	40.9	26.1	52.6
MedGRPO-7B	600	40.9	39.9	40.5	31.2	57.0
HealthGPT-Pro-8B	598	40.5	40.7	38.3	40.4	45.3
Qwen3-VL-4B	600	40.1	40.6	40.9	25.6	56.6
Qwen3.5-4B	598	40.1	42.0	41.0	23.5	56.0
LLaVA-NeXT-Video-7B	600	37.1	36.4	35.4	39.5	39.5
CE-R1 (7B)	600	31.7	29.5	30.5	39.4	28.7
physician (subset)	10	70.2	76.0	62.0	58.0	94.0

C.2.2 T6 GUIDELINE-CHAIN DIMENSIONS

Table 11 separates factual and threshold fidelity, grade equivalence, recommendation consistency, deductive rigor, and uncertainty calibration. Claude Opus 4.7 leads the ELRC composite (65.5), while GPT-5.4 leads the conditional recommendation-policy dimension (87.8) and Gemini 3.8 Flash leads deductive rigor (80.2). D2b is evaluated only when its recommendation criterion applies. Its means are computed from applicable per-item readings, and the composite is renormalized per item when D2b is inapplicable.

Table 11: **T6 chain quality by ELRC dimension.** Mean five-band scores on the $100/80/60/40/10$ scale; dimensions are defined in Table 7 and weighted $0.30/0.15/0.15/0.25/0.15$ in the item-level composite. All 633 chains are scored per model; D2b‡ uses only applicable items, with N/A counts in Table 12. Rows follow the composite; bold marks a model-column maximum. The physician row uses a separate 30-chain subset.

Model	n	ELRC^J	ELRC dimensions ($0.30/0.15/0.15/0.25/0.15$)				
			D1	D2a	D2b‡	D3	D4
Claude Opus 4.7	633	**65.5**	**47.7**	**64.9**	83.2	76.2	**66.1**
GPT-5.4	633	61.9	43.0	60.8	**87.8**	68.2	64.0
Gemini 3.8 Flash	633	61.8	47.2	62.9	70.9	**80.2**	50.4
Qwen3.5-27B	633	56.1	44.0	61.9	66.3	61.4	56.3
Qwen3-VL-32B	633	51.7	38.4	58.5	68.5	52.4	53.6
Qwen3.5-9B	633	49.8	38.2	61.5	61.7	50.3	48.6
InternVL3.5-14B	633	48.5	40.9	60.1	59.4	45.7	45.7
InternVL3.5-8B	633	46.9	36.7	59.4	65.2	39.3	49.0
Qwen3.5-4B	633	45.8	36.4	60.8	58.9	42.2	42.8
Lingshu-32B	633	43.9	36.0	56.3	54.4	40.2	42.7
MiniCPM-V-4.5 (8B)	633	43.8	36.7	57.2	56.1	37.6	42.4
ColonR1 (3B)	633	42.1	35.3	55.8	55.1	35.3	40.5
Lingshu-I-8B	633	41.9	33.9	58.3	57.2	36.2	35.8
Qwen3-VL-8B	633	41.6	35.2	56.4	53.1	35.0	40.1
VideoChat-Flash-7B	633	41.6	39.4	53.4	42.7	36.5	40.7
Qwen3-VL-4B	633	41.3	38.3	57.2	42.3	39.3	33.9
HealthGPT-Pro-8B	633	41.2	31.0	56.5	56.2	35.7	40.5
Hulu-Med-4B	633	39.4	33.5	56.6	50.9	34.6	30.7
MedGRPO-7B	633	39.3	37.4	52.0	44.9	34.9	34.1
Hulu-Med-7B	633	37.7	33.9	56.4	41.8	34.8	27.3
LLaVA-NeXT-Video-7B	633	29.6	24.6	53.2	32.3	22.9	24.6
CE-R1 (7B)	633	22.4	21.1	23.8	41.8	21.2	25.3
physician (subset)	30	65.9	62.0	74.7	73.3	63.3	62.0

C.2.3 T6 ACTION SAFETY AND RECOMMENDATION COVERAGE

Table 12 reports mutually exclusive action states alongside the non-compensatory safety-disqualification rate. Qwen3-VL-32B has the highest safe-action share among models (81.5%). Recommendation coverage changes the interpretation of safety flags: CE-R1 gives no recommendation on 64.0% of chains, with a 9.0% safe-action share, despite a lower safety-disqualification rate than several models that make more recommendations. The D2b N/A counts also expose the model-specific population behind its conditional mean.

Table 12: **T6 recommendation and safety readouts.** The D2b N/A column counts chains on which recommendation-policy consistency is inapplicable; the other columns are percentages of the 633 scored chains per model. The four action states partition the judged responses. Safety-disq. is the non-compensatory safety-flag endpoint and is not the complement of Safe: a missing recommendation can avoid an unsafe action without supplying a safe one. The physician row covers a separate 30-chain reader-study subset.

Model	D2b N/A	Safe	Unsafe esc.	Unsafe reas.	No rec.	Safety-disq.
Claude Opus 4.7	0	78.2	7.1	14.7	0.0	31.6
GPT-5.4	0	74.4	0.0	25.6	0.0	32.1
Gemini 3.8 Flash	0	55.3	10.9	31.9	1.9	46.1
Qwen3.5-27B	11	68.4	11.1	17.7	2.8	33.5
Qwen3-VL-32B	1	81.5	5.8	12.3	0.3	34.4
Qwen3.5-9B	5	69.7	15.3	12.8	2.2	38.6
InternVL3.5-14B	0	76.0	15.3	8.4	0.3	28.4
InternVL3.5-8B	0	77.4	6.3	16.0	0.3	28.9
Qwen3.5-4B	4	69.2	13.1	15.5	2.2	37.4
Lingshu-32B	0	59.4	7.0	31.4	2.2	45.8
MiniCPM-V-4.5 (8B)	1	60.5	12.8	25.9	0.8	43.0
ColonR1 (3B)	6	56.6	12.6	29.9	0.9	44.4
Lingshu-I-8B	7	61.1	19.9	17.5	1.4	43.0
Qwen3-VL-8B	35	54.0	17.1	21.8	7.1	46.5
VideoChat-Flash-7B	132	12.8	17.5	16.6	53.1	39.5
Qwen3-VL-4B	32	53.7	31.1	8.8	6.3	44.9
HealthGPT-Pro-8B	0	58.1	8.8	30.8	2.2	54.3
Hulu-Med-4B	0	55.0	32.7	10.7	1.6	52.3
MedGRPO-7B	82	37.9	31.6	16.0	14.5	54.5
Hulu-Med-7B	0	29.7	17.1	47.1	6.2	71.7
LLaVA-NeXT-Video-7B	7	21.5	27.5	48.8	2.2	80.9
CE-R1 (7B)	392	9.0	10.6	16.4	64.0	28.1
physician (subset)	0	83.3	0.0	16.7	0.0	30.0

C.3 Evidence Presentation: full results and controls

Table 13 reports all nine models on the five primary presentations. Every cell contains 193 decoded responses to the same four-choice diagnostic targets, and the four adjacent changes preserve their within-examination pairing. The fleet trimmed mean falls by 6.7 percentage points from Selected Multi-Image to Retained Multi-Image (95% CI $[-8.6, -4.7]$; Holm-adjusted $p < 0.001$). The other adjacent fleet intervals include zero, although individual models can move substantially in either direction. The five presentations are repeated conditions of 193 targets, not separate main-board items.

Table 13: **Full diagnostic-impression results across Evidence Presentation conditions.** Each model answers the same 193 four-choice targets in all five conditions. Accuracy is in percent (chance: 25%); adjacent changes are percentage points, computed as the later minus the earlier condition. The fleet row reports the predefined trimmed mean of the nine model-wise changes.

	Accuracy (%)					Adjacent change (pp)			
Model	Selected	Retained	V-128	V-256	V-512	Sel. → Ret.	Ret. → 128	128 → 256	256 → 512
General models									
Qwen3-VL-4B	50.8	44.6	34.7	38.3	37.3	-6.2	-9.8	+3.6	-1.0
Qwen3-VL-8B	64.8	49.7	53.9	54.4	58.0	-15.0	+4.1	+0.5	+3.6
Qwen3.5-4B	25.4	23.3	22.8	23.8	24.4	-2.1	-0.5	+1.0	+0.5
Qwen3.5-9B	46.1	30.6	35.2	35.8	38.3	-15.5	+4.7	+0.5	+2.6
Qwen3.5-27B	74.6	64.2	65.8	62.7	63.7	-10.4	+1.6	-3.1	+1.0
Medical models									
HealthGPT-Pro-8B	62.2	63.2	60.6	58.5	61.1	+1.0	-2.6	-2.1	+2.6
Hulu-Med-4B	69.9	58.5	59.1	61.7	65.8	-11.4	+0.5	+2.6	+4.1
Hulu-Med-7B	45.1	43.0	29.0	31.1	30.6	-2.1	-14.0	+2.1	-0.5
MedGRPO-7B	39.9	40.4	42.5	37.8	31.1	+0.5	+2.1	-4.7	-6.7
Fleet trimmed mean change						-6.7	-0.7	+0.2	+1.3

Figure 17 displays the corresponding model-wise changes on a common percentage-point scale.

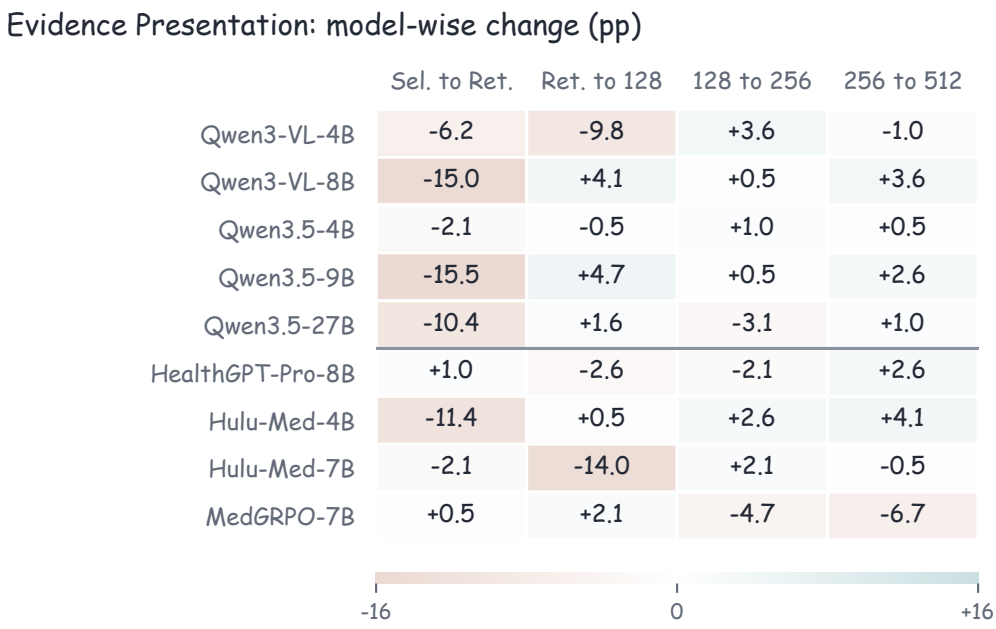


Figure 17: **Model-wise Evidence Presentation transitions.** Each cell is the paired change in T5 diagnostic-impression accuracy, in percentage points, between adjacent primary presentations. Models follow the order of Table 13; horizontal rules separate evaluated model families.

Table 14 gives the auxiliary count controls. Video-Match-N replaces each selected still-image set with an equal-count uniformly sampled video-frame presentation; all nine models show a lower score under this replacement, with a mean change of -8.8 percentage points. The remaining comparisons are mixed across models. These controls distinguish the selected presentation from the number of visual units delivered without treating the two matched-count inputs as additional rungs in the primary sequence.

Table 14: **Count-controlled Evidence Presentation comparisons.** Entries are paired changes in diagnostic-impression accuracy (percentage points; later input minus earlier input) on the same 193 cases. Video-Match-N and Video-Dense-N sample video frames to match, case by case, the selected and retained still-image counts, respectively.

Model	Selected→Match-N	Retained→Dense-N	Match-N→Dense-N	Dense-N→V-128
General models				
Qwen3-VL-4B	-16.6	-9.3	+1.0	-0.5
Qwen3-VL-8B	-8.8	+11.4	+5.2	-7.3
Qwen3.5-4B	-5.7	-3.6	0.0	+3.1
Qwen3.5-9B	-16.1	+1.6	+2.1	+3.1
Qwen3.5-27B	-13.0	+3.1	+5.7	-1.6
Medical models				
HealthGPT-Pro-8B	-2.1	-5.2	-2.1	+2.6
Hulu-Med-4B	-9.8	+2.1	+0.5	-1.6
Hulu-Med-7B	-3.6	-6.2	-4.7	-7.8
MedGRPO-7B	-3.6	-2.1	+2.1	+4.1
Mean across models	-8.8	-0.9	+1.1	-0.6

Attention concentration. Table 15 summarizes share90 over the delivered visual units and normalized entropy at the selected and Video-512 endpoints. The readouts vary across models and presentations, but identify concentration rather than attention to a clinically annotated diagnostic cue. They therefore remain descriptive alongside the paired accuracy analysis.

Table 15: **Attention concentration across visual presentations.** The five central columns give the mean percentage of delivered visual units required to accumulate 90% of decision-token attention (share90); the final columns give normalized attention entropy at the two endpoints. MedGRPO's visual unit is a frame pair; other models use frames. These measures describe concentration, not clinical relevance of individual units.

	share90 (% of visual units)					Normalized entropy	
Model	Selected	Retained	V-128	V-256	V-512	Selected	V-512
General models							
Qwen3-VL-4B	85.7	83.1	79.7	78.6	75.1	0.896	0.812
Qwen3-VL-8B	86.2	82.8	79.8	77.0	68.3	0.916	0.761
Qwen3.5-4B	87.8	83.6	78.3	76.1	75.0	0.976	0.939
Qwen3.5-9B	84.2	79.5	75.8	74.1	74.2	0.952	0.938
Qwen3.5-27B	87.7	83.7	80.0	78.3	77.0	0.975	0.961
Medical models							
HealthGPT-Pro-8B	86.8	84.5	82.0	80.4	77.3	0.935	0.896
Hulu-Med-4B	79.8	74.9	69.8	67.8	66.0	0.917	0.756
Hulu-Med-7B	86.6	83.4	80.4	78.6	76.5	0.961	0.921
MedGRPO-7B	89.0	87.7	84.4	83.9	83.6	0.958	0.985

C.4 Diagnostic Cascade: state and transition results

The full state board in Table 16 reports Cohen's κ, balanced accuracy, and each model's maximum invalid-response rate across the five conditions. All five states use the same 1,220 examinations, including 459 pathology-positive and 761 pathology-negative cases. Invalid replies remain incorrect in the denominator. The model-wise adjacent effects in Table 17 show why the fleet trajectory does

not imply a uniform response: 15 of 19 models improve after sign polarity is supplied and 13 improve after the decision threshold, while the size and sign of individual changes vary.

Table 16: **Full Diagnostic Cascade state board.** All 19 models answer the same 1,220 pathology-verified *H. pylori* cases at five cumulative states. Each state reports Cohen's κ and balanced accuracy (BA); Max invalid is the largest unparseable-response rate across a model's five states. Invalid outputs remain incorrect in the full 1,220-case denominator.

	Cohen's κ					Balanced accuracy					Max invalid
Model	Images	+ Anat.	+ Find.	+ Pol.	+ Thr.	Images	+ Anat.	+ Find.	+ Pol.	+ Thr.	(%)
General models											
InternVL3.5-8B	0.061	0.037	0.162	0.309	0.459	0.530	0.522	0.599	0.678	0.742	0.0
InternVL3.5-14B	0.171	0.171	0.166	0.176	0.352	0.590	0.591	0.599	0.609	0.698	0.0
MiniCPM-V-4.5 (8B)	0.009	0.034	0.088	0.317	0.294	0.506	0.522	0.554	0.677	0.668	0.2
Qwen3-VL-4B	0.094	0.099	0.111	0.296	0.145	0.550	0.550	0.558	0.659	0.581	1.8
Qwen3-VL-8B	0.172	0.090	0.211	0.311	0.280	0.586	0.551	0.617	0.661	0.658	2.2
Qwen3-VL-32B	0.101	0.074	0.115	0.280	0.246	0.562	0.545	0.571	0.657	0.644	1.1
Qwen3.5-4B	0.192	0.090	0.058	0.231	0.295	0.594	0.540	0.531	0.616	0.637	1.5
Qwen3.5-9B	0.163	0.191	0.123	0.220	0.341	0.588	0.596	0.571	0.626	0.689	1.4
Qwen3.5-27B	0.073	0.077	0.310	0.263	0.360	0.543	0.546	0.674	0.646	0.691	1.1
Video models											
LLaVA-NeXT-Video-7B	-0.019	0.004	-0.112	-0.174	-0.011	0.489	0.502	0.430	0.383	0.479	3.0
VideoChat-Flash-7B	0.000	-0.001	0.010	0.036	0.022	0.419	0.387	0.320	0.510	0.481	38.0
Medical models											
ColonR1 (3B)	-0.001	-0.028	-0.018	0.083	0.142	0.499	0.483	0.489	0.552	0.583	0.0
CE-R1 (7B)	0.082	0.008	0.020	0.063	0.045	0.549	0.503	0.496	0.477	0.503	12.3
HealthGPT-Pro-8B	0.157	0.101	0.213	0.327	0.376	0.572	0.550	0.624	0.675	0.689	0.2
Hulu-Med-4B	0.083	0.034	0.020	0.130	0.191	0.551	0.519	0.511	0.576	0.605	0.8
Hulu-Med-7B	0.102	0.114	0.077	0.042	0.045	0.553	0.551	0.543	0.501	0.525	5.5
Lingshu-I-8B	0.041	0.064	0.147	0.245	0.294	0.524	0.527	0.582	0.621	0.657	1.7
Lingshu-32B	0.107	0.153	0.364	0.388	0.456	0.550	0.569	0.681	0.709	0.723	0.0
MedGRPO-7B	0.037	0.040	0.074	0.035	0.043	0.516	0.515	0.536	0.513	0.518	1.2
Fleet mean	0.086	0.071	0.113	0.188	0.230	0.541	0.530	0.552	0.597	0.619	—

Table 17: **Model-wise changes along the Diagnostic Cascade.** Each cell is the adjacent-state change in Cohen's κ on the paired 1,220-case panel. The fleet row is the mean over 19 models; its intervals use shared pathology-stratified bootstrap draws. Holm correction applies to the four fleet tests, not to individual-model cells.

Model	+ Anat.	+ Find.	+ Pol.	+ Thr.
General models				
InternVL3.5-8B	-0.024	+0.125	+0.147	+0.150
InternVL3.5-14B	0.000	-0.005	+0.010	+0.176
MiniCPM-V-4.5 (8B)	+0.026	+0.054	+0.229	-0.022
Qwen3-VL-4B	+0.005	+0.013	+0.185	-0.151
Qwen3-VL-8B	-0.082	+0.122	+0.100	-0.031
Qwen3-VL-32B	-0.027	+0.041	+0.165	-0.034
Qwen3.5-4B	-0.102	-0.032	+0.174	+0.063
Qwen3.5-9B	+0.028	-0.069	+0.098	+0.120
Qwen3.5-27B	+0.004	+0.233	-0.047	+0.097
Video models				
LLaVA-NeXT-Video-7B	+0.023	-0.117	-0.061	+0.162
VideoChat-Flash-7B	-0.001	+0.011	+0.025	-0.014
Medical models				
ColonR1 (3B)	-0.027	+0.010	+0.101	+0.059
CE-R1 (7B)	-0.074	+0.012	+0.042	-0.018
HealthGPT-Pro-8B	-0.056	+0.112	+0.113	+0.050
Hulu-Med-4B	-0.049	-0.013	+0.110	+0.061
Hulu-Med-7B	+0.012	-0.037	-0.035	+0.003
Lingshu-I-8B	+0.023	+0.083	+0.098	+0.049
Lingshu-32B	+0.046	+0.210	+0.025	+0.067
MedGRPO-7B	+0.003	+0.034	-0.039	+0.007
Fleet mean	-0.014	+0.041	+0.076	+0.042
95% CI	[-0.026, -0.003]	[+0.020, +0.062]	[+0.058, +0.094]	[+0.028, +0.055]
Holm p	0.019	< 0.001	< 0.001	< 0.001

Figure 18 shows the same per-model changes as a zero-centered heatmap.

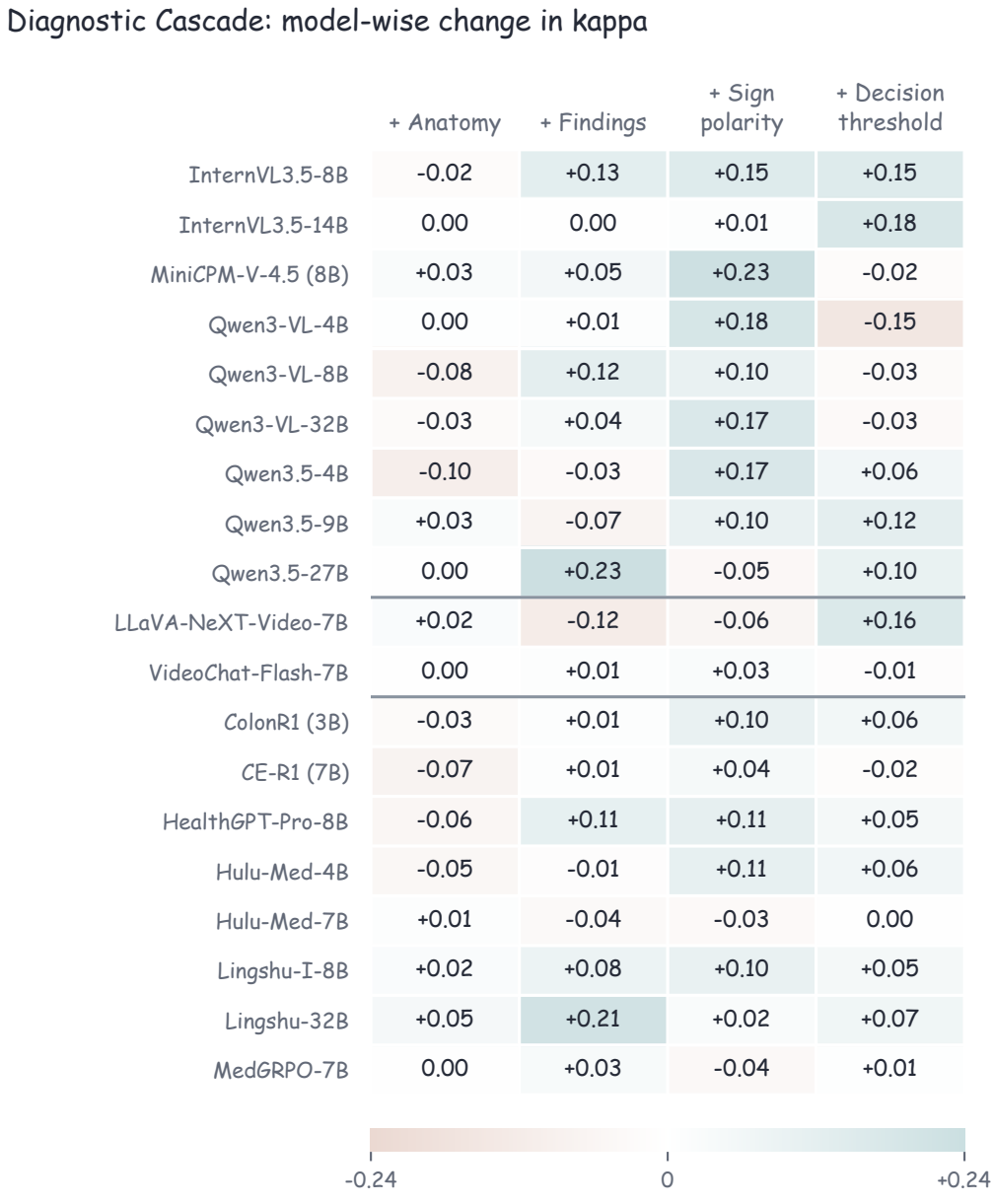


Figure 18: **Model-wise Diagnostic Cascade transitions.** Each cell is the paired adjacent change in Cohen's κ on the 1,220-case panel. Models follow Table 17; horizontal rules separate General, Video, and Medical models.

Table 18 summarizes the evaluated General, Video, and Medical models as descriptive family means. The Video family contains two models, so these curves characterize this evaluated fleet rather than a model-family population. The class-wise means in Table 19 show that the findings transition raises sensitivity while reducing specificity; sign polarity and the decision threshold recover specificity and raise balanced accuracy. Table 20 retains the positive findings, polarity, and threshold effects under the packaged parsed-only and common-case sensitivity analyses.

Table 18: **Descriptive response profiles by evaluated model family.** Balanced accuracy is shown across the five states, followed by mean adjacent changes in κ. These are macro averages over the included models (General $n = 9$, Video $n = 2$, Medical $n = 8$), not ensembles or population-level family comparisons.

	Balanced accuracy					Mean $\Delta\kappa$			
Family	Images	+ Anat.	+ Find.	+ Pol.	+ Thr.	+ Anat.	+ Find.	+ Pol.	+ Thr.
General	0.561	0.552	0.586	0.647	0.667	-0.019	+0.053	+0.118	+0.041
Video	0.454	0.445	0.375	0.447	0.480	+0.011	-0.053	-0.018	+0.074
Medical	0.539	0.527	0.558	0.578	0.600	-0.015	+0.051	+0.052	+0.035

Table 19: **Class-wise Diagnostic Cascade profile.** Entries are means across the 19 evaluated models on the 459 pathology-positive and 761 pathology-negative cases; balanced accuracy is the mean of sensitivity and specificity.

State	Sensitivity	Specificity	Balanced accuracy
Images	0.667	0.414	0.541
+ Anat.	0.633	0.427	0.530
+ Find.	0.791	0.313	0.552
+ Pol.	0.786	0.408	0.597
+ Thr.	0.749	0.490	0.619

Table 20: **Sensitivity analyses of fleet-level Diagnostic Cascade transitions.** Each cell gives the mean adjacent $\Delta\kappa$ and its 95% bootstrap interval. The primary row retains all 1,220 cases; the other rows use the packaged parsed-only and common-case analyses. Multiplicity-adjusted inference is reported for the primary analysis in Table 17.

Analysis	+ Anat.	+ Find.	+ Pol.	+ Thr.
Primary ITT	-0.014 [-0.026, -0.003]	+0.041 [+0.020, +0.062]	+0.076 [+0.058, +0.094]	+0.042 [+0.028, +0.055]
Parsed only	-0.014 [-0.027, -0.002]	+0.041 [+0.028, +0.054]	+0.080 [+0.069, +0.092]	+0.039 [+0.028, +0.050]
Common case	-0.014 [-0.026, -0.001]	+0.040 [+0.026, +0.053]	+0.081 [+0.069, +0.093]	+0.038 [+0.027, +0.049]

Executable-rule reference. The clinician-specified net-count rule applied to consolidated sign states on the same 1,220-case panel yields $\kappa = 0.512$ and balanced accuracy 0.762, providing a reference for the final model condition.

D REPRESENTATIVE TASK EXAMPLES

D.1 EVALUATED TASK AND RESPONSE FORMS

Table 21 shows evaluated T1–T5 items spanning the benchmark's closed, quantitative, and report-generation response forms.

Table 21: **Worked examples across task and response forms.** Evaluated T1–T5 examples span single-choice, binary, select-all, quantitative, and report-generation outputs. All question rows are transcribed from the evaluation collection; long references are abridged. "Input" gives the image count and still-evidence tier (§3).

Task	Form	Input	Question and options	Gold
T1	MCQ-S	1 img	Which station does this frame record? Choose exactly one option. (A) gastric angulus (B) gastric body (C) esophagus (D) gastric antrum	(C) esophagus
T2	TF	3 img	Is the following finding visible at this site: ulcer? Choose exactly one option. (A) visible (B) not visible	(A) visible
T2	MCQ-M	3 img	Select every abnormal finding visible in this image. More than one may be correct. (A) luminal dilatation (B) erosion (C) deformity (D) ulcer	(C) deformity; (D) ulcer
T3	MCQ-S	3 img	For the lesion in this gastric body image, what is the distribution? Choose exactly one option. (A) at a few sites (B) confluent (C) at several sites (D) solitary	(C) at several sites
T3	SA	3 img	In this duodenum image, what is the size of the ulcer?	approximately 1.2 cm in diameter
T4	RG	37 img	These images are from one upper-GI endoscopy. Write the endoscopic findings, organized site by site.	site-by-site findings report (gold = the endoscopist's own report)
T5	MCQ-S	18 img	Based on these endoscopic findings, which diagnosis is most appropriate? Choose exactly one option. (A) autoimmune gastritis (B) nodular gastritis (C) intestinal metaplasia (D) with bile reflux	(C) intestinal metaplasia

D.2 VISUAL EXAMPLES ACROSS TASKS

The examples pair T1 with T2 and T3 with T5 on shared pages, followed by full-page examples for T4 and T6.

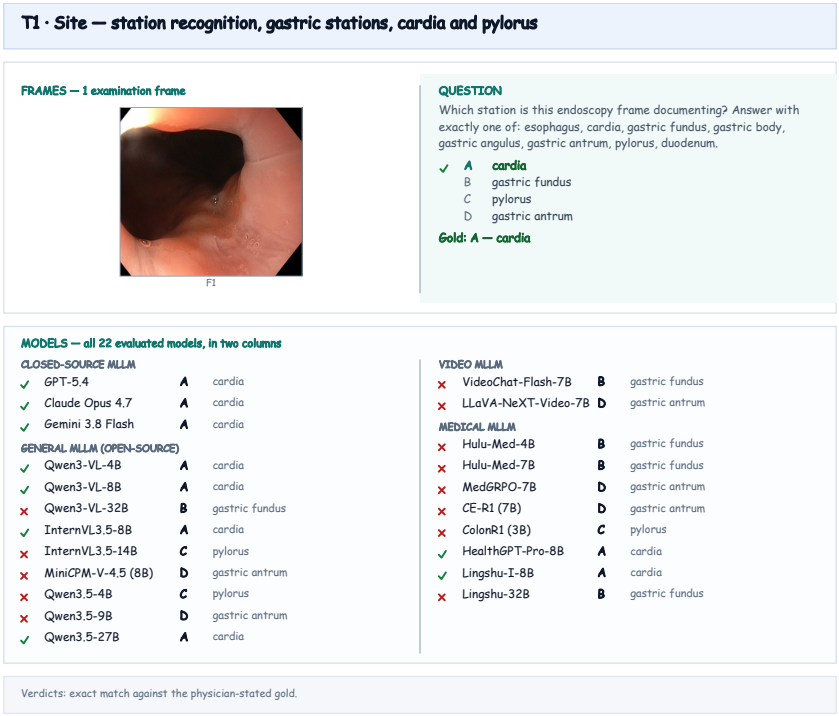


Figure 19: T1 Anatomy Recognition: representative frame, anatomy question, reference station, and model predictions.

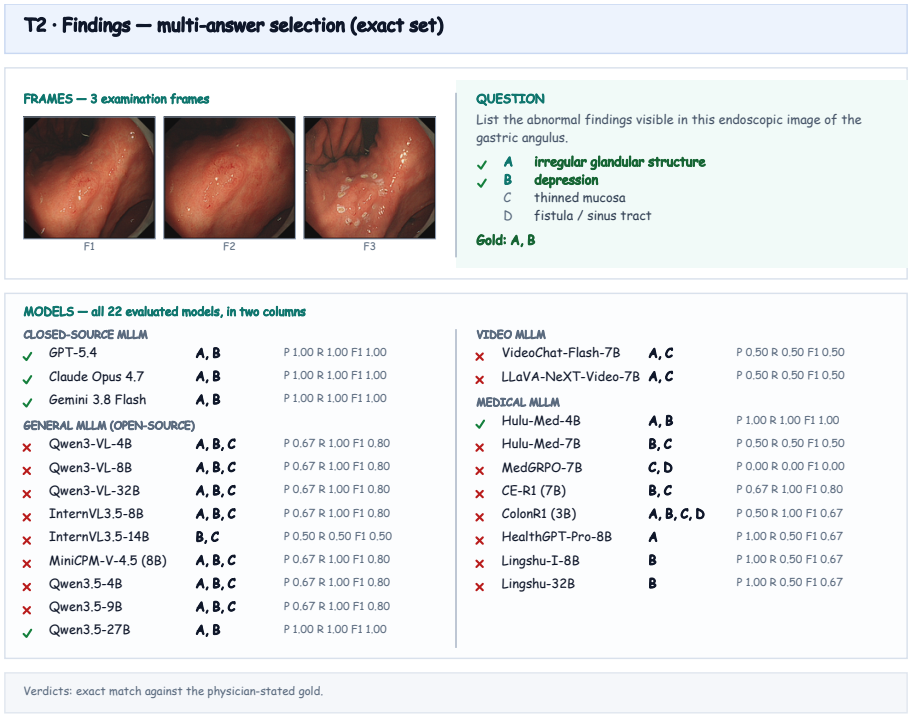


Figure 20: T2 Abnormal Finding Assessment: select-all example with three frames, the reference finding set, and model predictions.

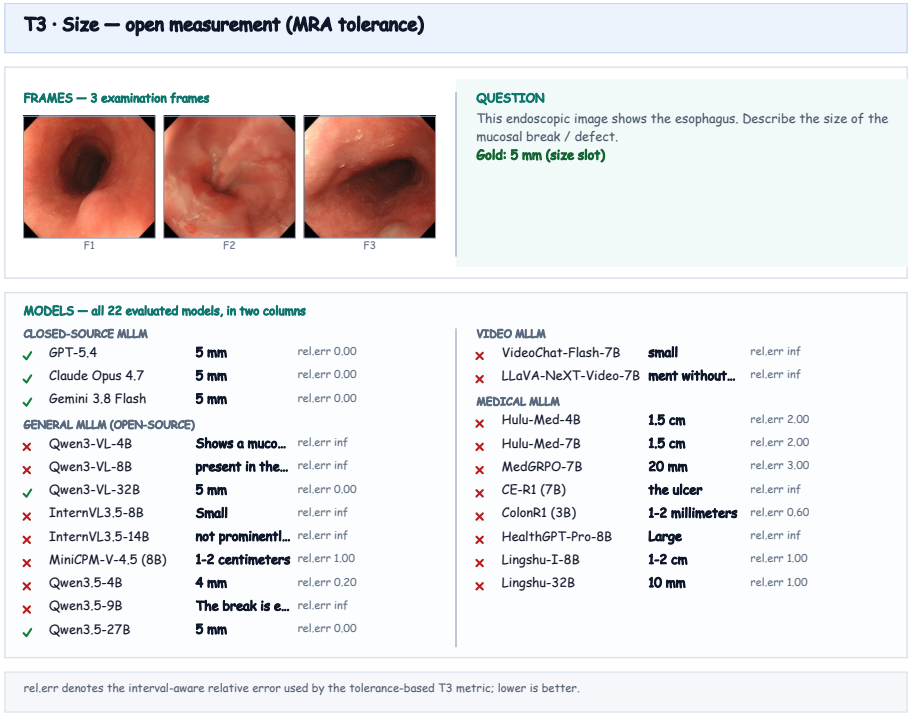


Figure 21: T3 Finding Attribute Specification: size-estimation example with three frames, reference measurement, and model predictions.

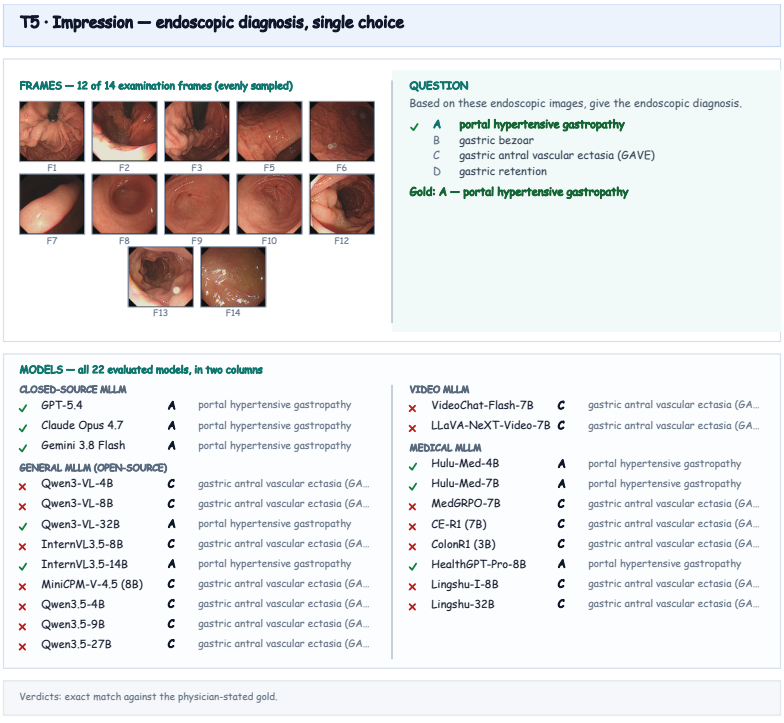


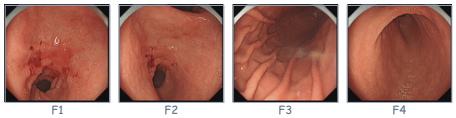
Figure 22: T5 Diagnostic Impression Prediction: single-choice impression example with selected images, reference diagnosis, and model predictions.



Figure 23: T4 Findings Report Generation: sampled examination frames, structured report reference, and representative model-generated reports.

T6 · Guideline CoT — grade + signs, with the reasoning chain

FRAMES — 4 examination frames



QUESTION

The following frames are an upper gastrointestinal endoscopy of one patient. Grade this case according to the Paris Classification + MESDA-G Guideline for Early Gastric Cancer, and state the signs your grade rests on.

Gold grade: 0-IIc

Gold signs: 0-III, 0-IIa, 0-IIb, irregular gland pattern, irregular vascular pattern, unclear border

MODELS — 6 representative models (four strongest and two lowest-scoring), verbatim

✓ GPT-5.4 — exact

grade 0-IIc · sign F1 0.29 (1/0/5) — 88 · A E E E E

[Visible findings]
F1 and F2 show a focal superficial depressed lesion in the distal stomach/antral-prepyloric region, with shallow mucosal depression, erythema, and small adherent blood/erosive change. The lesion appears non-polypoid and slightly excavated rather than protruded.
F1 and F2 suggest a localized abnormal area, but the margin is not sharply demarcated on these white-light views.
F3 and F4 show surrounding gastric mucosa without a clearly visible mass lesion; folds are preserved and there is no obvious diffuse severe atrophy or clear map-like metaplastic change in these frames.

[Sign analysis]
Irregular gland/pit pattern: absent/not assessable on these frames (F1, F2, F3, F4)
Irregular vascular pattern: absent/not assessable on these frames (F1, F2, F3, F4) …

✓ Claude Opus 4.7 — exact

grade 0-IIc · sign F1 0.40 (2/2/4) — 85 · A E E E S

[Visible findings]
F1: A depressed, irregular lesion with mucosal disruption, spontaneous bleeding/hemorrhagic spots, and converging folds around an ulcerated depression near the gastric angle/antrum. The lesion shows loss of normal mucosal pattern with a clearly demarcated depressed area.
F2: Similar lesion viewed from another angle, showing a well-defined depressed area with irregular margins, mucosal erosion, and hemorrhagic changes. The surrounding mucosa appears pale with visible vasculature suggesting atrophy.
F3: View of the gastric body showing preserved rugal folds with pale mucosa and increased visibility of submucosal vessels, consistent with background atrophic changes.
F4: Antrum/incisura view showing pale, thinned mucosa with loss of normal fold pattern, consistent with atrophic gastritis background.

[Sign analysis]
Irregular gland/pit pattern: present (F1, F2) …

✓ Gemini 3.8 Flash — exact

grade 0-IIc · sign F1 0.00 (0/1/6) — 80 · A E E S S

[Visible findings]
- F1: A well-demarcated, slightly depressed, erythematous lesion is visible in the prepyloric antrum with minor focal mucosal friability and bleeding.
- F2: Closer view of the antral lesion demonstrating a shallow depression with an irregular base and clear margin bordering the pyloric ring.
- F3: Normal gastric body showing preserved, regular longitudinal mucosal folds without suspicious focal lesions.
- F4: View of the gastric angle and lower body showing non-atrophic to mildly mottled background mucosa.

[Sign analysis]
Irregular gland/pit pattern: absent (F1-F4)
Irregular vascular pattern: absent (F1-F4)
Background atrophy: present (F1, F2, F4)
Background intestinal metaplasia: absent (F1-F4)
Unclear lesion border: absent (F1-F4) …

× Qwen3.5-27B — wrong

grade 0-IIc+IIa · sign F1 0.55 (3/2/3) — 65 · A D S S A

Based on the provided endoscopic images and the Paris Classification + MESDA-G Guideline for Early Gastric Cancer, here is the assessment:

[Visible findings]
The images display a lesion located in the gastric antrum/pyloric region. In F1 and F2, there is a distinct area of mucosal irregularity characterized by a reddish, slightly depressed or flat area with a rough surface texture. This area contrasts with the surrounding smoother, paler mucosa. There are visible linear erosions and small areas of hemorrhage (red streaks) within the lesion in F1 and F2. F3 shows the surrounding gastric folds which appear relatively normal, while F4 shows the adjacent mucosa which appears smooth and uniform, highlighting the localized nature of the abnormality seen in the first two frames.

[Sign analysis]
Irregular gland/pit pattern: present (F1, F2)
Irregular vascular pattern: present (F1, F2) …

× LLaVA-NeXT-Video-7B — wrong

grade 0-IIa · sign F1 0.00 (0/0/6) — 26 · F D D D F

[Visible findings]
F1: The mucosa appears normal, with no visible abnormalities.
F2: The mucosa appears normal, with no visible abnormalities.
F3: The mucosa appears normal, with no visible abnormalities.
F4: The mucosa appears normal, with no visible abnormalities.

[Sign analysis]
Irregular gland/pit pattern: Not present (F1-F4)
Irregular vascular pattern: Not present (F1-F4)
Background atrophy: Not present (F1-F4)
Background intestinal metaplasia: Not present (F1-F4)
Unclear lesion border: Not present (F1-F4)

[Guideline reasoning] …

× CE-R1 (7B) — wrong

grade – · sign F1 0.00 (0/0/6) — 10 · F F F F F

<start>
<answer>
<details>

<code>

<code>

<code>

<code>
 …

Verdicts: grade match against the physician-stated gold, sign F1 against the reference sign set (tp/fp/fn), and the judge composite (0-100) with band codes E/S/A/D/F.

Figure 24: T6 Guideline-Grounded Clinical Reasoning: grade-and-sign example with visual evidence and representative guideline reasoning chains.

E SCOPE AND INTERPRETATION

Clinical and institutional scope. TRACESCOPE is constructed from upper-GI examinations collected across multiple clinical centers, capturing variation in patient populations, acquisition practice, devices, and reporting conventions. This multi-center design provides a broader clinical evaluation setting than a single-institution benchmark, while the non-public source archive also reduces the risk that benchmark performance reflects overlap with publicly available pretraining data rather than task capability.

Interpretation of the diagnostic axes. The two diagnostic axes are designed for controlled within-case analysis. Evidence Presentation characterizes the response of a fixed diagnostic target to alternative visual presentations, while Diagnostic Cascade characterizes the response of the same pathology-backed decision as clinical state and guideline information are progressively exposed. Their estimand is therefore sensitivity to the information interface under a fixed downstream target. Axis-1 attention measures describe concentration over delivered visual units and are interpreted as model-attention descriptors rather than frame-level clinical relevance.

F DISCUSSION AND OUTLOOK

The stage-resolved results suggest that improving upper-GI multimodal systems is unlikely to reduce to a single model-scale or domain-specialization axis. Medical models lead selected finding and attribute readouts, whereas examination-level composition and rule-guided decisions are led by general-purpose or proprietary systems, highlighting that performance advantages remain stage-selective rather than workflow-wide.

The diagnostic axes further point to two concrete directions. First, the advantage of selected visual presentations persists under matched input count, motivating models that explicitly organize or retrieve examination evidence rather than relying on increasingly dense visual input. Second, the largest gain in the Diagnostic Cascade occurs when findings are mapped to guideline-sign semantics, motivating interfaces that represent clinical observations and decision rules explicitly before final prediction. Together, these results motivate upper-GI systems that combine evidence organization, structured clinical state, and rule-aware decision making.