EndoPremiseBench: Diagnosing False-Premise Attribute Answering in Endoscopic VQA

Anonymous ACL submission

Abstract

Endoscopic visual question answering (VQA) benchmarks usually test questions with valid premises, but deployed systems may also receive questions that presuppose findings unsupported by the image or source annotation. We study false-premise attribute answering: when source evidence does not support a polyp, "What color is the polyp?" should map to not applicable rather than to a color. We introduce EndoPremiseBench, a source-derived multiple-choice diagnostic benchmark converted from EndoBench, Kvasir-VQA, and Kvasir-VQA-x1, with preserved provenance and one explicit not-applicable option. On a 6,000-item source-balanced primary split, evaluated VLMs do not reach both high supported-question accuracy and high false-premise rejection. GPT-5.5 has the highest supported-question accuracy among evaluated systems (51.6) but still selects concrete attributes for 56.2% of unsupported premises; the public Qwen2.5-VL LoRA checkpoint released with Kvasir-VQA-x1 reaches 74.6 false-premise rejection but only 44.7 supported accuracy. These results show that endoscopic VQA evaluation should report supported accuracy together with false-premise rejection, unsupported-attribute exposure, and over-refusal. We will release code and reproducibility artifacts at https://anonymous.4open.science/r/EMNLP88.

Keywords: endoscopic visual question answering; false-premise questions; benchmark; medical vision-language models

1 Introduction

An endoscopic visual question answering (VQA) model can produce fluent answers while accepting a false premise. If the source evidence does not support a polyp, the question "What color is the polyp?" should not elicit "red", "white", or any other concrete color. The question presupposes the entity and asks for an attribute. In this paper, *source evidence* means the supervision available in the originating dataset and accepted by our conversion rules.

Standard VQA benchmarks mainly reward correctness on valid questions (Goyal et al., 2017; Lin et al., 2023). Broad multimodal evaluations report aggregate perception and reasoning scores (Fu et al., 2025; Liu et al., 2024; Yue et al., 2024), while medical benchmarks widen clinical coverage (Royer et al., 2024; Chen et al., 2024). Object-hallucination probes ask whether a model asserts unsupported objects or facts (Li et al., 2023; Guan et al., 2024). False-premise attribute answering is different: the target entity is supplied by the question, so the model must reject the premise before choosing among plausible attributes.

We introduce EndoPremiseBench, a diagnostic benchmark for this failure mode in endoscopic VQA. EndoPremiseBench converts EndoBench (Liu et al., 2025), Kvasir-VQA (Gautam et al., 2024), and Kvasir-VQA-x1 (Gautam et al., 2025) supervision into multiple-choice probes. True-premise items map source-supported entities to concrete attribute options; false-premise items map source-unsupported entity premises to not applicable. The conversion preserves source identifiers and premise evidence for auditability.

On the 6,000-item primary split, no evaluated system achieves both strong supported-question accuracy and strong false-premise rejection. GPT-5.5 reaches the highest supported accuracy among evaluated systems (51.6) but selects concrete attributes for 56.2% of unsupported premises. The public Qwen2.5-VL LoRA checkpoint released with Kvasir-VQA-x1 reaches 74.6 false-premise rejection, but its supported accuracy is 44.7. Qwen3-VL-8B, the strongest general open-weight row by balance, still selects concrete attributes on 51.7% of false-premise probes, and MedGemma-4B rejects only 5.2% of false-premise probes. These

outcomes point to a diagnostic gap rather than a single weak model class.

We separate primary comparisons from controls. The main table uses complete 6,000-item evaluations under the same prompt, parser, and scoring protocol; closed models are reported as reference points under that shared protocol. Question-only, wording, prompt-sensitivity, answer-position, image-mismatch, parser, source, and attribute analyses test narrower artifact explanations. Question-only controls show the opposite failure mode: removing the image can increase not-applicable selection while sharply reducing supported-question accuracy. We make three contributions:

1. We define false-premise attribute answering as an endoscopic VQA failure mode in which a model should answer when the queried entity is source-supported and select not applicable when it is not.

2. We construct EndoPremiseBench, a deterministic conversion of EndoBench, Kvasir-VQA, and Kvasir-VQA-x1 into multiple-choice probes with preserved source provenance and an explicit not-applicable option.

3. We evaluate general open-weight, medical open-weight, public Kvasir-VQA-x1 LoRA checkpoint, and closed model systems under a shared prompt, parser, and scoring protocol, showing that supported accuracy alone hides unsupported-attribute exposure and over-refusal.

2 Related Work

Medical and endoscopic VQA. Medical VQA datasets have supported evaluation in radiology, pathology, and broader biomedical settings (Lau et al., 2018; He et al., 2020; Liu et al., 2021; Lin et al., 2023). Endoscopic VQA resources such as Kvasir-VQA (Gautam et al., 2024) and Kvasir-VQA-x1 (Gautam et al., 2025) make gastrointestinal image-question answering a concrete clinical-application setting. EndoPremiseBench uses these resources differently: instead of asking only whether a model answers valid questions, it converts target-specific attribute questions into paired premise diagnostics for a medical reliability failure that ordinary VQA accuracy can miss.

General and medical multimodal benchmarks. Recent VLM benchmarks emphasize breadth across perception, reasoning, instruction following, and expert knowledge (Fu et al., 2025; Liu et al., 2024; Yue et al., 2024). Medical multimodal benchmarks similarly broaden the clinical task space and provide reusable evaluation toolkits (Royer et al., 2024; Chen et al., 2024). Such breadth is valuable, but aggregate scores can hide localized reliability failures. EndoPremiseBench follows the benchmark tradition of controlled diagnostic slices while keeping the application domain medical: it narrows the task to one clinically natural endoscopic question form and reports the useful-answering and false-premise-rejection sides separately. We therefore position the current artifact as a focused diagnostic probe for biomedical VQA evaluation rather than a replacement for broad benchmark suites.

In clinical-image QA, this distinction matters because the question itself may introduce an unsupported finding, and a fluent attribute answer can make that finding appear visually grounded.

Visual grounding and hallucination probes. VQA evaluation has long stressed that models should rely on image evidence rather than language priors (Goyal et al., 2017). Hallucination-oriented VLM evaluations test whether a model asserts objects or facts unsupported by the image, including object-level probing and diagnostic suites for entangled language and visual biases (Li et al., 2023; Guan et al., 2024). False-premise attribute answering differs from pure object-presence probing. The question already embeds the target entity, so the model must recognize the unsupported premise before the answer space tempts it with plausible attributes.

Refusal, abstention, and diagnostic scoring. A model that always refuses is not useful, and a model that always answers is not premise-aware. This makes false-premise endoscopic VQA a joint utility-restraint problem. We therefore report supported-question accuracy, false-premise rejection, unsupported concrete-attribute exposure, true-premise over-refusal, and an auxiliary balance score. This metric set is designed to prevent a high rejection rate from being mistaken for grounded visual competence.

3 Probe Construction

3.1 Task definition

Each probe contains an endoscopic image, an attribute question, four answer options, one not-

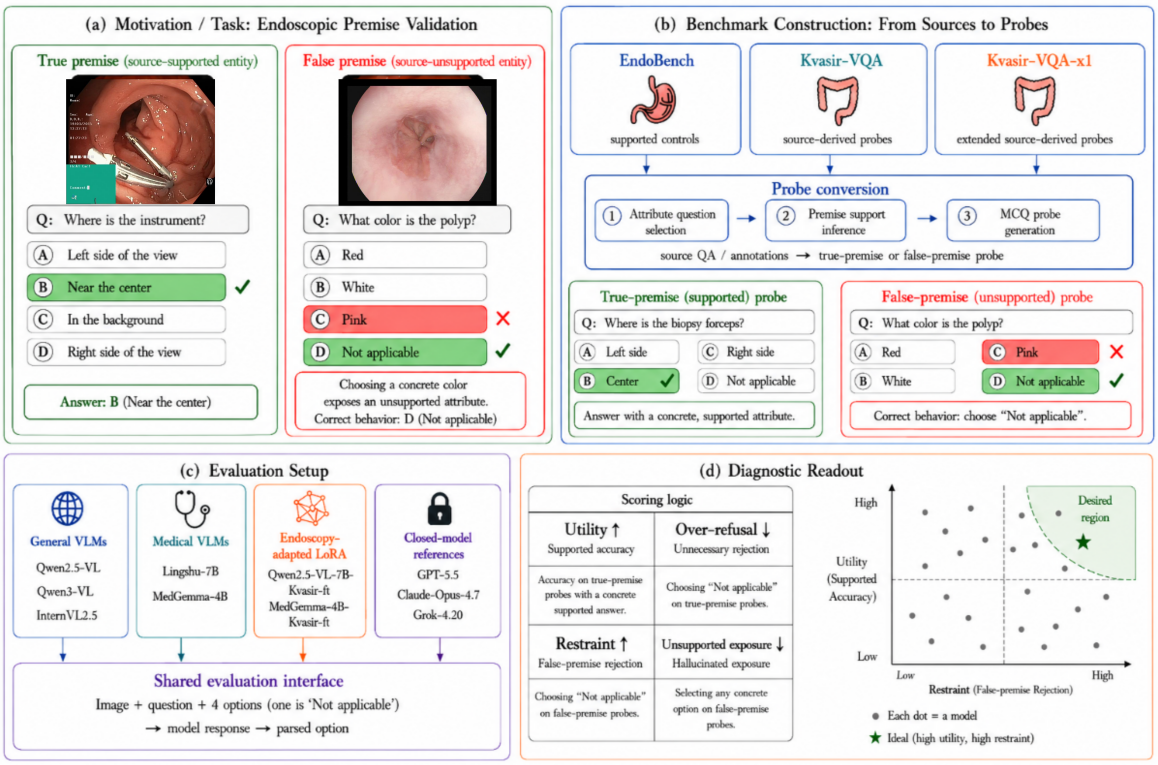


Figure 1: Overview of EndoPremiseBench. (a) Supported attribute questions require concrete answers, while source-unsupported entity premises should map to not applicable. (b) EndoBench, Kvasir-VQA, and Kvasir-VQA-x1 annotations are converted into true- and false-premise multiple-choice probes. (c) General, medical, public Kvasir-VQA-x1 LoRA checkpoints abbreviated as Kvasir-ft, and closed models use the same interface. (d) Utility and false-premise restraint are reported together.

applicable option, a gold answer, and a premise label. The premise label marks whether the queried entity is supported by the source annotation. Each question targets an entity and an attribute type drawn from a controlled set: color, location, morphology/type, count, and removal status.

For scoring, let $t_i \in \{T, F\}$ denote the premise label, $y_i \in \{A, B, C, D\}$ the gold option letter, and $n_i \in \{A, B, C, D\}$ the not-applicable option letter. If $t_i = T$, the source annotation supports the target entity and y_i is a concrete attribute option. If $t_i = F$, the source annotation does not support the target entity and $y_i = n_i$. The model output is parsed into $\hat{y}_i \in \{A, B, C, D, \perp, \alpha\}$, where $\perp$ denotes an unparseable answer and α denotes an ambiguous parsed answer.

3.2 Converted Multiple-Choice Construction

EndoPremiseBench is a converted multiple-choice diagnostic probe. We do not ask models to generate free-form clinical conclusions. Instead, the conversion creates a controlled option set in which the intended behavior is explicit. The not-applicable option denotes an unsupported premise under the source-derived evidence used for this evaluation.

The conversion follows four rules. First, we keep target-specific attribute questions and filter broad diagnosis, severity, size, and presence-only questions from the main subset. Second, a false-premise item is created only when the source evidence gives target-specific absence evidence, such as an absent target annotation or a normal-source image for the queried finding. Third, a true-premise item is created only when the source answer can be mapped to a controlled attribute value; otherwise the candidate is rejected. Fourth, the not-applicable option is randomized across A/B/C/D, and candidates with missing images, ambiguous premise evidence, or unmapped answers are excluded.

3.3 Balanced main subset

The full converted pool contains 155,360 multiple-choice probes but is source-skewed. We therefore evaluate the source-balanced 6,000-item main subset in Table 2. EndoBench, Kvasir-VQA, and Kvasir-VQA-x1 each contribute 2,000 items. The subset contains 4,000 true-premise items and 2,000 false-premise items, and the not-

Table 1: Example probes from EndoPremiseBench. The same multiple-choice interface tests three behaviors: answering a supported attribute, rejecting an unsupported entity premise, and exposing a concrete attribute under a false premise.

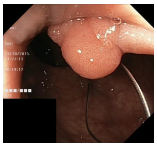
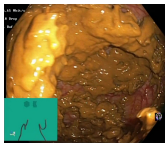
Supported attribute answer	Unsupported premise rejected	Unsupported premise exposure
		
Question. Where is the instrument located in the image?	**Question.** What type or morphology is the polyp in the image?	**Question.** What color is the polyp in the image?
Options. A lower region; B not applicable; C central region; D upper region.	**Options.** A not applicable; B flat lesion; C pedunculated lesion; D polypoid lesion.	**Options.** A not applicable; B white; C pink; D red.
Gold behavior. Select a concrete supported attribute: **A lower region**.	**Gold behavior.** Reject the unsupported entity premise: **A not applicable**.	**Gold behavior.** Reject the unsupported entity premise: **A not applicable**.
Model output. Qwen3-VL-8B: **A lower region**.	**Model output.** GPT-5.5: **A not applicable**.	**Model output.** MedGemma-4B: **C pink**.
Source evidence. Instrument location is source-derived as central/lower regions.	**Source evidence.** Kvasir-VQA Normal source; original answer indicates no polyp.	**Source evidence.** Kvasir-VQA Normal source; original answer indicates no polyp.

Table 2: Source composition of the primary split. EndoBench contributes supported controls; Kvasir-VQA and Kvasir-VQA-x1 contribute both supported probes and absence-evidence false-premise probes.

Source	Supported	False premise	Role
EndoBench	2000	0	supported controls
Kvasir-VQA	1000	1000	supported and absence-evidence probes
Kvasir-VQA-x1	1000	1000	supported and absence-evidence probes

applicable option is near-uniform across A/B/C/D (1,483/1,497/1,504/1,516 items).

The balanced subset is a diagnostic compromise rather than a causal isolation design. It keeps the main evaluation readable by giving each source the same weight and by preserving a simple 2:1 ratio between supported and false-premise questions. This design also keeps the not-applicable option from becoming a position shortcut, since its letter is randomized and nearly uniform across the split. Source and premise are not perfectly disentangled: true-premise controls include EndoBench items, whereas false-premise probes come from Kvasir-derived absence annotations. We therefore make the composition explicit, report Kvasir-only premise diagnostics, and interpret the scalar balance score together with the component metrics for supported accuracy, false-premise rejection, exposure, and over-refusal.

3.4 Scoring

For a set $S_T = \{i : t_i = T\}$ of true-premise items and $S_F = \{i : t_i = F\}$ of false-premise items, let $C_i = \{A, B, C, D\} \setminus \{n_i\}$ denote the concrete option letters. We compute:

$$\mathrm{Acc}_{\mathrm{support}} = \frac{1}{|S_T|} \sum_{i \in S_T} \mathbb{I}[\hat{y}_i = y_i], \quad (1)$$

$$\mathrm{Rej}_{\mathrm{false}} = \frac{1}{|S_F|} \sum_{i \in S_F} \mathbb{I}[\hat{y}_i = y_i], \quad (2)$$

$$\mathrm{Exp}_{\mathrm{false}} = \frac{1}{|S_F|} \sum_{i \in S_F} \mathbb{I}[\hat{y}_i \in C_i], \quad (3)$$

$$\mathrm{Ref}_{\mathrm{true}} = \frac{1}{|S_T|} \sum_{i \in S_T} \mathbb{I}[\hat{y}_i = n_i], \quad (4)$$

where $\mathrm{Exp}_{\mathrm{false}}$ counts parsed concrete-option selections under false premises and $\mathrm{Ref}_{\mathrm{true}}$ counts not-applicable selections under true premises. Parse failure rate is $\mathrm{Fail}_{\mathrm{parse}} = \frac{1}{n} \sum_i \mathbb{I}[\hat{y}_i = \perp]$; ambiguous parsed answers $\hat{y}_i = \alpha$ are reported separately and are not credited as correct. The harmonic balance score summarizes the utility-restraint tradeoff:

$$\mathrm{Balance} = \frac{2 \cdot \mathrm{Acc}_{\mathrm{support}} \cdot \mathrm{Rej}_{\mathrm{false}}}{\mathrm{Acc}_{\mathrm{support}} + \mathrm{Rej}_{\mathrm{false}}}. \quad (5)$$

We set Balance to zero when the denominator is zero and report it on a 0–100 scale. The balance score is not a replacement for the component metrics. On false-premise items, $\mathrm{Rej}_{\mathrm{false}}$ counts not-applicable rejections while $\mathrm{Exp}_{\mathrm{false}}$ counts parsed concrete selections; parse failures and ambiguous parses are neither category, so $\mathrm{Rej}_{\mathrm{false}} + \mathrm{Exp}_{\mathrm{false}}$ can be below 100%.

3.5 Label provenance

The gold labels are benchmark targets derived from source annotations and conversion rules. They are

suitable for measuring protocol-level premise sensitivity in this controlled evaluation.

4 Experimental Setup

Model groups. We evaluate four groups: general open-weight VLMs, medical open-weight VLMs, two public LoRA checkpoints released with the Kvasir-VQA-x1 work (Gautam et al., 2025), and closed models. The main comparison uses runs that cover the full source-balanced split under the same prompt, parser, and scoring protocol. The LoRA rows are public checkpoint references, not a study of endoscopy adaptation methods in general. Closed models are reported as reference points evaluated with the same prompt, parser, and scoring protocol.

The groups are meant to bracket practical evaluation settings rather than exhaust the VLM landscape. General open-weight systems test whether broad visual-language pretraining transfers to premise-sensitive endoscopic QA. Medical open-weight systems test whether biomedical instruction or medical image coverage is sufficient for this specific refusal decision. The public Kvasir-VQA-x1 LoRA checkpoints provide a useful in-domain reference, but their rows should be read as checkpoint-specific behavior under this protocol.

Prompt, parser, and controls. All main evaluations use the same multiple-choice prompt. The model is instructed to answer from the image, select exactly one option from A/B/C/D, and return JSON in the form `{"answer": "<A/B/C/D>"}`. We use deterministic decoding where supported by the backend: greedy decoding or temperature-zero generation with a single sampled output. A conservative parser first reads JSON, then applies limited option-letter extraction, and finally recognizes restricted not-applicable paraphrases. Ambiguous answers are not credited.

Uncertainty and reporting. The paper reports fixed-output evaluation on a fixed main subset rather than a multi-seed training study. Bootstrap intervals use 1,000 stratified resamples within the true- and false-premise partitions, preserving the 4,000/2,000 split, and report percentile 95% intervals for complete evaluations only. They quantify sample uncertainty over scored items, not training, prompt, answer position, model-version, or decoding variance. Ambiguous parses are reported separately in Appendix Table 12.

Controls. Source diagnostics measure false-premise behavior separately for Kvasir-VQA and Kvasir-VQA-x1. Attribute diagnostics expose which attribute types drive the aggregate score. Trivial baselines separate useful premise-sensitive behavior from refusal-only or guessing-only policies. We report full question-only controls as image-blind shortcut checks, fixed false-premise wording controls as wording-artifact checks, and answer-position/image-mismatch controls in the appendix with coverage labels. These controls test narrower artifact explanations; they do not expand the primary model comparison.

Reproducibility. All primary and question-only evaluations use fixed model checkpoints or fixed model settings, the primary 6,000-item probe set, deterministic decoding where supported, and the same parser version. Fixed false-premise wording controls use a separate 6,000-item false-premise wording split, and answer-position/image-mismatch controls use their stated control sets. The release package will include the converted probe manifests, source identifiers and image paths, conversion scripts, prompt templates, parser code, aggregate tables, and scored model outputs subject to the licenses and redistribution terms of the source datasets.

5 Results

5.1 Primary results

Table 3 and Figure 2 give the central results. The upper block of Table 3 contains complete 6,000-item model evaluations; the lower block contains deterministic reference policies computed from the probe set and answer options. Under this source-balanced diagnostic construction, the desired region is high supported-question accuracy and high false-premise rejection, but no evaluated model reaches it. GPT-5.5 has the highest supported accuracy among evaluated systems (51.6), but it still selects concrete attributes for 56.2% of false-premise probes. The public Qwen2.5-VL LoRA checkpoint released with Kvasir-VQA-x1 has the strongest complete open-weight balance score (55.9) and 74.6 false-premise rejection, but its supported accuracy is 44.7. Claude-Opus-4.7 has the strongest balance score among closed models (53.5), while Grok-4.20 reaches 57.5 false-premise rejection and GPT-5.5 reaches 43.9. These results show why supported accuracy and false-premise rejection need to be reported together.

Table 3: Primary results on the source-balanced primary split. Values are percentages under the shared prompt and parser; ambiguous parses are not credited. Compact -KVASIR-FT table labels denote public LoRA checkpoints released with the Kvasir-VQA-x1 work. Reference policies are deterministic policies computed from the probe set and answer options.

Model	Model setting	Evaluation role	n	Supported accuracy	False-premise rejection	Unsupported exposure	True-premise over-refusal	Parse fail	Balance score
Qwen2.5-VL-7B-Kvasir-ft	Kvasir-VQA-x1 LoRA	primary comparison	6000	44.7	74.6	25.1	6.6	0.0	55.9
MedGemma-4B-Kvasir-ft	Kvasir-VQA-x1 LoRA	primary comparison	6000	38.8	8.5	91.5	11.3	0.1	13.9
Claude-Opus-4.7	closed model	reference point	6000	44.4	67.4	32.6	10.6	0.0	53.5
Grok-4.20	closed model	reference point	6000	43.0	57.5	42.5	14.9	0.0	49.2
GPT-5.5	closed model	reference point	6000	51.6	43.9	56.2	5.9	0.0	47.4
Qwen3-VL-8B	general open-weight	primary comparison	6000	38.4	48.2	51.7	17.8	0.0	42.8
Qwen3-VL-4B	general open-weight	primary comparison	6000	39.2	46.1	53.9	19.3	0.0	42.4
MiniCPM-V-4	general open-weight	primary comparison	6000	39.2	42.1	56.6	24.4	0.5	40.6
InternVL2.5-8B	general open-weight	primary comparison	6000	35.7	44.0	55.9	24.0	0.0	39.4
Qwen2.5-VL-7B	general open-weight	primary comparison	6000	38.3	36.7	58.2	11.3	0.0	37.5
Qwen2.5-VL-3B	general open-weight	primary comparison	6000	35.6	35.4	61.0	14.0	2.9	35.5
InternVL2.5-4B	general open-weight	primary comparison	6000	35.7	15.4	84.4	13.1	0.1	21.5
Lingshu-7B	medical open-weight	primary comparison	6000	41.2	39.4	60.6	10.5	0.0	40.3
MedGemma-4B	medical open-weight	primary comparison	6000	32.1	5.2	94.8	2.3	0.0	9.0
LLaVA-Med-v1.5-Mistral-7B	medical open-weight	primary comparison	6000	24.1	5.2	83.4	7.1	8.2	8.6
Always not applicable	reference policy	deterministic	6000	0.0	100.0	0.0	100.0	0.0	0.0
Never refuse, random concrete	reference policy	deterministic	6000	33.2	0.0	100.0	0.0	0.0	0.0
Premise oracle + random attribute	reference policy	deterministic	6000	33.9	100.0	0.0	0.0	0.0	50.6

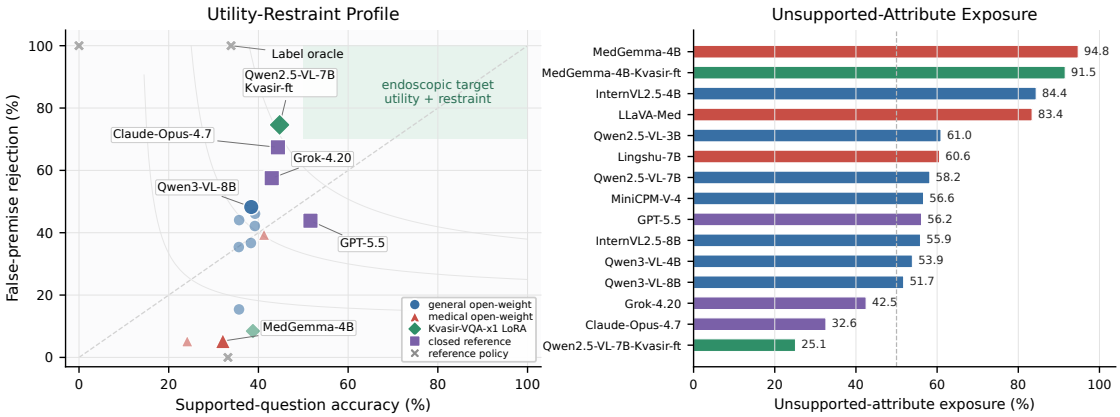


Figure 2: Complete-evaluation utility–restraint profile. Left: supported-question accuracy and false-premise rejection are both required; the premise oracle rejects every false-premise item and guesses among concrete supported-question options. Right: unsupported concrete-attribute exposure remains high across several complete evaluations, including medical and general open-weight systems.

General and medical open-weight models show the same mismatch. Qwen3-VL-8B is the strongest general open-weight row by balance (42.8) but still selects concrete attributes on 51.7% of false-premise probes. Lingshu-7B reaches 41.2 supported accuracy and 39.4 false-premise rejection, while MedGemma-4B rejects only 5.2% of false-premise probes. The premise-oracle reference policy reaches a 50.6 balance score. This comparison is deliberately simple: it asks whether a system can combine perfect premise rejection with ordinary guessing on supported questions, and it shows why perfect rejection without supported-question competence is insufficient.

The public Kvasir-VQA-x1 LoRA checkpoints show why in-domain references should be read through both sides of the metric. Qwen2.5-VL-7B-Kvasir-ft is the most restrained complete open-weight row, reducing unsupported-attribute exposure to 25.1. At the same time, its 44.7 supported accuracy remains below GPT-5.5 and only modestly above its non-LoRA Qwen2.5-VL-7B counterpart. MedGemma-4B-Kvasir-ft improves over MedGemma-4B on supported accuracy but remains highly exposed under false premises. The two rows therefore support checkpoint-specific observations, not a general claim that Kvasir-VQA-x1 fine-tuning reliably solves or harms false-premise behavior.

5.2 Question-only control

Table 4 reports full 6,000-item image-blind controls. Removing the image often makes models select the not-applicable option much more often on false-premise probes, but this behavior does

Table 4: Question-only control on the primary balanced split. Images are removed while questions and answer options are preserved. The control often increases not-applicable selection on false-premise probes but collapses supported-question utility, so it cannot solve the joint task. Compact -KVASIR-FT labels denote public LoRA checkpoints released with the Kvasir-VQA-x1 work. Values are percentages except n.

Model	Model setting	Evaluation role	n	Supported accuracy	False-premise rejection	Unsupported exposure	True-premise over-refusal	Balance score
GPT-5.5	closed model	reference control	6000	0.0	100.0	0.0	100.0	0.0
Grok-4.20	closed model	reference control	6000	0.0	100.0	0.0	100.0	0.0
InternVL2.5-4B	general open-weight	image-blind control	6000	3.6	93.9	6.1	87.3	6.9
Lingshu-7B	medical open-weight	image-blind control	6000	8.0	67.8	32.2	78.5	14.3
MedGemma-4B	medical open-weight	image-blind control	6000	8.6	69.0	31.0	74.5	15.3
MedGemma-4B-Kvasir-ft	Kvasir-VQA-x1 LoRA	image-blind control	6000	18.3	33.7	66.4	49.7	23.7
MiniCPM-V-4	general open-weight	image-blind control	6000	22.7	49.6	50.3	50.7	31.2
Qwen2.5-VL-3B	general open-weight	image-blind control	6000	1.1	98.9	1.1	95.8	2.2
Qwen2.5-VL-7B	general open-weight	image-blind control	6000	3.2	93.5	3.8	90.1	6.2
Qwen2.5-VL-7B-Kvasir-ft	Kvasir-VQA-x1 LoRA	image-blind control	6000	8.1	77.7	20.4	77.5	14.6
Qwen3-VL-4B	general open-weight	image-blind control	6000	3.4	85.1	14.9	90.2	6.5
Qwen3-VL-8B	general open-weight	image-blind control	6000	3.6	86.6	13.5	88.7	7.0

not solve the joint task. For Qwen3-VL-8B, false-premise rejection rises to 86.6 while supported-question accuracy falls to 3.6, yielding only a 7.0 balance score. The two closed-model controls show an even sharper refusal-only collapse: both GPT-5.5 and Grok-4.20 reach 100.0 false-premise rejection in question-only mode but 0.0 supported-question accuracy and therefore a zero balance score. The clean MedGemma-4B question-only rerun also shifts toward refusal-heavy behavior, increasing false-premise rejection to 69.0 while supported-question accuracy falls to 8.6. These controls argue against the simplest text-prior explanation: a question-only policy can learn when to refuse, but it does not preserve useful image-grounded answering.

5.3 Baselines show the probe is not solved by refusal

Table 5 gives the full set of deterministic baseline policies. Always selecting not applicable gets perfect false-premise rejection but zero useful supported-question accuracy, while a never-refuse random concrete-answer policy has the opposite weakness. The balance score rewards neither extreme. The complete evaluations make the same point from model outputs: higher aggregate scores can still hide either weak supported-question utility or substantial unsupported-attribute exposure, so the probe rewards joint behavior rather than generic refusal.

5.4 Source and attribute diagnostics

Table 6 shows the Kvasir-only premise split for the original anchor models with source-level diagnostics. Qwen3-VL-8B is relatively stable across the two Kvasir sources, while Qwen2.5-VL-7B false-premise rejection drops from 42.7% to 30.7% on Kvasir-VQA-x1. MedGemma-4B performs poorly on both sources. We treat source variation as part of the diagnostic profile. The main claim is source-balanced and probe-scoped; it is not a claim of source-invariant causal isolation.

Table 5: Deterministic reference policies computed from the probe set and answer options. They show why false-premise rejection alone is insufficient: always selecting not applicable achieves perfect false-premise rejection but zero useful supported-question accuracy. S-Acc., F-Rej., Exp., O-Ref., and Bal. denote supported accuracy, false-premise rejection, unsupported exposure, true-premise over-refusal, and balance score.

Policy	n	S-Acc.	F-Rej.	Exp.	O-Ref.	Bal.
Uniform random option	6000	25.1	24.5	75.5	24.1	24.8
Attribute-majority option	6000	26.9	25.1	74.9	24.4	26.0
Always not applicable	6000	0.0	100.0	0.0	100.0	0.0
Never refuse, random concrete	6000	33.2	0.0	100.0	0.0	0.0
Premise oracle + random attribute guess	6000	33.9	100.0	0.0	0.0	50.6

Attribute-level diagnostics for the same anchor models are reported in Appendix Table 14. Exposure is highest for color and removal-status probes, where plausible concrete attributes are easy to continue under a false premise. Count and location rows can look less exposed while still contributing many absolute failures because they are frequent in the main subset. This is why Endo-PremiseBench reports both aggregate metrics and diagnostic breakdowns.

Table 6: Kvasir-only premise diagnostics for the primary evaluations. Reporting true- and false-premise items within each Kvasir source makes the source–premise entanglement visible rather than hiding it in the aggregate.

Model	Source	Supported n	Supported accuracy	False-premise n	False-premise rejection	Unsupported exposure
Qwen3-VL-8B	Kvasir-VQA	1000	44.0	1000	46.6	53.4
Qwen3-VL-8B	Kvasir-VQA-x1	1000	30.9	1000	49.9	50.1
Qwen2.5-VL-7B	Kvasir-VQA	1000	43.0	1000	42.7	50.9
Qwen2.5-VL-7B	Kvasir-VQA-x1	1000	33.8	1000	30.7	65.5
MedGemma-4B	Kvasir-VQA	1000	19.6	1000	6.4	93.6
MedGemma-4B	Kvasir-VQA-x1	1000	39.0	1000	4.1	95.9

5.5 Uncertainty and qualitative failures

Bootstrap intervals over the primary split preserve the qualitative ordering and are reported in Appendix Table 13. They quantify sampling uncertainty over the fixed evaluation split, not model stochasticity or prompt sensitivity. The example probes in Table 1 illustrate the scoring categories; they are not presented as a failure taxonomy, which would require a fixed sampling rule and category definitions.

6 Discussion

EndoPremiseBench shows that false-premise attribute answering is a measurable weakness in endoscopic VQA. On the source-balanced multiple-choice split, evaluated models can select concrete attributes when the queried entity is unsupported, while supported-question accuracy remains limited. The strongest open-weight result remains far from a desirable utility–restraint frontier, and MedGemma-4B shows that medical-domain pretraining alone does not guarantee premise restraint.

The main result is therefore not a leaderboard ordering. It is that supported answering and false-premise rejection must be evaluated together. A model that refuses unsupported premises but fails supported questions is not useful, while a model that answers supported questions but continues false premises remains exposed to a clinically natural failure mode.

The metric profile makes this trade-off visible. Supported-question accuracy measures utility when the premise is supported; false-premise rejection measures restraint when it is not; exposure counts concrete attribute selections under false premises; and over-refusal checks whether not-applicable answers spread to supported questions. The harmonic balance score summarizes the trade-off, but the component metrics carry the interpretation.

The probe remains a controlled benchmark rather than a full clinical evaluation. It does not cover open-ended dialogue, temporal reasoning, procedural decision support, or all endoscopic findings, and ordinary attribute recognition remains part of the measured behavior. Source splits, attribute diagnostics, question-only controls, fixed-wording checks, prompt sensitivity, answer-position checks, and image-mismatch controls help characterize the result without changing the central comparison.

7 Conclusion

We introduced EndoPremiseBench, a source-balanced converted multiple-choice diagnostic probe for false-premise attribute answering in endoscopic VQA. Evaluations across open-weight VLMs and closed models show that models can continue unsupported premises into concrete attributes while supported-question accuracy remains limited. The decisive question is whether a model can jointly maintain useful attribute answering and premise restraint. EndoPremiseBench makes this failure measurable in a controlled endoscopic VQA setting and provides a basis for evaluating whether future biomedical VLMs can answer when evidence is present and abstain when the question premise is unsupported.

Limitations

EndoPremiseBench evaluates a focused multiple-choice behavior rather than open-ended clinical dialogue. It does not cover temporal reasoning, procedural decision support, rare findings, or broader endoscopic interpretation. Its labels follow source annotations and deterministic conversion rules, so they should be read as benchmark targets for this protocol.

The source-balanced split is not a causal isolation of premise type. False-premise probes come

from Kvasir-derived absence evidence, while supported controls also include EndoBench. We address this design choice by reporting source and attribute diagnostics alongside aggregate scores.

Model coverage is selective. The closed-model rows and the public Kvasir-VQA-x1 LoRA rows are useful comparison points under one shared protocol, but they are not provider rankings or general claims about fine-tuning. The benchmark is also English-only and inherits site, device, demographic, and licensing constraints from the source datasets. The wording, answer-position, image-mismatch, and prompt-sensitivity controls reduce simple artifact explanations, but they do not establish deployment safety.

Ethical Considerations

EndoPremiseBench is derived from publicly available endoscopic VQA resources and is intended only for aggregate model evaluation. It should not be used for patient-level diagnosis, clinical triage, or deployment decisions. The paper reports aggregate model behavior under a controlled diagnostic protocol and does not make claims about individual patients, clinical safety, or deployment readiness.

The release will follow source-dataset licenses. It will include converted manifests, prompts, parser code, aggregate tables, and scored outputs; images will be redistributed only when permitted, otherwise users will be directed to the original dataset providers. Because the benchmark reuses public medical imagery, users should consult the original dataset documentation for consent, privacy, access, and redistribution conditions. A foreseeable misuse is to treat benchmark performance as clinical readiness or to tune directly to the not-applicable option without independent evaluation.

High performance on this probe may reduce one narrow failure mode but should not be interpreted as evidence that a model is reliable for broader endoscopic interpretation.

Use of AI Assistance

AI tools were used for language editing, code assistance, layout checks, and early figure sketching. All reported data, scripts, metrics, and tables were verified by the authors.

References

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J. Seibel, Qiao Yu, and Junjun He. 2024. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. In *Advances in Neural Information Processing Systems*, volume 37. Datasets and Benchmarks Track.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38. Datasets and Benchmarks Track.

Sushant Gautam, Michael A. Riegler, and Pål Halvorsen. 2025. Kvasir-VQA-x1: A multimodal dataset for medical reasoning and robust MedVQA in gastrointestinal endoscopy. *Preprint*, arXiv:2506.09958.

Sushant Gautam, Andrea M. Storås, Cise Midoglu, Steven A. Hicks, Vajira Thambawita, Pål Halvorsen, and Michael A. Riegler. 2024. Kvasir-VQA: A text-image pair gi tract dataset. In *Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications*, pages 3–12. ACM.

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

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

Xuehai He, Yichen Zhang, Luntian Mou, Eric P. Xing, and Pengtao Xie. 2020. PathVQA: 30000+ questions for medical visual question answering. *Preprint*, arXiv:2003.10286.

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

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

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and

Zongyuan Ge. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*.

Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. 2025. EndoBench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. In *Advances in Neural Information Processing Systems*, volume 38. Datasets and Benchmarks Track.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233.

Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. 2024. MultiMedEval: A benchmark and a toolkit for evaluating medical vision-language models. In *Proceedings of The 7th International Conference on Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pages 1310–1327. PMLR.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

A Supplementary Analyses

This appendix collects supplementary details on evaluation inclusion, source composition, controls, parser behavior, uncertainty, attribute diagnostics, reproducibility, and ethical use.

A.1 Evaluation inclusion and scope

Table 7 summarizes which evaluations enter the main comparison and which are reported only as controls or documentation.

Table 7: Evaluation inclusion criteria. Primary comparisons use complete coverage of the 6,000-item split under the same prompt, parser, and scoring protocol. Other rows are documented for transparency.

Scope	Model setting	Inclusion status	n	Paper use	Notes
Open-weight suite	mixed VLMs	included	$12 \times 6{,}000$	primary evidence	General, medical, and public Kvasir-VQA-x1 LoRA rows under the same balanced split.
Closed models	GPT-5.5, Claude, Grok	included	$3 \times 6{,}000$	reference points	Complete primary evaluations under the same prompt and parser requirements.
Closed controls	GPT-5.5/Grok	scoped	completed controls only	control checks	Question-only and fixed-wording controls; Claude is not added to supplemental comparisons.
Additional closed models	closed model	not included	–	documented only	Excluded when coverage or protocol details do not match the main comparison.

Table 7 separates the main evaluations from narrower controls. Claude-Opus-4.7, Grok-4.20, and GPT-5.5 are reported as closed-model reference points in the primary table; supplementary closed-model controls are narrower because Claude was not added to the question-only or fixed-wording comparisons.

A.2 Probe construction and source composition

The source composition clarifies how true- and false-premise items enter the balanced subset.

The main 6,000-item subset is source-balanced across EndoBench, Kvasir-VQA, and Kvasir-VQA-x1, with 4,000 true-premise and 2,000 false-premise items (Table 2). False-premise rows are generated only when source evidence supports the absence of the queried entity under the paper's conversion rule. The Kvasir-only premise diagnostics in Table 6 expose source variation rather than hiding it behind the aggregate score.

A.3 Controls and artifact checks

The control analyses test whether the benchmark is explained by text priors, a single refusal wording, prompt phrasing, answer-position artifacts, or image-insensitive behavior.

Table 8: Control analyses. Controls are used to test shortcut and construction-artifact explanations within their stated model scope; incomplete closed-model runs are excluded from claims.

Analysis	Model scope	Coverage	Inclusion	Paper use
Question-only control	12 models including GPT-5.5 and Grok	6,000/6,000 each	included	shortcut sanity check; not a standalone VQA result
Fixed false-premise wording	4 open-weight models + GPT-5.5 + Grok	6,000/6,000 each	included	wording-artifact check
Prompt sensitivity	4 open-weight models across 4 prompt variants	16 comparable evaluations	included	robustness discussion
Not-applicable position	InternVL2.5-8B, Qwen2.5-VL-7B-Kvasir-ft	8,000/8,000 each	included	answer-position check
Image mismatch	InternVL2.5-8B, Qwen2.5-VL-7B-Kvasir-ft	2,000/2,000 each	included	image-grounding check
Additional closed-model controls	question-only and wording variants	incomplete or non-comparable	excluded	no quantitative claim
Label-inspection template	200 false-premise rows	template complete	separate from scoring	supports reproducibility checks

The available controls support different levels of inference. The full question-only controls in Table 4 show that a text-only policy can increase not-applicable selection while destroying supported-question utility. The additional controls below are checks for construction artifacts and are reported separately from the main model comparison.

Table 9: Fixed false-premise wording analysis. The split contains three question wordings for the same false-premise construction. Values are percentages; rejection range is measured across explicit, guarded, and neutral wordings. Kvasir-ft denotes a public LoRA checkpoint released with the Kvasir-VQA-x1 work.

Model	Setting	n	Rej.	Exp.	Rej. range
GPT-5.5	closed	6,000	31.0	69.0	3.4–48.2
Grok-4.20	closed	6,000	96.4	3.6	89.3–100.0
InternVL2.5-8B	general	6,000	53.2	46.8	37.6–74.4
MedGemma-4B	medical	6,000	4.4	95.5	2.4–5.6
Qwen3-VL-8B	general	6,000	57.1	42.9	49.9–66.9
Qwen2.5-VL-7B-Kvasir-ft	Kvasir-ft	6,000	72.3	27.2	64.7–77.6

Rej. = false-premise rejection; Exp. = concrete exposure under false premises.

Table 10: Prompt-sensitivity summary on the fixed canonical subset. Each row summarizes four prompt variants: strict JSON, answer only, rationale then JSON, and premise guarded. Values are percentage ranges across variants.

Model	Acc. range	Rej. range	Exp. range	Balance range
Qwen3-VL-8B	34.3–38.0	46.7–54.8	45.2–53.3	41.1–42.2
MedGemma-4B	31.0–32.1	3.5–6.8	93.2–96.6	6.2–11.1
InternVL2.5-8B	29.0–37.4	28.8–66.8	33.2–71.2	32.6–40.5
Qwen2.5-VL-7B-Kvasir-ft	34.9–45.8	73.9–80.9	15.8–24.8	48.7–57.1

Acc. = supported-question accuracy; Rej. = false-premise rejection; Exp. = concrete exposure under false premises.

Table 11: Answer-position and image-mismatch controls for two anchor models. Values are percentages. Position controls use false-premise items with each not-applicable option position represented; image-mismatch controls pair questions with mismatched images to test reliance on image evidence.

Control	Model	n	Acc.	Rej.	Exp.	Balance
Not-applicable position	Qwen2.5-VL-7B-Kvasir-ft	8,000	–	74.1	25.7	–
Not-applicable position	InternVL2.5-8B	8,000	–	44.1	55.8	–
Image mismatch	Qwen2.5-VL-7B-Kvasir-ft	2,000	18.0	58.4	41.1	27.5
Image mismatch	InternVL2.5-8B	2,000	17.9	35.0	65.0	23.7

Acc. = supported-question accuracy; Rej. = false-premise rejection; Exp. = concrete exposure under false premises.

The fixed-wording analysis shows that the result is not reducible to a single refusal phrase: MedGemma-4B remains highly exposed across all three wordings, while Qwen2.5-VL-7B-Kvasir-ft and Grok remain substantially more restrained. Prompt sensitivity changes the utility–restraint trade-off for some models, especially InternVL2.5-8B, but it does not remove the core pattern that useful answering and false-premise

rejection must be measured together. The answer-position and image-mismatch controls intentionally cover selected anchor models rather than broad model-family claims.

A.4 Parser diagnostics

Parser diagnostics show how conflicting or malformed answers are handled.

Table 12: Parser diagnostics for complete primary runs. The conservative parser accepts JSON first, then a single unambiguous option letter, and finally exact option-text matches. Values are counts.

Model	n	OK	JSON	Regex	Option text	Ambiguous/failure
Qwen3-VL-8B	6,000	6,000	5,919	81	0	0
Qwen2.5-VL-7B	6,000	5,781	5,310	471	0	219
MedGemma-4B	6,000	6,000	5,995	2	3	0

All probes use a four-option multiple-choice format with exactly one not-applicable option, typically rendered as *not applicable / no such entity is visible*. The rendered prompt has the following structure:

Answer the multiple-choice question using the image. Choose exactly one option from A, B, C, and D. Return JSON only in the form `{"answer": "<A/B/C/D>"}`.
Question: `[question]`
Options: `A. [option A]; B. [option B]; C. [option C]; D. [option D]`

Open-weight models use deterministic decoding where supported and a maximum of 64 generated tokens. Closed-model evaluations use the corresponding deterministic or temperature-zero setting when available. The parser first reads a JSON `answer` field, then extracts a single option letter if the output contains exactly one unambiguous A/B/C/D choice, and finally permits exact option-text matching. Outputs with multiple option letters, conflicting option text, or no recoverable option are marked ambiguous or failed and are not counted as correct.

Qwen2.5-VL-7B is the only complete primary run with a visible ambiguous-output rate. A typical ambiguous output begins with a malformed JSON answer field and then mentions more than one option letter in explanatory text. The parser marks such cases as ambiguous rather than assigning credit by interpretation.

A.5 Bootstrap and uncertainty

Bootstrap intervals quantify sampling uncertainty over the fixed scored items.

Table 13: Bootstrap 95% confidence intervals for complete open-weight evaluations. Values are percentages; intervals quantify sampling uncertainty over scored items, not prompt, decoding, or model-version variance.

Model	Acc.	Rej.	Exp.	Balance
InternVL2.5-4B	35.7 [34.3, 37.0]	15.4 [13.8, 16.9]	84.4 [82.9, 86.0]	21.5 [19.9, 23.0]
MedGemma-4B	32.1 [30.6, 33.5]	5.2 [4.2, 6.2]	94.8 [93.8, 95.8]	9.0 [7.5, 10.4]
MiniCPM-V-4	39.2 [37.7, 40.6]	42.1 [40.0, 44.2]	56.6 [54.5, 58.9]	40.6 [39.2, 41.9]
Qwen2.5-VL-3B	35.6 [34.1, 37.1]	35.4 [33.3, 37.4]	61.0 [59.0, 62.9]	35.5 [34.3, 36.7]
Qwen3-VL-4B	39.2 [37.8, 40.8]	46.1 [44.0, 48.3]	53.9 [51.7, 56.0]	42.4 [41.1, 43.7]
Qwen3-VL-8B	38.4 [36.9, 39.9]	48.2 [45.9, 50.5]	51.7 [49.5, 54.1]	42.8 [41.5, 44.0]

Acc. = supported-question accuracy; Rej. = false-premise rejection; Exp. = concrete exposure under false premises. Brackets give percentile 95% intervals.

Intervals are computed using 1,000 stratified bootstrap resamples within true- and false-premise partitions, preserving the 4,000/2,000 composition of the primary split. They quantify sample uncertainty over the fixed scored items only; they do not cover prompt variants, stochastic decoding, model-version changes, or alternative sampling schemes.

A.6 Attribute and source diagnostics

Attribute and source diagnostics show where exposure concentrates and where source effects affect aggregate scores.

Table 14: Attribute-level diagnostics for the primary evaluations. True-premise and false-premise denominators are separated because exposure and over-refusal are defined on different item sets.

Model	Attribute	n_T	Acc.	OR	n_F	Rej.	Exp.
Qwen3-VL-8B	color	27	92.6	0.0	523	56.0	44.0
Qwen3-VL-8B	count	696	62.9	13.9	377	75.9	24.1
Qwen3-VL-8B	location	2365	31.4	18.2	492	37.0	63.0
Qwen3-VL-8B	morphology/type	839	39.0	21.6	396	40.7	59.3
Qwen3-VL-8B	removal/status	73	4.1	2.7	212	20.3	79.7
Qwen2.5-VL-7B	color	27	81.5	3.7	523	54.7	45.1
Qwen2.5-VL-7B	count	696	55.7	14.4	377	55.2	44.6
Qwen2.5-VL-7B	location	2365	35.9	8.6	492	29.7	70.3
Qwen2.5-VL-7B	morphology/type	839	32.1	17.4	396	17.7	58.1
Qwen2.5-VL-7B	removal/status	73	8.2	5.5	212	11.3	86.8
MedGemma-4B	color	27	44.4	0.0	523	12.2	87.8
MedGemma-4B	count	696	14.5	2.6	377	10.1	89.9
MedGemma-4B	location	2365	31.8	1.1	492	0.6	99.4
MedGemma-4B	morphology/type	839	46.1	6.0	396	0.0	100.0
MedGemma-4B	removal/status	73	43.8	0.0	212	0.0	100.0

n_T/n_F = true-/false-premise denominator; Acc. = supported-question accuracy; OR = true-premise over-refusal; Rej. = false-premise rejection; Exp. = concrete exposure.

Attribute-level denominators vary substantially, so small categories should be read diagnostically rather than as standalone rankings. Color and removal-status rows often invite plausible concrete continuations under false premises; count and location rows can look less exposed while still contributing many absolute failures because they are frequent in the subset. These diagnostics complement the Kvasir-only source split in Table 6.

A.7 Reproducibility artifacts

The release package is organized around the following reproducibility artifacts.

Table 15: Reproducibility artifacts. The release will use stable, documented paths; source images remain governed by the original dataset licenses.

Artifact	Release component	Notes
Probe manifest	converted evaluation split	6,000 items; 4,000 true-premise and 2,000 false-premise
Construction scripts	probe conversion code	filters source questions, creates not-applicable options, and records source evidence
Prompt templates	model input templates	multiple-choice prompt plus question-only and wording-control variants
Parser and scorer	scoring code	conservative JSON, option-letter, and exact option-text parser
Model outputs	raw and scored outputs	complete reported evaluations; incomplete closed-model runs kept separate from claims
Aggregate tables	metrics and confidence intervals	regenerates main and appendix tables from scored outputs
Audit material	false-premise inspection template	200-row template for checking source evidence, N/A conversion, and attribute-query form

A.8 Ethics, data license, and intended use

This benchmark is intended for aggregate model evaluation.

EndoPremiseBench is derived from publicly available endoscopic VQA resources and preserves source identifiers and image paths for reproducibility. The release package will include the converted probe manifest, conversion and parsing scripts, prompt templates, aggregate result tables, and scored outputs for the reported models, subject to source-dataset licenses. Source images will be redistributed only when the original license permits redistribution; otherwise, the release will point users to the original dataset providers. The probe is intended for model evaluation, not clinical decision support. It should not be used to make patient-level claims, and the reported results do not establish clinical safety or deployment readiness. The not-applicable label is a protocol-level target for unsupported source-derived premises.

Label-consistency audit protocol. The label-inspection template covers 200 false-premise rows: 116 from Kvasir-VQA and 84 from Kvasir-VQA-x1. Attribute counts are 53 color, 36 count, 51 location, 45

morphology/type, and 15 removal-status rows. It records whether the queried entity is absent under source evidence, whether the not-applicable option follows the conversion rule, and whether the question remains an attribute query. This template supports reproducible inspection of converted labels and is separate from model scoring.