

Measuring and Improving Complex-Atomic Answer Consistency in Endoscopic VQA

Yuhao Liu, Cheng Zhao, Guanghui Yue

School of Biomedical Engineering, Shenzhen University, Shenzhen, China

Abstract—Endoscopic visual question answering (VQA) increasingly asks complex questions that combine several endoscopic answer components rather than isolated factual queries. Such complex answers may be scored as correct even when the same model fails on associated atomic questions. We introduce EndoCA, a paired complex-atomic answer consistency benchmark for evaluating whether complex answers remain consistent with same-image atomic answers. EndoCA contains two suites: EndoCA-Core evaluates compact question-complexity patterns commonly seen in practical endoscopic VQA, and EndoCA-Diagnostic supports controlled analysis across increasing question complexity. We evaluate 11 VLMs spanning open, medical, endoscopy-adapted, and closed-source models on EndoCA. Some VLMs achieve high complex-answer accuracy, yet their atomic-answer accuracy and complex-atomic answer consistency remain substantially lower. To reduce this complex-atomic inconsistency, we introduce Atomic-Support Reconciliation (ASR), a training-free mechanism that uses model-generated atomic answers as contextual premises for answer revision and consistency-guided selective answering. On four selected publicly available models, ASR-Revise improves paired complex-atomic correctness with modest changes in complex-answer accuracy, while ASR-Selective improves accuracy on answered cases by allowing the model to abstain from less reliable cases. Together, EndoCA and ASR provide a consistency-aware benchmark and a training-free mechanism for answer reconciliation and selective answering in endoscopic VQA.

Index Terms—Medical visual question answering, endoscopy, vision-language models, answer consistency, selective prediction.

I. INTRODUCTION

Visual question answering (VQA) provides a task-directed way to evaluate visual understanding: a model must interpret an image under a natural-language query and return an answer grounded in the relevant visual evidence [1]. Medical VQA extends this paradigm to clinically oriented image interpretation [2]. With stronger vision-language models (VLMs), endoscopic VQA is moving from isolated recognition prompts, such as whether a polyp is present, toward complex questions that combine multiple visual components. We use the term *atomic question* to denote a same-image component check for one visual commitment, such as an anatomical site, a polyp count, or an instrument type. A complex question can combine several such checks into one natural answer.

Recent endoscopic VLM benchmarks reflect this broader evaluation trend. For example, EndoBench evaluates models across multiple endoscopic scenarios, clinical tasks, and visual prompting granularities [3]. Such benchmarks provide important final-answer and task-performance signals. At the same time, when a complex question involves multiple component judgments, each component can be checked through a simpler

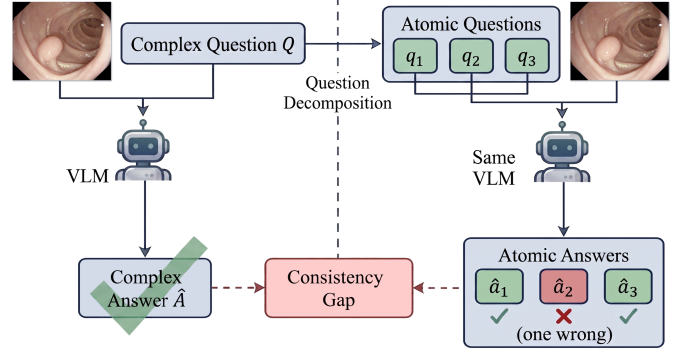


Fig. 1. Paired complex-atomic evaluation. A correct complex answer can still disagree with associated atomic answers on the same image.

atomic question. If the same model gives a correct complex answer but fails corresponding atomic questions under the same image and related question context, its image understanding and evidence use remain unstable in a way that complex-answer accuracy can obscure. The motivating example in Fig. 1 illustrates this gap: the complex answer is accepted while one corresponding atomic answer is wrong. In medical settings, this hidden inconsistency matters because apparently correct answers may mask unreliable interpretation of the underlying endoscopic evidence.

We construct EndoCA from the two-level QA structure formed by Kvasir-VQA [4] and Kvasir-VQA-x1 [5]. Kvasir-VQA provides the component-level atomic QA pairs, while Kvasir-VQA-x1 combines corresponding same-image pairs into complex questions and naturalized complex answers. EndoCA keeps this link: each complex question is evaluated together with its associated atomic questions. EndoCA-Core covers common compact cases with one or two associated atomic questions and serves as the main comparison setting. EndoCA-Diagnostic follows the same protocol but adds a controlled three-atomic-question stratum to test how paired correctness changes as more endoscopic visual judgments are combined. Across both suites, EndoCA reports complex-answer accuracy, atomic-answer accuracy, joint accuracy, and complex-atomic inconsistency to separate final-answer correctness from same-image paired consistency.

Our study evaluates 11 VLMs from general open, medical open, endoscopy-adapted, and closed-source groups. Across both EndoCA suites, complex-answer rankings change once every associated atomic prediction must also be correct, and higher question complexity places stronger pressure on all-

atomic accuracy. These results support reporting complex-answer accuracy together with atomic-answer accuracy, joint accuracy, and complex-atomic inconsistency whenever associated atomic questions are available. Building on this paired evaluation signal, we introduce Atomic-Support Reconciliation (ASR), a training-free mechanism that uses model-generated atomic answers as contextual premises for answer revision and consistency-guided selective answering.

This paper makes three contributions:

- We introduce EndoCA, a paired complex-atomic answer consistency benchmark for endoscopic VQA, with EndoCA-Core for common compact question patterns and EndoCA-Diagnostic for controlled higher-complexity analysis.
- We evaluate 11 VLMs across general open, medical open, endoscopy-adapted, and closed-source groups, showing that complex-answer accuracy alone can hide substantial same-image complex-atomic answer inconsistency.
- We propose ASR, a training-free mechanism that uses model-generated atomic answers as contextual premises for answer revision and consistency-guided selective answering.

II. RELATED WORK

A. Medical and Endoscopic VQA

Medical VQA has progressed from modality-specific QA resources to broader clinical VLM benchmarks. VQA-RAD [6], SLAKE [7], and PathVQA [8] provide foundations for radiology, semantically labeled medical QA, and pathology, respectively. In gastrointestinal endoscopy, Kvasir-VQA [4] and Kvasir-VQA-x1 [5] provide concise endoscopic QA and complex reasoning-oriented QA. Recent efforts further broaden the evaluation context: EndoBench evaluates VLMs across multiple endoscopic scenarios, clinical tasks, and visual prompting granularities [3]; Gastric-X studies multimodal gastric cancer analysis across VQA, report generation, retrieval, classification, and localization [9]; and MEDVQA-GI explores challenge systems for gastrointestinal endoscopy VQA [10]. These studies expand the task scope and clinical context of medical and endoscopic VQA; their primary evaluation signals still center on final-answer accuracy and task performance, leaving model-internal consistency and stability less directly assessed.

B. Consistency Evaluation and Answer Reconciliation

Consistency-aware VQA and atomic evaluation study whether model answers remain reliable across related questions, structured reasoning steps, or smaller factual units. SQuINTing probes VQA models with lower-level perception sub-questions [11], and GQA provides structured real-world visual reasoning questions that support compositional evaluation [12]. In generation evaluation, FActScore decomposes long-form text into atomic facts [13], while FaithScore and POPE assess VLM hallucination and image-grounded faithfulness at finer granularity [14], [15]. Reliability-oriented work has also appeared in surgical and endoscopic settings, such as

SurgViVQA for temporally grounded surgical VideoQA robustness [16]. EndoCA brings this consistency-aware atomic checking perspective to endoscopic VQA by focusing on whether accepted complex answers remain supported by their associated atomic answers.

Answer correction and self-checking methods use related predictions or multiple generations to improve reliability after an initial response. SOrT uses sub-question relevance to improve VQA consistency [17]; self-consistency decoding samples multiple reasoning paths and selects an answer supported by agreement across paths [18]; and SelfCheckGPT detects hallucination by comparing consistency across sampled passages [19]. In endoscopic diagnosis, EndoCogniAgent uses closed-loop self-consistency validation for agentic reasoning [20]. ASR differs from prior self-checking methods by using model-generated atomic answers as explicit premises for complex-answer revision and as consistency evidence for selective answering.

III. ENDOCA BENCHMARK

A. Source Dataset

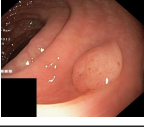
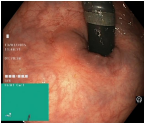
EndoCA uses the paired QA structure of Kvasir-VQA and Kvasir-VQA-x1. Kvasir-VQA contains concise QA pairs over gastrointestinal endoscopic images, covering common endoscopic findings, landmarks, instruments, and artifacts. Kvasir-VQA-x1 records complex questions and naturalized complex answers formed from corresponding same-image QA pairs. In EndoCA, the associated atomic questions are these same-image component questions linked to a complex question. This recorded association links each complex question to its atomic answer annotations, providing the paired units used for complex-atomic evaluation.

B. Benchmark Construction

The construction process starts from samples that contain an endoscopic image, a non-empty complex question, a complex answer annotation, and at least one associated atomic question-answer pair. Fixed parsing, answer normalization, consistency checks, and validity filtering standardize QA fields, canonicalize answer forms, require complete image/question/answer fields, well-formed atomic-question sets, and supported question types, and verify agreement between the complex annotation and its associated atomic annotations under the same scoring rules. The resulting paired units are illustrated in Table I.

EndoCA is organized into two complementary suites. Question complexity is the number of associated atomic questions in a complex answer. EndoCA-Core is the main evaluation suite, with 12,000 complex samples, 15,736 atomic QA items, and 27,736 total QA items per model; it focuses on complexity-1 and complexity-2 cases for stable cross-model comparison. EndoCA-Diagnostic contains 6,000 complex samples, 9,300 atomic QA items, and 15,300 total QA items per model; it keeps compact questions as the majority and adds 900 complexity-3 cases as a controlled higher-complexity stratum. The benchmark atomic questions cover high-confidence endoscopic question types over common findings, landmarks, instruments, and

TABLE I
REPRESENTATIVE ENDOCA COMPLEX-ATOMIC QA EXAMPLES.

Question-complexity	Image	Complex QA	Atomic QA
1		Q: How many polyps are present in the gastrointestinal tract shown? A: one polyp observed	Polyp Count q_1: How many polyps are in the image? Atomic Answer a_1: 1
2		Q: Are there any green or black box artifacts and is there visible text on the image? A: Evidence of green and black box artifacts along with visible text is present.	Box Artifact Presence q_1: Is there a green/black box artifact? Atomic Answer a_1: yes Text Presence q_2: Is there text? Atomic Answer a_2: yes
3		Q: Where in the image are the instruments located, is there a green/black box artifact, and what instruments are visible? A: A tube is visible in the center-right and upper-right regions, with a green/black box artifact present.	Instrument Location q_1: Where in the image is the instrument? Atomic Answer a_1: center-right; upper-right Box Artifact Presence q_2: Is there a green/black box artifact? Atomic Answer a_2: yes Instrument Presence q_3: Are there any instruments in the image? Check all that are present. Atomic Answer a_3: tube

artifacts. The suites use distinct sample identities: Core provides the main benchmark distribution, and Diagnostic provides a separate setting for question-complexity analysis.

C. Answer Scoring and Metrics

For each sample i , EndoCA stores an endoscopic image V_i , a complex question Q_i , a complex answer annotation A_i , and an associated atomic-question list U_i :

$$x_i = (V_i, Q_i, A_i, U_i), \quad (1)$$

$$U_i = ((q_{i,1}, a_{i,1}), \dots, (q_{i,m_i}, a_{i,m_i})).$$

where q_{ij} and a_{ij} are the j -th associated atomic question and its answer annotation. The quantity $m_i = |U_i|$ is the question complexity, defined as the number of atomic questions associated with Q_i .

EndoCA evaluates answers with a fixed question-type-aware scorer. The same concise-answer prompt template is used for complex and atomic probes, with the question field replaced by Q_i or q_{ij} . Model outputs may be short phrases or complete sentences, so the scorer normalizes common surface variation, including case, spacing, punctuation, yes/no forms, count words, and canonical labels. For separate atomic predictions, the expected question type determines the matching rule: binary questions are mapped to yes/no values, count questions to integers, categorical questions to canonical labels, and multi-label questions to normalized label sets. For complex-answer scoring, EndoCA uses the verified associated atomic annotations as component-level ground truth for the original complex question. A complex prediction is counted as correct only when the normalized answer matches every associated atomic answer annotation under the corresponding question-type rule. Empty outputs, execution errors, ambiguous answers, and non-committal responses are counted as incorrect. This fixed protocol enables reproducible comparison across open-source and closed-source VLMs and keeps the evaluation independent of an additional LLM judge.

For a model output, let $\hat{A}_i$ be the complex answer and $\hat{a}_{ij}$ be the answer to atomic question q_{ij} . Let $S_C(\hat{A}_i, x_i) \in \{0, 1\}$

TABLE II
EXAMPLE SCORING OF AN ENDOCA PAIRED UNIT.

Step	Scoring process and result
Paired unit	$x_i = (V_i, Q_i, A_i, U_i)$, where U_i contains two associated atomic QA annotations with $a_{i,1} = \text{yes}$ and $a_{i,2} = \text{yes}$.
Prompt template	Complex and atomic probes use the same template, replacing only $\langle Q \rangle$: <i>Answer the question based on the endoscopic image. Respond with a concise short answer only. Do not explain. Question: $\langle Q \rangle$</i>
Complex QA	$\langle Q \rangle = Q_i$: <i>Are there any green/black box artifacts and is there visible text in the image?</i> Raw output $\hat{A}_i$: <i>A green/black box artifact is present, and visible text is present.</i>
Complex score	The scorer extracts component labels: box artifact = yes and visible text = yes. Both match $(a_{i,1}, a_{i,2})$, so $C_i = 1$.
Atomic QA	$\langle Q \rangle = q_{i,1}$: <i>Is there a green/black box artifact?</i> Raw output $\hat{a}_{i,1} = \text{yes}$. $\langle Q \rangle = q_{i,2}$: <i>Is there visible text?</i> $\hat{a}_{i,2} = \text{no}$.
Atomic score	$\hat{a}_{i,1}$ matches $a_{i,1}$, so $\mathcal{A}_{i,1} = 1$. $\hat{a}_{i,2}$ does not match $a_{i,2}$, so $\mathcal{A}_{i,2} = 0$.
Indicators	After normalization, $C_i = 1$, $\mathcal{A}_{i,1} = 1$, $\mathcal{A}_{i,2} = 0$, and $\mathcal{A}_i^* = 0$.

denote complex-answer correctness, and $S_A(\hat{a}_{ij}, q_{ij}, a_{ij}) \in \{0, 1\}$ denote atomic-answer correctness. We define

$$C_i = S_C(\hat{A}_i, x_i), \quad \mathcal{A}_{ij} = S_A(\hat{a}_{ij}, q_{ij}, a_{ij}),$$

$$\mathcal{A}_i^* = \prod_{j=1}^{m_i} \mathcal{A}_{ij}. \quad (2)$$

Here, $C_i = 1$ when the complex prediction matches all associated atomic answer annotations for sample i , $\mathcal{A}_{ij} = 1$ when the model's answer to q_{ij} matches its associated annotation, and $\mathcal{A}_i^* = 1$ only when all atomic questions for the sample are answered correctly. In question-complexity analysis, all-atomic accuracy averages $\mathcal{A}_i^*$ within each complexity group.

EndoCA reports four metrics. Complex-answer accuracy measures the accuracy of model answers to complex questions. Atomic-answer accuracy averages correctness across the model's separate answers to the associated atomic questions. Joint accuracy requires both the complex prediction and every separately generated atomic prediction to match their associated answer annotations for the same sample. Complex-atomic inconsistency is computed among correct complex-answer

cases and measures the share with at least one wrong atomic prediction:

$$\text{Complex-answer Accuracy} = \frac{1}{N} \sum_{i=1}^N C_i, \quad (3)$$

$$\text{Atomic-answer Accuracy} = \frac{\sum_{i=1}^N \sum_{j=1}^{m_i} \mathcal{A}_{ij}}{\sum_{i=1}^N m_i}, \quad (4)$$

$$\text{Joint Accuracy} = \frac{1}{N} \sum_{i=1}^N C_i \mathcal{A}_i^*, \quad (5)$$

$$\text{Complex-Atomic Inconsistency} = \frac{\sum_{i=1}^N C_i (1 - \mathcal{A}_i^*)}{\sum_{i=1}^N C_i}. \quad (6)$$

If a model has no correct complex-answer cases, this conditional metric is undefined and should be reported as N/A; in our evaluated rows, every model has at least one correct complex answer, so the denominator is nonzero. Result tables abbreviate the three accuracy metrics as Complex Acc., Atomic Acc., and Joint Acc.; the inconsistency metric is written out.

IV. METHOD

A. Motivation

EndoCA exposes complex-atomic inconsistency in endoscopic VQA: a model may give an acceptable complex answer while failing one supporting same-image atomic question. This observation suggests a direct inference-time mitigation strategy. ASR uses only the image, the complex question, the direct model answer, and model-generated atomic answers; the associated atomic annotations are used only for EndoCA scoring. Atomic-Support Reconciliation (ASR) treats the model-generated atomic answers as explicit component-level evidence for answer revision and selective answering.

Workflow 1: Atomic-Support Reconciliation

Input: image V_i , complex question Q_i , associated atomic questions $\{q_{ij}\}$, VLM M .

1. Direct answer: $\hat{A}_i \leftarrow M(V_i, Q_i)$.

2. Atomic answers: for $j = 1, \dots, m_i$ do
 $\hat{a}_{ij} \leftarrow M(V_i, q_{ij})$.

end for

3. Atomic premises: $P_i \leftarrow \{(q_{ij}, \hat{a}_{ij})\}_{j=1}^{m_i}$.

4. Revise: $\tilde{A}_i \leftarrow M(V_i, Q_i, \hat{A}_i, P_i)$.

5. Select: $d_i \leftarrow M(V_i, Q_i, \tilde{A}_i, P_i)$,
checking consistency between $\tilde{A}_i$ and P_i ,
where $d_i \in \{\text{answer}, \text{abstain}\}$.

Return:

ASR-Revise: $\tilde{A}_i$.

ASR-Selective: $\tilde{A}_i$ if $d_i = \text{answer}$, otherwise abstain.

Fig. 2. ASR workflow for revision and selective answering.

B. ASR-Revise

ASR-Revise revises complex answers with atomic answers through a training-free answer reconciliation procedure. For each sample, the model first produces a direct complex answer $\hat{A}_i$ to the complex question Q_i . The same model then answers every associated atomic question q_{ij} , yielding $\hat{a}_{ij}$. ASR organizes the atomic question-answer pairs into an atomic

premise set P_i , which records the model’s own same-image component-level evidence:

$$P_i = \{(q_{ij}, \hat{a}_{ij})\}_{j=1}^{m_i}. \quad (7)$$

The revise step asks the model to produce a reconciled complex answer $\tilde{A}_i$ using the image, the complex question, the model’s direct complex answer, and the model-generated atomic answers as contextual premises. This preserves the original complex-answer task while exposing the model to a structured summary of its atomic responses. All ASR calls use a fixed prompt template per variant: ASR-Revise requests a concise revised answer, and ASR-Selective requests an explicit answer/abstain decision together with the revised answer and a coarse consistency label. Selective-answering coverage counts only samples with an explicit answer decision; revised text is scored by the same EndoCA protocol as direct complex answers. By conditioning on the model-generated atomic answers and applying a consistency-based answer decision, ASR can improve joint accuracy and reduce complex-atomic inconsistency even when full-coverage complex-answer accuracy changes only modestly.

C. ASR-Selective

ASR-Selective performs selective answering under complex-atomic consistency. After ASR-Revise produces $\tilde{A}_i$, the model is asked to judge whether the revised complex answer is consistent with the atomic premise set P_i . The output is a compact decision object containing an “answer”/“abstain” decision, the revised answer, and a coarse consistency label. If the decision is “answer”, ASR-Selective returns $\tilde{A}_i$; otherwise, it abstains for that sample.

This selective-answering step provides a controlled way to withhold low-consistency responses. We evaluate it with coverage, the percentage of samples for which an answer is returned, and selective accuracy, the complex-answer accuracy on the answered subset. In endoscopic VQA, this reliability–coverage view is useful because the model returns more answers when its complex answer and model-generated atomic answers agree, and fewer answers when the paired evidence is unstable.

V. EXPERIMENTS

A. Experimental Setup

EndoCA evaluates whether a model’s complex prediction matches the associated atomic answer annotations and remains consistent with the model’s separate answers to the associated atomic questions on the same image. We evaluate 11 VLMs from four model groups. The General open group includes InternVL2.5-8B [21], Qwen2.5-VL-7B [22], and Qwen3-VL-8B [23]. The Medical open group includes MedGemma-4B [24], Lingshu-7B [25], and LLaVA-Med [26]. The endoscopy-adapted group includes MedGemma-FT and Qwen2.5-VL-FT [5], which are LoRA-adapted on the endoscopic VQA domain [27]. We use them as in-domain adaptation references: they are expected to be strong on complex-answer accuracy, and the analysis asks whether that strength also carries over to paired atomic correctness. The Closed-source group includes

TABLE III
ENDOCA-CORE PAIRED EVALUATION.

Group	Model	Params	QA		Atomic	Joint	Complex-Atomic
			Count	Acc. (%)			Inconsistency (%)
General open	InternVL2.5-8B	8B	27,736	51.4	44.0	31.5	38.7
	Qwen2.5-VL-7B	7B	27,736	54.1	55.5	38.0	29.8
	Qwen3-VL-8B	8B	27,736	55.1	67.3	47.6	13.7
Medical open	MedGemma-4B	4B	27,736	35.4	29.1	17.1	51.8
	Lingshu-7B	7B	27,736	51.6	53.7	36.9	28.6
	LLaVA-Med	7B	27,736	28.7	17.0	4.9	83.0
Endoscopy -adapted	MedGemma-FT	4B	27,736	74.5	73.0	55.8	25.1
	Qwen2.5-VL-FT	7B	27,736	72.1	59.3	41.6	42.4
Closed source	GPT-5.5	–	27,736	60.8	63.0	47.4	22.1
	Grok-4.20	–	27,736	34.8	36.9	19.7	43.3
	Claude Opus 4.7	–	27,736	63.9	76.1	56.0	12.5

GPT-5.5 [28], Grok-4.20 [29], and Claude Opus 4.7 [30]. Each model answers every complex question and every associated atomic question on the same image, and all outputs are scored with the fixed question-type-aware scorer defined in Section III.

All local open-source models are evaluated with PyTorch on a server with 8 NVIDIA RTX 4090 GPUs. All prompts are zero-shot and contain one endoscopic image, one user question, and the fixed concise-answer instruction described in Section III, with no system prompt or in-context examples. Open-source models use deterministic decoding with `do_sample=False` and a maximum generation length of 80 tokens. Closed-source models are evaluated with `temperature=0` and `max_tokens=80`; their reasoning settings are set to `xhigh` for GPT-5.5 and Grok-4.20, and `max` for Claude Opus 4.7. Images are loaded in RGB format and processed by each model’s default visual processor, without manual cropping or global resizing; closed-source API calls use base64-encoded original images. All reported runs cover the expected QA Count, and empty or invalid outputs are scored as incorrect under the fixed EndoCA scoring protocol.

B. EndoCA Benchmark Results

On EndoCA-Core, complex-answer accuracy and paired atomic correctness diverge substantially (Table III). The in-domain adapted references achieve the highest complex-answer accuracy, with MedGemma-FT and Qwen2.5-VL-FT reaching 74.5% and 72.1%, respectively. Their joint accuracy differs by 14.2 percentage points, indicating that similar complex-answer performance can correspond to different levels of atomic-answer correctness. Across open and closed-source models, rankings based on complex-answer accuracy change when every separately generated atomic prediction must also be correct. This supports the motivating concern that an accepted endoscopic complex answer may still hide unsupported component judgments.

The diagnostic suite confirms the same reliability gap under a controlled complexity axis. MedGemma-FT and Qwen2.5-VL-FT obtain nearly identical complex-answer accuracy, 70.4% and 70.5%, yet their joint accuracy differs substantially, 50.7% versus 36.8% (Table IV). Their complex-atomic inconsistency also differs by 19.8 percentage points. This contrast is central

TABLE IV
ENDOCA-DIAGNOSTIC PAIRED EVALUATION.

Group	Model	Params	QA		Atomic	Joint	Complex-Atomic
			Count	Acc. (%)			Inconsistency (%)
General open	InternVL2.5-8B	8B	15,300	46.1	43.6	26.7	42.0
	Qwen2.5-VL-7B	7B	15,300	48.1	55.0	33.0	31.4
	Qwen3-VL-8B	8B	15,300	50.3	66.0	42.1	16.3
Medical open	MedGemma-4B	4B	15,300	32.1	28.4	14.8	54.0
	Lingshu-7B	7B	15,300	46.3	54.1	32.9	28.9
	LLaVA-Med	7B	15,300	22.6	15.3	3.8	83.1
Endoscopy -adapted	MedGemma-FT	4B	15,300	70.4	71.2	50.7	28.0
	Qwen2.5-VL-FT	7B	15,300	70.5	58.0	36.8	47.8
Closed source	GPT-5.5	–	15,300	55.2	61.0	41.7	24.4
	Grok-4.20	–	15,300	30.1	35.8	16.7	44.6
	Claude Opus 4.7	–	15,300	54.2	56.6	36.0	33.5

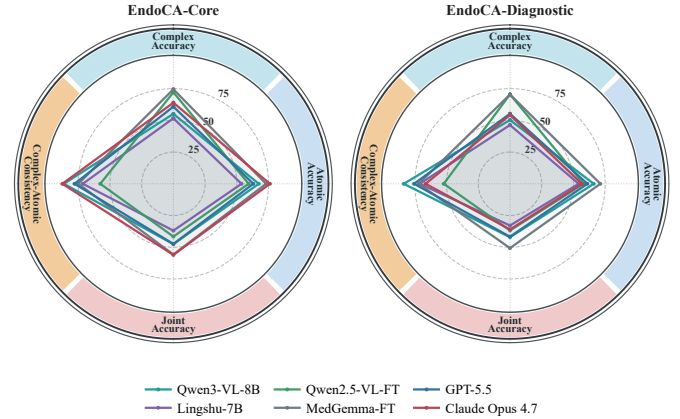


Fig. 3. Accuracy-consistency radar chart for six representative VLMs on EndoCA-Core and EndoCA-Diagnostic. Complex-atomic consistency denotes the complement of inconsistency.

to EndoCA: paired metrics expose whether a correct complex response remains aligned with the model’s same-image atomic predictions.

The six-model radar summary reinforces this separation between accepted complex answers and paired atomic correctness. Several models remain competitive on complex-answer accuracy but drop on joint accuracy or complex-atomic consistency. For visualization, Fig. 3 plots the complement of complex-atomic inconsistency so that all axes follow a higher-is-better direction.

Question complexity makes the same effect visible at the sample level. All-atomic accuracy uses $\mathcal{A}_i^*$ from Eq. (2); a single wrong atomic prediction sets the sample-level indicator to zero. At complexity 3, several models retain non-trivial complex-answer accuracy while all-atomic accuracy remains much lower (Fig. 4). As a complex question combines more atomic judgments, all-atomic accuracy becomes a paired correctness stress test.

C. Evaluation of ASR

We next evaluate whether ASR can reduce complex-atomic inconsistency by using model-generated atomic answers as contextual premises. We apply ASR to four stronger publicly available models selected from the general open, medical open,

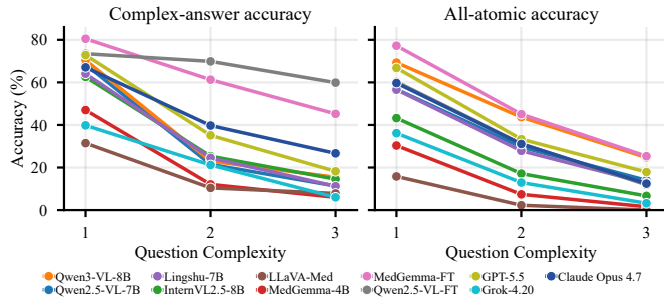


Fig. 4. Question-complexity analysis on EndoCA-Diagnostic. All-atomic accuracy requires every separate atomic prediction to match its associated answer annotation.

and endoscopy-adapted groups: Qwen3-VL-8B, Lingshu-7B, MedGemma-FT, and Qwen2.5-VL-FT. Relative to the direct EndoCA-Core results, ASR-Revise improves joint accuracy for all four models, with gains from +3.1 to +8.1 percentage points, and reduces complex-atomic inconsistency by 10.4 points on average (Table V). Its effect on full-coverage complex-answer accuracy is mixed, with the clearest benefit appearing in paired reconciliation.

TABLE V
ASR-REVISE AND ASR-SELECTIVE RESULTS ON ENDOCA-CORE.

Model	Direct Complex Acc. (%)	Revise Complex Acc. (%)	Joint Complex Acc. (%)	Complex-Atomic Inconsistency Δ	Coverage (%)	Selective Complex Acc. (%)
Qwen3-VL-8B	55.1	54.7	+5.3	-10.3	73.0	68.3
MedGemma-FT	74.5	76.0	+8.1	-9.2	95.7	76.8
Qwen2.5-VL-FT	72.1	69.6	+3.1	-6.6	76.1	74.9
Lingshu-7B	51.6	48.4	+5.2	-15.5	58.9	63.9
Average	63.3	62.2	+5.4	-10.4	75.9	71.0

ASR-Selective characterizes a reliability–coverage trade-off for the answered subset. On average, it returns answers for 75.9% of samples and reaches 71.0% selective complex-answer accuracy. Models with less stable paired behavior improve answered-subset accuracy by abstaining more often. Fig. 5 compares Direct, ASR-Revise, and ASR-Selective complex-answer accuracy together with the coverage reported in Table V.

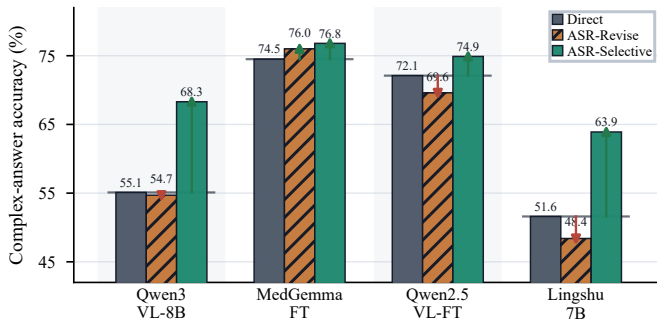


Fig. 5. Direct, ASR-Revise, and ASR-Selective complex-answer accuracy on EndoCA-Core. Direct and ASR-Revise accuracies use all samples; ASR-Selective accuracy is computed over answered samples only.

D. Ablation Study

We isolate the role of atomic premises using no-atomic self-check controls on the same four models and EndoCA-Core samples. Direct is the original complex answer. For Complex Acc., Joint Acc., and Complex-Atomic Inconsistency, w/o atomic is Self-Revise, which revises using the image, complex question, and direct answer without model-generated atomic premises. For Coverage and Selective Complex Acc., w/o atomic is Self-Selective, which makes the answer/abstain decision without atomic consistency evidence. ASR adds atomic premises to the corresponding revise and selective variants.

TABLE VI
ATOMIC-PREMISE ABLATION ON ENDOCA-CORE.

Method	Complex Acc. (%)	Joint Acc. (%)	Complex-Atomic Inconsistency (%)	Coverage (%)	Selective Complex Acc. (%)
Direct	63.3	45.5	27.4	100.0	63.3
w/o atomic	60.6	43.0	28.2	84.5	66.8
ASR	62.2	50.9	17.1	75.9	71.0

The ablation shows that generic self-checking alone does not recover the paired-correctness gains of ASR. Self-Revise without atomic premises reduces full-coverage complex-answer accuracy from 63.3% to 60.6% and does not improve joint accuracy or inconsistency, whereas ASR keeps complex-answer accuracy close to Direct at 62.2%, improves joint accuracy to 50.9%, and reduces complex-atomic inconsistency to 17.1% (Table VI). The selective columns show a related pattern: Self-Selective filters some unreliable answers, reaching 66.8% answered-subset accuracy at 84.5% coverage, while ASR-Selective achieves 71.0% at 75.9% coverage by using atomic premises as a stronger consistency signal. These results indicate that conditioning revision and selection on model-generated atomic answers is the key source of the paired-correctness gains.

VI. CONCLUSION

This paper presented EndoCA, a paired complex-atomic evaluation protocol for endoscopic VQA that scores complex predictions together with same-image atomic predictions. Across 11 VLMs, complex-answer accuracy gives an incomplete view of paired correctness: models with strong complex-answer performance can still show lower atomic-answer accuracy, lower joint accuracy, and non-negligible complex-atomic inconsistency. ASR further shows that model-generated atomic answers can serve as contextual premises for answer revision and consistency-guided selective answering, exposing a reliability–coverage trade-off. The fixed EndoCA scorer enables reproducible comparison across model groups, while future work should extend paired evaluation to free-form clinical questions through reliable automatic decomposition and careful comparison of decomposition strategies.

REFERENCES

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “VQA: Visual question answering,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

- [2] Z. Lin, D. Zhang, Q. Tao, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge, "Medical visual question answering: A survey," *Artificial Intelligence in Medicine*, vol. 143, p. 102611, 2023.
- [3] R. Liu *et al.*, "EndoBench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis," 2025, arXiv:2505.23601.
- [4] S. Gautam, A. Storås, C. Midoglu, S. A. Hicks, V. Thambawita, P. Halvorsen, and M. 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*, 2024.
- [5] S. Gautam, M. A. Riegler, and P. Halvorsen, "Kvasir-VQA-x1: A multimodal dataset for medical reasoning and robust MedVQA in gastrointestinal endoscopy," 2025, arXiv:2506.09958.
- [6] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific Data*, vol. 5, p. 180251, 2018.
- [7] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *2021 IEEE 18th International Symposium on Biomedical Imaging*, 2021.
- [8] X. He, Y. Zhang, L. Mou, E. P. Xing, and P. Xie, "PathVQA: 30000+ questions for medical visual question answering," 2020, arXiv:2003.10286.
- [9] S. Lu *et al.*, "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*, 2026.
- [10] P. K. Gaihare, A. More, R. Kusumba, M. Bhalerao, S. Sinha, C. Bhatt, G. Thakur, D. K. Yadav, and A. K. Jaiswal, "Multimodal AI for gastrointestinal diagnostics: Tackling VQA in MEDVQA-GI 2025," 2025, arXiv:2507.14544.
- [11] R. R. Selvaraju, P. Tendulkar, D. Parikh, E. Horvitz, M. T. Ribeiro, B. Nushi, and E. Kamar, "SQuINTing at VQA models: Introspecting VQA models with sub-questions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 003–10 011.
- [12] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [13] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "FActScore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 12 076–12 100.
- [14] L. Jing, R. Li, Y. Chen, and X. Du, "FaithScore: Fine-grained evaluations of hallucinations in large vision-language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 5042–5063.
- [15] Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 292–305.
- [16] M. O. Drago, L. Carlini, P. C. Balyemez, D. Pierantozzi, C. Lena, C. Hassan, D. Stoyanov, E. De Momi, S. Bano, and M. I. Hoque, "SurgViVQA: Temporally grounded video question answering for surgical scene understanding," *International Journal of Computer Assisted Radiology and Surgery*, 2026.
- [17] S. Dharur, P. Tendulkar, D. Batra, D. Parikh, and R. R. Selvaraju, "SOrTing VQA Models: Contrastive gradient learning for improved consistency," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 3103–3111.
- [18] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *International Conference on Learning Representations*, 2023.
- [19] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 9004–9017.
- [20] S. Tang *et al.*, "EndoCogniAgent: Closed-loop agentic reasoning with self-consistency validation for endoscopic diagnosis," 2025, arXiv:2508.07292.
- [21] Z. Chen *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," 2024, arXiv:2412.05271.
- [22] S. Bai *et al.*, "Qwen2.5-VL technical report," 2025, arXiv:2502.13923.
- [23] S. Bai *et al.*, "Qwen3-VL Technical Report," 2025, arXiv:2511.21631.
- [24] A. Sellergren *et al.*, "MedGemma technical report," 2025, arXiv:2507.05201.
- [25] LASA Team, W. Xu *et al.*, "Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning," 2025, arXiv:2506.07044.
- [26] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 28 541–28 564, datasets and Benchmarks Track.
- [27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://arxiv.org/abs/2106.09685>
- [28] OpenAI, "Introducing GPT-5.5," <https://openai.com/index/introducing-gpt-5-5/>, Apr. 2026.
- [29] xAI, "Grok 4.20," xAI Docs. <https://docs.x.ai/developers/models/grok-4-20-2026>.
- [30] Anthropic, "Introducing Claude Opus 4.7," Anthropic News. <https://www.anthropic.com/news/claude-opus-4-7>, Apr. 2026.