EVERYTHING A CORRECT ANSWER NEVER TOLD YOU: LEARNING FROM HUMAN-CORRECTED EXPLANATIONS IN COLONOSCOPY VIDEO QA

Anonymous authors
Paper under double-blind review

ABSTRACT

In colonoscopy video question answering, a correct answer can rest on the wrong visual evidence: across the general-purpose open-source models we evaluate, 10.60% of correct responses are justified by hallucinated observations. Motivated by this finding, we introduce ColonExplainVQA, a dataset for evaluating multimodal models and for training them to answer from corrected visual evidence. It contains 7,962 video-question pairs across nine tasks from three public datasets. We minimally correct hallucinated claims in conclusions generated by Qwen 3.7 Plus while preserving supported content and retaining the original responses. These corrections provide supervision for learning both the reference answer and the visual evidence that supports it. Using ColonExplainVQA, we develop Qwen-Colon-8B by fine-tuning Qwen3-VL-8B-Instruct through supervised fine-tuning followed by direct preference optimization in modules selected through Colon Function Layer (CFL) analysis. On our manually reviewed test set, Qwen-Colon-8B leads all fifteen baselines, including proprietary models far larger than itself, with 79.15% answer accuracy; among its correct answers, the hallucination rate falls to 0.95%. Answer accuracy alone is an incomplete target; the evidence behind an answer can be curated, measured, and improved. Upon acceptance, we will publicly release the ColonExplainVQA annotations, clip-reconstruction metadata, and all associated code, subject to source-data access and licensing terms.

1 INTRODUCTION

Multimodal large language models (MLLMs) are extending gastrointestinal visual question answering (VQA) from recognizing findings to explaining observations in endoscopic examinations. Kvasir-VQA and EndoBench support gastrointestinal image understanding, while SurgViVQA and Colon-Bench broaden the study of temporal reasoning and lesion-centered video analysis (Gautam et al., 2024; Liu et al., 2025; Drago et al., 2026; Hamdi et al., 2026). These developments raise a central question: how well are model answers supported by the visual evidence? Answer accuracy alone is insufficient to establish this connection. SurgCheck, for example, shows that surgical VQA models can exploit linguistic cues even when visual understanding is required (Shin et al., 2026).

Our evaluation reveals a further gap: even a correct answer can be justified by hallucinated visual evidence. Across the eight general-purpose open-source models we evaluate, 10.60% of correct responses still contradict the curated reference conclusion (Section 4.2). In one colonoscopy example, a model correctly identifies snare-related content but claims that the snare has engaged a polyp, although the clip shows only a wire beside the lesion. The treatment category is correct, yet the claimed action is unsupported. During curation of our corpus, 27.2% of examples require conclusion revisions while retaining the answer (Section 3.1.2). Such cases show why evaluating the answer alone misses errors in the observations used to justify it. They also provide a learning opportunity: correcting these observations supplies supervision about which visual evidence supports an answer. We therefore investigate whether learning from corrected visual evidence can reduce hallucinations while improving colonoscopy VQA accuracy.

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

Existing work provides complementary tools for this goal. Synthetic instruction tuning enables biomedical adaptation (Li et al., 2023a), while expert corrections and automated verification improve the visual support of model responses (Yu et al., 2024; Khanal et al., 2025; Yu et al., 2025). Preference learning further uses corrected responses, clinically informed preferences, or spatiotemporal perturbations to align image and video models (Yang et al., 2025; Zhu et al., 2025; Sarkar & Etemad, 2025). Building on these advances, we study two coupled design choices in colonoscopy VQA: which visual claims provide supervision and which modules receive preference updates. The first concerns learning answers together with their visual justification; the second concerns directing adaptation toward computations relevant to the domain.

We introduce ColonExplainVQA, a resource for evaluating visual-evidence hallucinations and training models with corrected evidence. It contains 7,962 video-question pairs across nine tasks from three public datasets. Two human annotators for endoscopic video analysis, assisted by GPT-5.5, review Qwen 3.7 Plus conclusions regardless of answer correctness and minimally revise hallucinated claims, with complex cases referred to one clinical reviewer with endoscopy expertise. Supported content and wording are preserved to limit stylistic differences between original and corrected conclusions, and both versions are retained. Figure 2 illustrates these corrections. The resulting resource pairs reference answers with reviewed visual evidence and includes 796 manually verified test questions.

Using this supervision, we develop Qwen-Colon-8B by adapting Qwen3-VL-8B-Instruct (Bai et al., 2025) in two stages. Supervised fine-tuning (SFT) first teaches the model to produce answers with supported visual conclusions. We then apply direct preference optimization (DPO), using the SFT model as the initialization and frozen reference (Rafailov et al., 2023). Our experiments show that distributing DPO updates across all language layers does not improve upon curated SFT. We therefore introduce Colon Function Layer (CFL) selection, adapting Vision Function Layer analysis (Shi et al., 2025) to identify sensitive modules and select their combinations through DPO screening. Preference updates are restricted to the selected modules, using curated completions and retained erroneous responses. On the 796-question test set, Qwen-Colon-8B achieves 79.15% answer accuracy, compared with 45.73% for the backbone, while reducing the human-verified visual-evidence hallucination rate, assessed against curated references, from 52.39% to 15.58%. Among correctly answered questions, this rate decreases from 13.46% to 0.95%.

Our contributions are threefold:

- We introduce ColonExplainVQA, comprising 7,962 video-question pairs across nine tasks from three public datasets. By pairing reference answers with original and minimally corrected visual conclusions, it supports evaluation beyond answer correctness and provides supervision for learning from reliable visual evidence.
- We develop Qwen-Colon-8B through visual-evidence SFT followed by CFL-guided DPO. CFL combines functional sensitivity analysis with DPO screening to select adaptation modules, connecting curated evidence supervision with targeted preference updates for colonoscopy understanding.
- We show that answer correctness can conceal visual hallucinations and that our approach improves answer accuracy while reducing hallucinations. Qwen-Colon-8B leads all 15 baselines on both answer accuracy and evidence consistency, reducing the hallucination rate among correct answers to 0.95%.

2 Related Work

2.1 General and Gastrointestinal VQA

VQA, VQA v2, and GQA establish visual QA, reduce language shortcuts, and support compositional reasoning, respectively (Antol et al., 2015; Goyal et al., 2017; Hudson & Manning, 2019). NExT-QA targets causal and temporal reasoning, while GroundVQA jointly learns answering and temporal localization (Xiao et al., 2021; Di & Xie, 2024).

Gastrointestinal benchmarks span image QA, compositional questions, and colonoscopy reasoning (Gautam et al., 2024; 2026; Liu et al., 2025; Ji et al., 2025). SurgViVQA and Colon-Bench add clip-based QA and reviewed lesion-window annotations (Drago et al., 2026; Hamdi et al., 2026). Surgical

VideoQA explores text-guided memory, question-conditioned sampling, and structured perception (Li et al., 2026; Guo et al., 2026; Shen et al., 2026). Medico 2025 evaluates explanations; multi-task GI methods combine QA with synthetic explanations, appearance descriptions, and localization (Gautam et al., 2025; Safwan et al., 2025; 2026). ColonExplainVQA retains original and minimally corrected answer–conclusion pairs. Table 3 in Appendix C.3 compares dataset scale, tasks, visual inputs, and paired corrections.

2.2 Hallucination Evaluation and Preference Learning

POPE, VidHalluc, and ARGUS evaluate object hallucinations, video inconsistencies, and unsupported or omitted visual details (Li et al., 2023b; 2025; Rawal et al., 2025). e-ViL and MIMIC-NLE assess explanations beyond answer accuracy (Kayser et al., 2021; 2022). Gut-VLM provides corrected gastrointestinal reports; subsequent benchmarks evaluate GI hallucination detectors, while VIHD probes visual dependence (Khanal et al., 2025; Lawal et al., 2026; Chen et al., 2026).

Mitigation includes VCD decoding, RLHF-V weighting of human-corrected spans, and RLAIF-V AI feedback (Leng et al., 2024; Yu et al., 2024; 2025). Preference methods address policy matching, visual conditioning, and phrase-level errors, with medical and video adaptations (Yang et al., 2025; Wang et al., 2024; Sarkar et al., 2025; Zhu et al., 2025; Sarkar & Etemad, 2025). MCR uses image corruption and synthetic reasoning–answer mismatches (Jiang et al., 2026). Selective adaptation includes layer sampling in LISA, rank allocation in AdaLoRA, and decoder visual-function localization in VFL (Pan et al., 2024; Zhang et al., 2023; Shi et al., 2025). Our SFT uses minimally corrected visual conclusions, including correct-answer cases; DPO contrasts curated responses with retained model errors. CFL measures task-balanced sensitivity across language, visual-encoder, and merger modules and screens update configurations through development-set DPO.

3 Method

We first describe the composition, curation, and partitions of ColonExplainVQA (Section 3.1). We then present the two-stage model adaptation pipeline: evidence-curated SFT followed by selective DPO, with Colon Function Layer (CFL) selection connecting the two stages (Section 3.2).

3.1 ColonExplainVQA Benchmark

ColonExplainVQA is a colonoscopy video question-answering resource for studying answer correctness and the grounding of accompanying explanations. It contains 7,962 examples across nine tasks, combining clinical findings, treatment-related observations, and temporal understanding. Each example couples a video clip, or an ordered pair of clips, with a multiple-choice question and a concise explanatory conclusion. The construction pipeline preserves original model responses alongside curated answers and explanations, making the changes introduced during curation traceable. Figure 1 summarizes the construction pipeline and representative examples of the nine tasks.

3.1.1 Dataset Composition

ColonExplainVQA contains 7,962 VQA pairs derived from three public datasets. REAL-Colon contributes 3,763 examples covering polyp detection, identity matching, apparent motion, and instrument motion (Biffi et al., 2024). HyperKvasir contributes 3,878 examples covering bowel preparation, endoscopic findings, and treatment-related appearance and content (Borgli et al., 2020). SUN-SEG contributes 321 examples for apparent-motion analysis and pathology classification using source labels (Ji et al., 2022). The resource contains 6,816 unique video clips, with a mean duration of 6.82 s, a median of 5.00 s, and a range of 1.36–14.80 s. Task-specific inputs preserve temporal sequences, with paired clips used for identity matching.

The nine tasks form four groups. Lesion recognition includes polyp detection, main endoscopic finding identification, and pathology classification using source-provided labels. Quality assessment assigns the observed segment to a binary bowel-preparation group. Procedure understanding covers treatment appearance, treatment content, and instrument motion, while temporal reasoning addresses apparent polyp motion and polyp identity across clips. Apparent polyp motion denotes



Figure 1: ColonExplainVQA construction pipeline and representative examples of its nine tasks. VQA pairs derived from source annotations undergo model response generation and human review, with unsupported visual claims minimally corrected in Plus conclusions.

displacement in the image plane, whereas instrument advancement and retraction are defined relative to the working-channel entry.

3.1.2 VQA Construction and Evidence Curation

We construct multiple-choice video-question pairs from source annotations, which provide task-specific reference answers. For HyperKvasir, source-video labels are reviewed for their applicability to each selected clip. Qwen 3.7 Flash and Qwen 3.7 Plus each receive the video, question, and answer options and generate an answer and a conclusion describing the claimed visual evidence and rationale for the selected option. Two human annotators for endoscopic video analysis independently review all 7,962 VQA pairs, checking both models' answers against the reference labels and their conclusions against the videos, regardless of answer correctness. Their initial judgments on hallucination presence in the original Plus conclusions achieve 94.25% agreement and Cohen's $\kappa = 0.884$, measured before discussion or clinical adjudication. With GPT-5.5 assistance, they propose minimal revisions to hallucinated or unsupported claims, restricting textual correction to Plus conclusions and preserving supported content, wording, and structure. Clinical adjudication covers 458 cases with conflicting hallucination judgments and another 76 cases in which both annotators identify hallucinations but cannot determine a sufficiently supported correction. For these 534 difficult cases (6.71% of the dataset), one clinical reviewer with endoscopy expertise works with the two annotators to review the corresponding videos and jointly determine the final judgments and corrections. Among the 458 disputed cases, 304 are confirmed to contain hallucinations and 154 are judged not to contain hallucinations. Original Plus conclusions are retained alongside the final curated versions for subsequent training and analysis. The minimal-edit policy reduces stylistic differences between these versions, helping training focus on correcting visual evidence. Following curation, all 796 test pairs undergo a second round of manual verification of reference-answer correctness and the visual support for their conclusions.

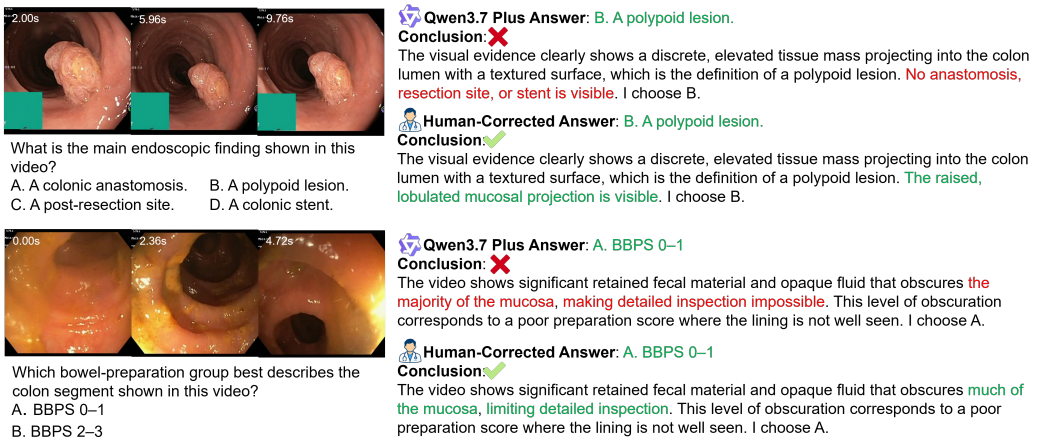


Figure 2: Examples of visual-evidence hallucination correction in ColonExplainVQA. Correct answers are retained while unsupported exclusions in the main-finding example (top) and overstated limits on mucosal inspection in the bowel-preparation example (bottom) are corrected. Red highlights unsupported or overstated claims; green marks corrected evidence and correct answer choices.

The final review identifies hallucinations in 4,405 original Plus conclusions (55.3%). Comparison with the curated responses after formatting normalization yields 2,242 cases (28.2%) in which the reference answer differs from the Plus prediction, 2,163 (27.2%) in which the answer is unchanged but the conclusion is revised, and 3,557 (44.7%) with neither changed. The conclusion-only cases motivate reviewing visual evidence even when the selected answer is correct. Figure 7 in Appendix C summarizes the dataset sources, task frequencies, and correction profile. Figure 2 illustrates a conclusion-only hallucination correction: the correct polypoid-lesion answer is retained, while unsupported exclusions of other findings are replaced with a description of the visible raised, lobulated tissue. Further corrections appear in Appendix C (Figures 8, 9, and 10).

The curated corpus is organized into training, validation, and test partitions, with 796 VQA pairs reserved for testing. We connect records sharing a source-video identifier, a parent SUN-SEG case, or an individual clip, including either member of a clip pair. Entire connected groups are assigned to one partition while approximating task proportions under fixed split sizes with seed 42. All 21 identity examples originate from a single independent source group and are assigned to training, so validation and test evaluation cover the remaining eight tasks.

3.2 Model Adaptation

Figure 3 summarizes the adaptation pipeline for Qwen-Colon-8B, which is fine-tuned from Qwen3-VL-8B-Instruct (Bai et al., 2025). The three illustrated steps are evidence-curated SFT, CFL selection, and selective DPO. SFT learns the correct reference answers paired with reviewed Plus conclusions, minimally human-corrected where needed. Task-balanced functional interventions and development-set screening then select the DPO update sites. Finally, DPO favors these curated answer–conclusion responses over retained incorrect-answer responses, using the SFT model as both the initialization and frozen reference. SFT and DPO constitute the two training stages.

3.2.1 Evidence-Curated Supervised Fine-Tuning

For an input $x = (v, q, o)$ consisting of video input, a question, and answer options, supervision comprises the reference answer a^* and a curated conclusion c^*. Conclusions follow the human-reviewed, model-assisted protocol in Section 3.1.2: accepted Plus conclusions are retained, while unsupported visual claims are minimally corrected using the reference annotations and video evidence. Revisions preserve compatible factual content and wording. We serialize the curated conclusion c^* and reference answer a^* into a target assistant response $y^* = (y_1^*, \ldots, y_T^*)$, where T denotes the number of response tokens. We introduce LoRA adapters (Hu et al., 2022) into the attention and MLP projections of all language-decoder layers, keeping the pretrained weights, visual encoder, and

visual mergers frozen. With trainable adapters ϕ and frozen parameters θ_0, we minimize

$$\mathcal{L}_{\mathrm{SFT}}(\phi) = -\mathbb{E}_{(x,y^*)\sim\mathcal{D}_{\mathrm{SFT}}} \sum_{t=1}^{|y^*|} \log \pi_{\theta_0,\phi}(y_t^* \mid x, y_{<t}^*). \tag{1}$$

Only assistant-response tokens contribute to the loss. We select the SFT checkpoint by validation strict QA accuracy, which requires a correct answer, valid response format, and agreement between the answer field and the conclusion's terminal option. Merging its adapters into the backbone yields π_{SFT}.

3.2.2 Colon Function Layer Selection

Our Colon Function Layer (CFL) selection aims to direct preference updates toward computations that support colonoscopy answering from visual evidence. We hypothesize that modules whose disruption reduces correct-answer probability provide useful candidates for evidence-focused adaptation. Building on VFL (Shi et al., 2025), we therefore use functional sensitivity to prioritize candidate modules and short DPO runs to determine which configurations improve performance. The procedure below expands the selection step in Figure 3.

CFL selection

1. **Probe modules.** Starting from the SFT model, temporarily remove visual attention entries or zero outputs in language attention, language MLPs, visual-encoder attention, and visual mergers, with parameters fixed.
2. **Score sensitivity.** Use Eq. 2 to measure reference-answer log-probability loss, weighting tasks equally.
3. **Rank and combine.** Form nested top-k layer sets within each family; combine these sets and merger subsets into language–vision–merger, language–vision, and language–merger configurations.
4. **Screen with DPO.** Compare short runs initialized from the same SFT model (Section 3.2.3). Select using task-macro strict QA accuracy on a development pool held out from DPO updates by source group.

For intervention m, let $\widetilde{p}_m(a \mid x)$ denote the answer probability obtained by applying softmax to logits over valid option letters, and let $\widetilde{p}_0$ denote the unmodified SFT model. For task set $\mathcal{T}$ and probe examples $\mathcal{P}_\tau$, we define

$$s(m) = \frac{1}{|\mathcal{T}|} \sum_{\tau\in\mathcal{T}} \frac{1}{|\mathcal{P}_\tau|} \sum_{(x,a^*)\in\mathcal{P}_\tau} \left[\log \widetilde{p}_0(a^* \mid x) - \log \widetilde{p}_m(a^* \mid x)\right]. \tag{2}$$

A larger positive score indicates stronger dependence of reference-answer prediction on the intervened computation. Sensitivity therefore prioritizes candidates, while DPO screening measures their adaptation utility. The development pool is drawn from the SFT training split. The selected configuration updates attention modules in three visual-encoder layers and MLP modules in 18 language-decoder layers. Appendix B provides the intervention details and screening results. Figure 5 presents the sensitivity profiles, with experimental analysis in Section 4.4.

3.2.3 Selective Preference Optimization

For each eligible question, y^+ combines the correct reference answer with its curated conclusion. The rejected response y^- is the original Plus response if its answer is incorrect or its conclusion contains verified visual-evidence hallucinations. For conclusion-only Plus pairs, the correct answer is preserved, while minimal corrections to hallucinated evidence provide the preference signal. The two responses share the same video, question, and options, forming the preferred and rejected completions illustrated in Figure 3 (step 3). For a retained module set $\mathcal{S}$, we initialize the policy from π_{SFT} and introduce fresh LoRA adapters $\psi_{\mathcal{S}}$. The merged SFT parameters remain frozen and define the reference $\pi_{\mathrm{ref}} = \pi_{\mathrm{SFT}}$. Writing

$$\Delta_\psi(x,y) = \log \pi_{\mathrm{SFT},\psi_{\mathcal{S}}}(y \mid x) - \log \pi_{\mathrm{ref}}(y \mid x),$$

we optimize (Rafailov et al., 2023)

$$\mathcal{L}_{\mathrm{DPO}}(\psi_{\mathcal{S}}) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}_{\mathrm{pref}}} \log \sigma\big(\beta\left[\Delta_\psi(x,y^+) - \Delta_\psi(x,y^-)\right]\big), \tag{3}$$

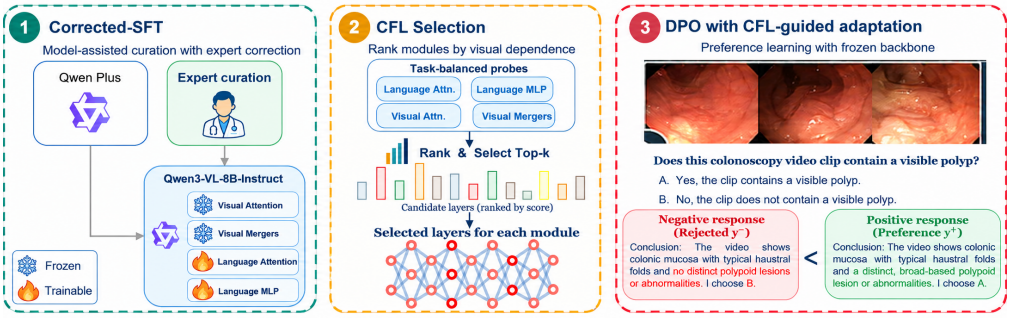


Figure 3: Training pipeline of Qwen-Colon-8B. (1) Reviewed and minimally corrected Plus conclusions supervise language-decoder SFT. (2) CFL probes visual dependence across four module families and screens candidate update sites. (3) DPO favors curated responses over incorrect responses, updating only LoRA adapters in the selected modules with the SFT backbone frozen.

Table 1: Results on ColonExplainVQA (%). Best and second-best results across all models are bold and underlined, respectively.

MLLMs	Answer Acc.↑	Task-macro Acc.↑	RCR↓	$CR_{correct}$ ↓	Joint Acc.↑
Open-Source MLLMs					
Qwen3-VL-8B-Instruct	45.73	32.30	52.39	13.46	39.57
Qwen3.5-9B	46.61	41.02	44.22	10.78	41.58
Qwen3.8-27B	64.82	49.27	31.28	5.81	61.06
InternVL3.5-8B	61.43	41.74	27.64	5.32	58.17
LLaVA-Video-7B	53.39	37.39	28.39	10.59	47.74
VideoLLaMA3-7B	51.01	39.53	52.14	19.21	41.21
MiniCPM-V 4.5	48.37	30.92	37.44	8.83	44.10
Molmo2-8B	42.34	26.83	49.50	13.95	36.43
Proprietary MLLMs					
Gemini 3.1 Pro	<u>68.09</u>	<u>52.56</u>	27.89	5.90	<u>64.07</u>
Qwen3.8-Max	67.59	52.06	31.03	7.06	62.81
Doubao-Seed-2.0-Pro	65.20	49.67	<u>26.01</u>	<u>5.20</u>	61.81
Qwen-VL-Max	60.55	45.02	33.79	8.92	55.15
Open-Source Medical-Domain MLLMs					
Hulu-Med-32B	48.37	39.26	42.84	9.35	43.84
Hulu-Med-30A3	47.74	38.00	38.19	8.68	43.59
Lingshu-32B	44.72	31.26	55.03	13.48	38.69
Qwen-Colon-8B (Ours)	**79.15**	**68.10**	**15.58**	**0.95**	**78.39**

where σ is the sigmoid function and $\beta = 0.1$. Sequence log-probabilities sum over the serialized assistant response, including the answer, conclusion, and response terminator, without length normalization. Prompt and padding tokens are masked. Only adapters in $\mathcal{S}$ are updated. Each retained configuration is trained independently, and its checkpoint is selected by task-macro strict QA accuracy on a separate validation set. Test data are excluded from module and checkpoint selection.

4 Experiments

4.1 Experimental Setup and Evaluation Metrics

We compare Qwen-Colon-8B against 15 baseline MLLMs, comprising eight general-purpose open-source, three open-source medical-domain, and four proprietary models, on 796 test questions from eight ColonExplainVQA tasks. All models follow the same generation protocol: given the video, question, and answer options, each first generates a conclusion describing the visual evidence and then outputs its answer. Model details and the generation protocol appear in Appendix A.1.

We report answer accuracy and task-macro accuracy, which weights the eight tasks equally. GPT-5.5 (medium) compares predicted and curated conclusions as text, followed by human verification of the judgments and supporting evidence. The reference contradiction rate (RCR) measures the fraction

Table 2: Ablation results on ColonExplainVQA (%). All DPO variants start from SFT–Curated; Qwen-Colon-8B uses CFL-DPO. Best and second-best results are bold and underlined, respectively.

Configuration	Answer Acc.↑	Task-macro Acc.↑	RCR↓	$CR_{correct}$ ↓	Joint Acc.↑
Qwen3-VL-8B-Instruct	45.73	32.30	52.39	13.46	39.57
SFT–Plus	68.47	57.04	31.03	10.09	61.56
SFT–Curated	72.61	60.34	23.49	3.98	69.72
Curated → Vanilla DPO	72.11	60.08	24.12	4.53	68.84
Curated → Random-layer DPO	<u>75.50</u>	<u>64.14</u>	22.86	3.66	72.74
Curated → Last-layer DPO	<u>75.50</u>	<u>64.14</u>	<u>22.61</u>	<u>3.49</u>	<u>72.86</u>
Qwen-Colon-8B (Ours)	**79.15**	**68.10**	**15.58**	**0.95**	**78.39**

containing at least one explicit contradiction; $CR_{correct}$ conditions this rate on a correct answer. Joint accuracy requires a correct answer and no detected reference contradiction. The final metrics use human-verified labels; absence of contradiction does not establish complete visual grounding. Formal definitions and judge procedures are given in Appendix A.2.

4.2 Benchmark Results

Qwen-Colon-8B leads all five metrics in Table 1. Medical-domain baselines do not consistently outperform general-purpose models: the most accurate medical baseline, Hulu-Med-32B (48.37%), falls below InternVL3.5-8B (61.43%). Moreover, across the eight general-purpose open-source models, 349 of 3,293 correct responses still contain a reference contradiction (10.60%), reinforcing the need to evaluate explanations separately from answers.

Qwen-Colon-8B outperforms SFT–Curated and Gemini 3.1 Pro, achieving 79.15% answer accuracy versus 72.61% and 68.09%, respectively. It also reduces RCR to 15.58% from 23.49% and 27.89%, while $CR_{correct}$ falls to 0.95% from 3.98% and 5.90%. These gains support improved answering and reduced visual-evidence hallucinations beyond curated SFT and the strongest accuracy baseline.

Qualitative hallucination analysis. Figure 4 illustrates two failure modes that complement the aggregate comparisons. In the motion example, Qwen Plus selects the correct leftward direction but describes the polyp as largely stationary, contradicting both its answer and the reviewed temporal evidence. In the pathology example, MiniCPM-V 4.5 predicts high-grade adenoma using claims of a large, broad-based lesion with a central depression, whereas the reviewed description reports no distinct visible polyp. These cases separate answer correctness from visual support; Figure 11 in Appendix C.4 provides further examples.

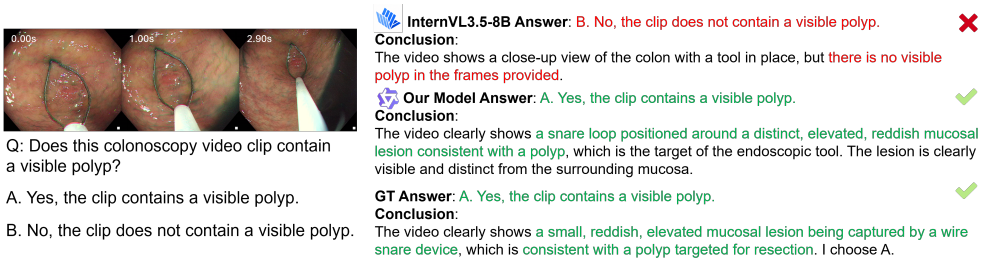


Figure 4: Illustrative answer–explanation failures. Top: Qwen3.7 Plus gives the correct motion answer with a contradictory explanation. Bottom: MiniCPM-V 4.5 supports a pathology prediction with lesion features absent from the reviewed description. Red highlights erroneous claims; green marks correct answers or reference content. Pathology targets are source labels, not visible histology.

4.3 Ablation Study

Table 2 compares the unadapted backbone, two SFT supervision sources, and four DPO variants initialized from SFT–Curated, using the same 796-question test set and evaluation protocol.

Effect of curated supervision. SFT–Plus learns the original Qwen 3.7 Plus answers and conclusions, including incorrect choices. Replacing these responses with reference answers and curated conclusions improves all five metrics: answer accuracy rises from 68.47% to 72.61%, while $\mathrm{CR}_{\mathrm{correct}}$ falls from 10.09% to 3.98%. This supports curating visual evidence with answer labels.

Effect of CFL-guided preference optimization. Vanilla DPO slightly worsens all five metrics relative to SFT–Curated, whereas random-layer and last-layer DPO improve them. CFL-DPO performs best across all five metrics. Compared with last-layer DPO, it raises joint accuracy from 72.86% to 78.39% and reduces $\mathrm{CR}_{\mathrm{correct}}$ from 3.49% to 0.95%, supporting task-relevant layer selection for improving answer correctness and reducing visual-evidence hallucinations.

4.4 CFL Sensitivity and Layer Selection

CFL sensitivity analysis. Figure 5 shows the task-balanced intervention scores of the SFT model (Eq. 2). Visual attention exhibits pronounced early- and late-layer peaks, while language-MLP sensitivity is concentrated in a subset of layers and merger responses vary across fusion modules. Under the respective interventions, language-attention scores are approximately two orders of magnitude smaller than those of the other module families, revealing substantially lower measured sensitivity in this probe. These heterogeneous profiles motivate ranking candidate update sites within each family.

Layer selection. Following step 2 of Figure 3, we screen sensitivity-ranked layer sets through short DPO runs on a 600-question development pool held out from DPO updates (Section 3.2.2). Appendix B details this pool and the archived single-family and joint screening results, with the configuration layout shown in Figure 6. Screening accuracy does not improve monotonically with the number of adapted layers. Our final model updates three visual-attention layers and 18 language-MLP layers without merger updates; these layers are ringed in Figure 5.

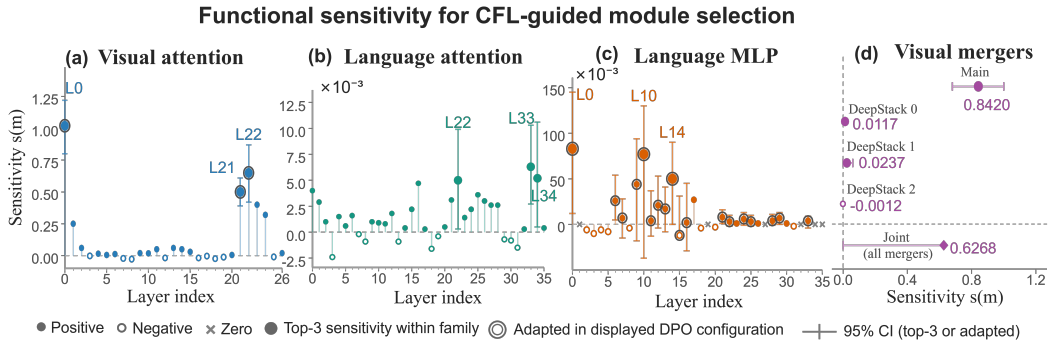


Figure 5: CFL sensitivity analysis before DPO layer selection. Panels show visual attention, language attention, language MLPs, and visual mergers. Rings mark the three visual-attention and 18 language-MLP layers selected for DPO; error bars show 95% confidence intervals.

5 Conclusion

We introduced ColonExplainVQA, with 7,962 video-question pairs across nine tasks and three public datasets, retaining original and minimally corrected conclusions, and Qwen-Colon-8B, trained through evidence-curated SFT and CFL-guided preference optimization. All 796 test questions are manually verified. Our results highlight two lessons: correct answers can conceal hallucinated evidence, and both supervision quality and update location matter for correcting it. At the same data scale, curated supervision improves accuracy and more than halves hallucinations among correct answers, while CFL-selected updates outperform preference optimization across all language layers. Qwen-Colon-8B leads all 15 baselines, achieving 79.15% answer accuracy and a visual-evidence hallucination rate of 0.95% among correct answers. These findings support evaluating and training models on the evidence behind their answers. Future work will extend this approach to other endoscopic procedures and modalities.

AI USE STATEMENT

Large language models are a component of our method and are described in Sections 3.1.2 and 4.1. Qwen 3.7 Flash and Qwen 3.7 Plus generated the candidate answers and visual-evidence conclusions that form the raw material of ColonExplainVQA; GPT-5.5 assisted annotators during correction and produced the initial reference-consistency judgments. Every conclusion and every judgment used in our reported results was verified by human reviewers, and the reference answers come from source-dataset annotations rather than model output. Separately, the authors used AI assistance for language editing of the manuscript. No experimental design, analysis, or interpretation was delegated to an AI system, and the authors take full responsibility for the content and scientific claims of this paper.

ETHICS STATEMENT

ColonExplainVQA builds on REAL-Colon and HyperKvasir, released under CC BY 4.0 (Biffi et al., 2024; Borgli et al., 2020), and SUN-SEG, whose source SUN videos are governed by non-commercial research and educational-use terms. The SUN terms prohibit redistribution of original or processed data, including derived clips.[1] Our release will provide VQA annotations, source identifiers, clip-reconstruction metadata, and code; users must obtain SUN media through the official access process. We will preserve source attribution, license notices, and applicable dataset-specific use conditions. Human review of reference conclusions helps reduce unsupported visual claims. Our benchmark and models are intended for research; clinical applications require independent validation, assessment of source-data biases, and clinician oversight.

REPRODUCIBILITY STATEMENT

Section 3.1 describes the data sources, task definitions, curation workflow, and data partitions. Section 3.2 specifies the SFT and DPO objectives and CFL selection procedure. Appendix A details the evaluated models, generation protocol, metric definitions, and human-verified judging process, while Appendix B documents the 600-question development pool and layer-selection screening. Additional dataset statistics and correction examples are provided in Appendix C. Upon acceptance, we will publicly release the ColonExplainVQA annotations, clip-reconstruction metadata, and all associated code, subject to source-data access and licensing terms.

We fine-tune Qwen3-VL-8B-Instruct using ms-swift 4.4.2 on two NVIDIA A100 40GB GPUs. Videos are sampled at 6 fps with 12–76 frames and a spatial budget of 128 visual tokens per frame. Both training stages use LoRA with rank 16 and alpha 32, an effective batch size of 16, BF16 precision, a maximum sequence length of 6,144 tokens, and fused AdamW with weight decay 0.1 and gradient clipping at 1.0. During SFT, adapters cover all 36 language-attention and MLP layers, while the visual encoder and mergers remain frozen. We use a learning rate of 5×10^{-5} with cosine decay and 5% warmup. The curated SFT checkpoint at step 650 is selected by constrained answer-choice accuracy on 200 validation examples. After merging its adapters, we optimize a fresh selective adapter on 2,383 preference pairs, holding out 600 examples from the SFT training partition for DPO development. DPO updates three visual-attention blocks and 18 language-MLP layers, totaling 14.49M trainable parameters, with $\beta = 0.1$, a learning rate of 10^{-5}, five warmup steps, and 300 optimization steps. Training uses answer-first JSON responses, whereas evaluation requests the conclusion before the answer. All runs use seed 42. The completed DPO run takes 5.39 hours (10.77 GPU-hours), including periodic preference validation.

REFERENCES

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Antol_VQA_Visual_Question_ICCV_2015_paper.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei

[1] https://github.com/GewelsJI/VPS#9-license

Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Carlo Biffi, Giulio Antonelli, Sebastian Bernhofer, Cesare Hassan, Daizen Hirata, Mineo Iwatate, Andreas Maieron, Pietro Salvagnini, and Andrea Cherubini. REAL-Colon: A dataset for developing real-world AI applications in colonoscopy. *Scientific Data*, 11, 2024. doi: 10.1038/s41597-024-03359-0.

Hanna Borgli, Vajira Thambawita, Pia H. Smedsrud, Steven Hicks, Debesh Jha, Sigrun L. Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, Dag Johansen, Carsten Griwodz, Håkon K. Stensland, Enrique Garcia-Ceja, Peter T. Schmidt, Hugo L. Hammer, Michael A. Riegler, Pål Halvorsen, and Thomas de Lange. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283, 2020. doi: 10.1038/s41597-020-00622-y. URL https://www.nature.com/articles/s41597-020-00622-y.

Jiayi Chen, Benteng Ma, Zehui Liao, Winston Chong, Yasmeen George, and Jianfei Cai. VIHD: Visual intervention-based hallucination detection for medical visual question answering, 2026. URL https://arxiv.org/abs/2605.20772v2.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28652–28668, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Clark_Molmo2_Open_Weights_and_Data_for_Vision-Language_Models_with_Video_CVPR_2026_paper.html.

Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12934–12943, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Di_Grounded_Question-Answering_in_Long_Egocentric_Videos_CVPR_2024_paper.html.

Mauro Orazio Drago, Luca Carlini, Pelinsu Celebi Balyemez, Dennis Pierantozzi, Chiara Lena, Cesare Hassan, Danail Stoyanov, Elena De Momi, Sophia Bano, and Mobarak I. Hoque. SurgViVQA: Temporally grounded video question answering for surgical scene understanding. *International Journal of Computer Assisted Radiology and Surgery*, 21(6):1369–1377, 2026. doi: 10.1007/s11548-026-03695-z.

Sushant Gautam, Andrea Storås, Cise Midoglu, Steven A. Hicks, Vajira Thambawita, Pål Halvorsen, and Michael A. Riegler. Kvasir-VQA: A text-image pair GI tract dataset. In *Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications*, pp. 3–12. Association for Computing Machinery, 2024. doi: 10.1145/3689096.3689458.

Sushant Gautam, Vajira Thambawita, Michael Riegler, Pål Halvorsen, and Steven Hicks. Medico 2025: Visual question answering for gastrointestinal imaging. In *Working Notes of the MediaEval 2025 Workshop*, 2025. URL https://2025.multimediaeval.com/paper2.pdf.

Sushant Gautam, Michael Riegler, and Pål Halvorsen. Kvasir-VQA-x1: A multimodal dataset for medical reasoning and robust MedVQA in gastrointestinal endoscopy. In *Data Engineering in Medical Imaging – DEMI 2025*, volume 16191 of *Lecture Notes in Computer Science*, pp. 53–63. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-08009-7_6. URL https://

link.springer.com/chapter/10.1007/978-3-032-08009-7_6. First published online October 12, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Goyal_Making_the_v_CVPR_2017_paper.html.

Diandian Guo, Xikai Yang, Ruiyang Li, Jialun Pei, and Pheng-Ann Heng. SurgLQA: Scalable long-horizon surgical video question answering, 2026. URL https://arxiv.org/abs/2605.17915. arXiv version 1; author record reports MICCAI 2026 Early Accept.

Abdullah Hamdi, Changchun Yang, and Xin Gao. Colon-bench: An agentic workflow for scalable dense lesion annotation in full-procedure colonoscopy videos, 2026. URL https://arxiv.org/abs/2603.25645v3. Version 3, 28 June 2026. arXiv record reports MICCAI 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Drew A. Hudson and Christopher 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 (CVPR)*, pp. 6700–6709, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Machine Intelligence Research*, 19(6):531–549, 2022. doi: 10.1007/s11633-022-1371-y. URL https://link.springer.com/article/10.1007/s11633-022-1371-y.

Ge-Peng Ji, Jingyi Liu, Deng-Ping Fan, Huazhu Fu, and Nick Barnes. Colon-x: Advancing intelligent colonoscopy toward clinical reasoning, 2025. URL https://arxiv.org/abs/2512.03667v2. Original arXiv submission 2025; version 2 read, revised 14 March 2026.

Songtao Jiang, Yuan Wang, Ruizhe Chen, Yan Zhang, Ruilin Luo, Bohan Lei, Yeying Jin, Sibo Song, Zhibo Yang, Jimeng Sun, Jian Wu, and Zuozhu Liu. Act as you think: Reinforcing consistent reasoning in medical visual question answering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1788–1805. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.81. URL https://aclanthology.org/2026.acl-long.81/.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. E-ViL: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1244–1254, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Kayser_E-ViL_A_Dataset_and_Benchmark_for_Natural_Language_Explanations_in_ICCV_2021_paper.html.

Maxime Kayser, Cornelius Emde, Oana-Maria Camburu, Guy Parsons, Bartlomiej Papiez, and Thomas Lukasiewicz. Explaining chest X-Ray pathologies in natural language. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13435 of *Lecture Notes in Computer Science*, pp. 701–713. Springer, 2022. doi: 10.1007/978-3-031-16443-9_67. URL https://conferences.miccai.org/2022/papers/204-Paper2558.html.

Bidur Khanal, Sandesh Pokhrel, Sanjay Bhandari, Ramesh Rana, Nikesh Shrestha, Ram Bahadur Gurung, Cristian Linte, Angus Watson, Yash Raj Shrestha, and Binod Bhattarai. Hallucination-aware multimodal benchmark for gastrointestinal image analysis with large vision-language models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*,

volume 15969 of *Lecture Notes in Computer Science*, pp. 235–245, 2025. doi: 10.1007/978-3-032-05127-1_23.

Aminu Lawal, Niyoj Oli, Sachin Acharya, Prashnna Gyawali, Maria Carmen Romano, and Binod Bhattarai. A benchmark for hallucination detection in VLMs for gastrointestinal endoscopy, 2026. URL https://arxiv.org/abs/2606.24115.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2025.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.

Shi Li, Vinkle Srivastav, Nicolas Chanel, Saurav Sharma, Nabani Banik, Lorenzo Arboit, Kun Yuan, Pietro Mascagni, and Nicolas Padoy. SurgTEMP: Temporal-aware surgical video question answering with text-guided visual memory for laparoscopic cholecystectomy, 2026. URL https://arxiv.org/abs/2603.29962v4. Preprint, version 4.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023b. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. EndoBench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0076. Datasets and Benchmarks Track.

Rui Pan, Xiang Liu, Shizhe Diao, Renjie Pi, Jipeng Zhang, Chi Han, and Tong Zhang. LISA: Layerwise importance sampling for memory-efficient large language model fine-tuning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1817. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/687163285b8affc8ee933bdca8e75747-Abstract-Conference.html.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5. Official release blog; citation recommended by the Qwen3.5-9B model card. Accessed 2026-09-17.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8. Official release blog; citation recommended by the Qwen3.8-27B model card. Accessed 2026-09-17.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023. URL https://arxiv.org/abs/2305.18290.

Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. ARGUS: Hallucination and omission evaluation in video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20280–20290, 2025. URL

https://openaccess.thecvf.com/content/ICCV2025/html/Rawal_ARGUS_Hallucination_and_Omission_Evaluation_in_Video-LLMs_ICCV_2025_paper.html.

Itbaan Safwan, Muhammad Annas Shaikh, Muhammad Haaris, Ramail Khan, and Muhammad Atif Tahir. Multi-task learning for visually grounded reasoning in gastrointestinal VQA, 2025. URL https://arxiv.org/abs/2511.04384. MediaEval 2025 working paper; arXiv version 1.

Itbaan Safwan, Ramail Khan, Muhammad Annas Shaikh, and Muhammad Atif Tahir. Towards grounded GI endoscopy VQA via multi-task learning on small VLMs, 2026. URL https://arxiv.org/abs/2607.27122. arXiv version 1; accepted at the Workshop on Efficient Medical AI, EMA4MICCAI 2026.

Pritam Sarkar and Ali Etemad. Self-alignment of large video language models with refined regularized preference optimization. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3cbf627fa24fb6cb576e04e689b9428b-Abstract-Conference.html.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Arik, and Tomas Pfister. Mitigating object hallucination in MLLMs via data-augmented phrase-level alignment. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b73228d026ef49bfa49f95b7e330513e-Abstract-Conference.html.

Yiqing Shen, Han Zhang, and Mathias Unberath. Training LLMs with reinforcement learning over digital twin representations for reasoning-intensive surgical VideoQA, 2026. URL https://arxiv.org/abs/2606.17279. Version 1, 15 June 2026; preprint.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision function layer in multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 27510–27537. Curran Associates, Inc., 2025. doi: 10.52202/085713-0926. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/27c4e15d9af120d7fef04432c7db577f-Paper-Conference.pdf.

Jongmin Shin, Ka Young Kim, Eunki Cho, Seong Tae Kim, and Namkee Oh. SurgCheck: Do vision–language models really look at images in surgical VQA? *International Journal of Computer Assisted Radiology and Surgery*, 2026. doi: 10.1007/s11548-026-03739-4. URL https://doi.org/10.1007/s11548-026-03739-4.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.460. URL https://aclanthology.org/2024.emnlp-main.460/.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, 2021.

URL https://openaccess.thecvf.com/content/CVPR2021/html/Xiao_NExT-QA_Next_Phase_of_Question-Answering_to_Explaining_Temporal_Actions_CVPR_2021_paper.html.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via DPO: On-policy data hold the key. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10610–10620, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Mitigating_Hallucinations_in_Large_Vision-Language_Models_via_DPO_On-Policy_Data_CVPR_2025_paper.html.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024. doi: 10.1109/CVPR52733.2024.01310. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yu_RLHF-V_Towards_Trustworthy_MLLMs_via_Behavior_Alignment_from_Fine-grained_Correctional_CVPR_2024_paper.html.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: Open-source AI feedback leads to super GPT-4V trustworthiness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19985–19995, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yu_RLAIF-V_Open-Source_AI_Feedback_Leads_to_Super_GPT-4V_Trustworthiness_CVPR_2025_paper.html.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11704–11715, June 2026. URL https://openaccess.thecvf.com/CVPR2026?day=2026-06-05.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding, 2025a. URL https://arxiv.org/abs/2501.13106.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, June 2025b. URL https://openreview.net/forum?id=EELFGut39K.

Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. MMedPO: Aligning medical vision-language models with clinical-aware multimodal preference optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80207–80222, 2025. URL https://proceedings.mlr.press/v267/zhu25v.html.

A Evaluation Details

A.1 Models and Generation Protocol

We compare Qwen-Colon-8B with eight general-purpose open-source, three open-source medical-domain, and four proprietary MLLM baselines. The general-purpose open-source models use their official released checkpoints without additional fine-tuning on ColonExplainVQA and include Qwen3-VL-8B-Instruct (Bai et al., 2025), Qwen3.5-9B (Qwen Team, 2026a), Qwen3.8-27B (Qwen Team, 2026b), InternVL3.5-8B (Wang et al., 2025), LLaVA-Video-7B-Qwen2 (Zhang et al., 2025b), VideoLLaMA3-7B (Zhang et al., 2025a), MiniCPM-V 4.5 (Yu et al., 2026), and Molmo2-8B (Clark et al., 2026). Qwen3-VL-8B-Instruct provides the initialization for Qwen-Colon-8B; LLaVA-Video-7B-Qwen2 is abbreviated as LLaVA-Video-7B in the table. The medical-domain models are Hulu-Med-32B, Hulu-Med-30A3, and Lingshu-32B. The proprietary models are Gemini 3.1 Pro, Qwen3.8-Max, Doubao-Seed-2.0-Pro, and Qwen-VL-Max.

The evaluation uses 796 test questions across eight tasks, with 28–277 questions per task. Polyp identity is confined to training, as described in Section 3.1.2. The evaluation version uses four options for apparent-polyp-motion questions; across all tasks, questions contain two, three, four, or six options. The general-purpose open-source models use a common frame-sampling policy with their native visual processors and input templates. All evaluated models, including general-purpose open-source, medical-domain, and proprietary models, follow the same generation protocol. Given the video, question, and answer options, each model first generates a conclusion describing the visual evidence and then outputs its selected answer.

A.2 Judge and Metric Definitions

We assess answer correctness separately from reference consistency of the conclusion. Let $a_i = \mathbf{1}[\hat{y}_i = y_i]$ indicate whether the predicted option matches the reference answer, and let $\mathcal{D}_t$ contain the examples of task t. Answer accuracy averages correctness over all $N = 796$ questions, whereas task-macro accuracy gives equal weight to each of the $T = 8$ tasks:

$$\begin{aligned} \mathrm{Acc} &= \frac{1}{N}\sum_{i=1}^{N} a_i, \\ \mathrm{Acc}_{\mathrm{macro}} &= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{|\mathcal{D}_t|}\sum_{i\in\mathcal{D}_t} a_i. \end{aligned} \tag{4}$$

We measure reference contradiction using GPT-5.5 with medium reasoning effort, followed by human verification. Each request contains one predicted conclusion and its manually curated reference conclusion, without the video. The judge decomposes the prediction into factual claims and labels each as entailment, contradiction, or neutral relative to the reference. A contradiction requires an explicit conflict under the same object, temporal scope, and semantic conditions. Details absent from the reference are neutral rather than automatically contradictory. Each detected contradiction must include verbatim evidence from both conclusions and a short justification; returned fields and quoted spans are validated before aggregation. Human reviewers then verify the model judgments and supporting evidence, correcting errors before aggregation.

Let $c_i = 1$ when at least one claim in example i is confirmed as contradictory after human verification and $c_i = 0$ otherwise. The reference contradiction rate (RCR) measures the proportion of conclusions containing a detected conflict. The conditional rate $\mathrm{CR}_{\mathrm{correct}}$ measures how often such a conflict remains despite a correct answer. Joint accuracy requires both a correct answer and no

detected reference contradiction:

$$\begin{aligned} \mathrm{RCR} &= \frac{1}{N}\sum_{i=1}^{N} c_i, \\ \mathrm{CR}_{\mathrm{correct}} &= \frac{\sum_{i=1}^{N} a_i c_i}{\sum_{i=1}^{N} a_i}, \\ \mathrm{JointAcc} &= \frac{1}{N}\sum_{i=1}^{N} a_i(1-c_i). \end{aligned} \tag{5}$$

All metrics are reported as percentages. None remain among the reported contradiction results.

B Layer-Selection Screening Results

Functional interventions. CFL probes the SFT model without updating its parameters. For language attention, we remove visual key–value entries from the selected layer's attention computation during a single cached query at the final token of a fixed answer prefix, preserving the prefill cache. For language MLPs, we zero outputs at visual-token positions during prefill. Visual-attention branch outputs and merger outputs are ablated during a fresh visual forward pass.

Development pool and screening. We use a fixed 600-question DPO development pool drawn from the SFT training split. Source-connected groups are selected with seed 42 to approximately preserve task proportions across all nine tasks. These samples were seen during SFT but are excluded from DPO optimization and replay by source group. The pool is distinct from the 200-question checkpoint-validation set and the 796-question test set. Each candidate starts from the same SFT model and receives 20 DPO updates. Screening uses strict QA accuracy, requiring the correct option, valid response format, and answer–conclusion option consistency.

Figure 6 compares single-family layer budgets and joint configurations. On this 600-question pool, the SFT baseline achieves 84.33% accuracy, while jointly updating three visual-attention layers and 18 language-MLP layers without merger updates reaches 91.17%, an improvement of 6.84 percentage points. The corresponding single-family configurations achieve 89.17% and 89.33%. The joint configuration performs best among the reported candidates, while larger layer budgets do not consistently improve accuracy, supporting selective adaptation.

C Additional Dataset Statistics and Curation Examples

C.1 Dataset Composition and Correction Profile

Figure 7 characterizes ColonExplainVQA by data source, task, and correction type. The 7,962 video-question pairs combine HyperKvasir (3,878; 48.7%), REAL-Colon (3,763; 47.3%), and SUN-SEG (321; 4.0%). Their complementary annotations support questions about endoscopic findings, examination quality, treatment-related content, and temporal observations. This coverage makes visual-evidence supervision central to the benchmark: an explanation must identify the appearance, track changes across frames, and connect the observed lesion or instrument to the selected answer.

The nine tasks span different levels of frequency and visual reasoning. Polyp detection and main endoscopic findings contribute 2,744 and 2,113 pairs, respectively, together accounting for 61.0% of the corpus, followed by treatment appearance with 1,157 pairs. The remaining tasks cover apparent polyp motion (548), instrument motion (487), bowel preparation (321), treatment content (287), pathology classification (284), and polyp identity (21). Together, they extend the learning target from recognizing findings to interpreting motion, procedural content, and identity across clips. This diversity motivates task-balanced evaluation and adaptation. We report task-macro accuracy alongside overall accuracy and give equal weight to tasks during CFL sensitivity analysis, connecting module selection to visual functions across the task collection.

The correction profile highlights the value of reviewing explanations alongside answers. Relative to the original Plus responses after formatting normalization, 2,242 examples (28.2%) have a changed

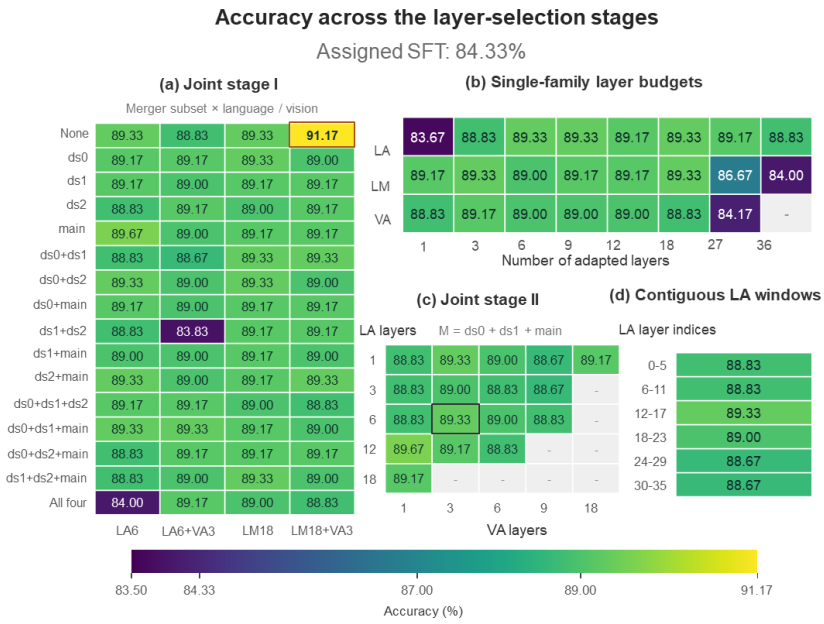


Figure 6: Accuracy (%) on the 600-question DPO development pool: (a) merger combinations, (b) single-family layer budgets, (c) joint attention budgets, and (d) contiguous language-attention windows. LA, LM, and VA denote language attention, language MLPs, and visual attention. The outlined cell marks the selected VA3+LM18 configuration; dashes denote unreported entries.

answer, 2,163 (27.2%) retain the answer but revise the conclusion, and 3,557 (44.7%) remain unchanged. Overall, 4,405 examples (55.3%) receive a revision, with conclusion-only corrections accounting for 49.1% of this group. These cases directly motivate our focus on explanatory evidence: retaining the correct option can still require correcting unsupported claims about lesion appearance, instrument action, or procedural state, as illustrated in Figure 8. Minimal edits target these claims while preserving compatible content and wording. The reviewed answers and conclusions supervise SFT, while retained incorrect teacher responses supply negative examples for subsequent preference learning. ColonExplainVQA thus supports learning both which answer to select and which visual evidence can justify it.

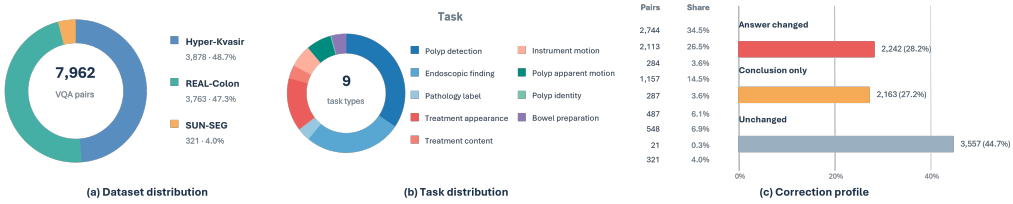


Figure 7: Dataset sources, task distribution, and correction profile of ColonExplainVQA. Changes are measured against the original Plus responses after formatting normalization; conclusion-only revisions retain the selected answer.

C.2 Examples of Visual-Evidence Correction

Figures 8 and 9 contrast original teacher responses with reviewed targets. The first preserves correct answer options while correcting unsupported claims about instrument placement and action. The second revises both the selected answer and its supporting conclusion. These examples complement the correction statistics and illustrate why supervision must distinguish answer correctness from the evidence and strength of claims in an explanation.



Q: What is the main endoscopic finding shown in this video?
A. A colonic anastomosis. B. A colonic stent.
C. A polypoid lesion. D. A post-resection site.

Qwen3.7 Plus Answer: C. A polypoid lesion.
Conclusion: ✗
The clip clearly displays a large, dye-lifted polypoid lesion with a snare loop placed around its base, characteristic of endoscopic mucosal resection preparation. No anastomosis, stent, or open resection defect is visible. I choose C.

Human-Corrected Answer: C. A polypoid lesion.
Conclusion: ✓
The clip clearly displays a large, blue-stained elevated polypoid lesion with a metallic snare-like wire adjacent to it. The lesion remains rounded and prominently elevated across the frames. I choose C.



Q: Which category of treatment-related content is primarily shown in this video?
A. Snare resection-related content.
B. Endoscopic clip placement-related content.
C. A colonic self-expanding stent.

Qwen3.7 Plus Answer: A. Snare resection-related content.
Conclusion: ✗
The visual evidence clearly shows a large polyp being engaged by a wire snare loop, which is the defining instrument for snare resection. There are no clips or stents visible to suggest alternative treatments. I choose A.

Human-Corrected Answer: A. Snare resection-related content.
Conclusion: ✓
The visual evidence clearly shows a large polyp with a thin wire consistent with a snare, which is the defining instrument for snare resection. A thin dark wire is visible beside the lobulated polyp in the late frames. I choose A.

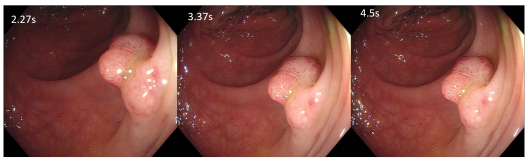
Figure 8: Conclusion-only corrections for main findings (top) and treatment content (bottom). Correct answers are retained while unsupported claims about snare placement or engagement are revised. Red marks unsupported claims; green marks correct answers and reviewed descriptions.



Which category of treatment-related content is primarily shown in this video?
A. Snare resection-related content.
B. Endoscopic clip placement-related content.
C. A colonic self-expanding stent.

Qwen3.7 Plus Answer: A. Snare resection-related content. ✗
Conclusion:
The video clearly displays a snare device, identifiable by its thin wire loop structure, interacting with colonic tissue. No clips or stents are visible. The visual evidence strongly supports snare resection-related content as the primary category. I choose A.

Human-Corrected Answer: C. A colonic self-expanding stent. ✓
Conclusion:
The video clearly displays a colonic stent, identifiable by its metallic mesh lattice structure, interacting with colonic tissue. A stent-like mesh is visible. The visual evidence strongly supports colonic self-expanding stent content as the primary category. A yellow metallic lattice tube with large openings is visible beside mucosa. I choose C.



Q: What is the most likely pathological category of the polyp shown in this colonoscopy video?
A. Hyperplastic polyp. B. Low-grade adenoma.
C. Invasive carcinoma. D. Sessile serrated lesion.
E. Traditional serrated adenoma. F. High-grade adenoma.

Qwen3.7 Plus Answer: B. Low-grade adenoma. ✗
Conclusion:
The polyp appears as a lobulated, pink, non-ulcerated mass consistent with a low-grade adenoma based on visible morphology. I choose B.

Human-Corrected Answer : F. High-grade adenoma. ✓
Conclusion:
The polyp appears as a lobulated, pink, non-ulcerated mass, with pathology most consistent with a high-grade adenoma. I choose F.

Figure 9: Answer and conclusion corrections for treatment content (top) and pathology classification (bottom). The reviewed treatment example identifies stent-related content. The pathology target follows the source annotation; histological grade is not directly observable in the clip. Red marks original errors; green marks corrected content.

C.3 Comparison with Existing VQA Resources

Compared with the resources in Table 3, ColonExplainVQA offers two complementary advantages for studying visual-evidence hallucinations. Most listed benchmarks use individual images, whereas our video inputs preserve temporal information for interpreting instrument actions, apparent polyp motion, and polyp identity across clips. Relative to video QA resources such as REAL-Colon-VQA, we emphasize paired evidence curation: original conclusions are retained alongside minimally human-corrected versions, making unsupported claims and their targeted corrections directly inspectable. Preserving supported wording is intended to focus supervision on factual corrections while reducing stylistic variation. The 2,163 conclusion-only revisions capture errors that answer accuracy alone overlooks. This supports separate evaluation of answers and visual evidence, while providing curated responses for SFT and retained incorrect-answer responses for DPO.

Table 3: Comparison with endoscopic VQA resources. Baseline statistics follow EndoBench (Liu et al., 2025), with REAL-Colon-VQA added from its paper and released annotations (Drago et al., 2026).

Benchmark	Size	Tasks	Data sources	Visual input	Paired corrections
OmniMedVQA*	1.88k	3	3P	Image	NR
GMAI-MMBench*	3.75k	7	16P	Image	NR
Kvasir-VQA	6.50k[a]	6	2P	Image	NR
Surgical-VQA	54k	5	2P	Image /multi-frame[b]	NR
SSG-VQA	960k	5	3P	Image	NR
EndoChat	396k	5	3P	Image	NR
ColonINST	300k+[a]	4	19P	Image	NR
EndoVQA-Instruct	446k+	12	20P+1I	Image	NR
EndoBench	6.83k	12	20P+1I	Image	NR
REAL-Colon-VQA	5.20k	20[c]	1P	Video	NR
ColonExplainVQA	**7.96k**	**9**	**3P**	**Video**	**Yes**

k = 1,000; P/I: public/in-house data sources. *Endoscopic subsets. [a]Image counts; other sizes follow the reported QA/instruction counts. Visual input describes the VQA input rather than its source recording. [b]Frame-level QA with an additional three-frame temporal variant. [c]20 distinct question types in the released annotations; the paper reports 18 categories across six reasoning domains. Paired corrections retain original and minimally human-corrected visual-evidence conclusions; NR: not reported.

C.4 Additional Qualitative Examples

Figure 10 supplements the curation examples in Section 3.1.2. Both answers remain correct, while the conclusions overstate mucosal visibility or assert an unsupported dye-injection action. Minimal revisions retain supported observations and replace unsupported claims, illustrating the value of correcting evidence independently of answer labels.

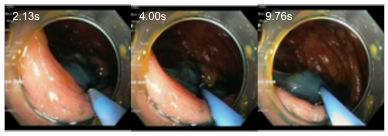
Figure 11 supplements the qualitative analysis in Section 4. Qwen 3.7 Plus selects the correct leftward direction but describes a stationary polyp; MiniCPM-V 4.5 cites lesion features absent from the reviewed description. These cases show how hallucinations can affect temporal evidence and lesion appearance, even when the selected answer is correct. The pathology target follows the source label and does not establish visually observable histology.



Q: Which bowel-preparation group best describes the colon segment shown in this video?
A: BBPS 0–1
B: BBPS 2–3

Qwen3.7 Plus Answer: B. BBPS 2–3
Conclusion: ✗
The segment shows clear visualization of the mucosal surface with only minor residual fluid and debris that does not impair inspection. This corresponds to a good preparation score where the lining is fully discernible. I choose B.

Human-corrected Answer: B. BBPS 2–3
Conclusion: ✓
The segment shows clear visualization of the mucosal surface with small residual fluid and focal adherent yellow debris, with most mucosa visible. This corresponds to adequate preparation with most of the lining discernible. Clear pink mucosa is visible despite focal yellow adherent residue along one wall. I choose B.

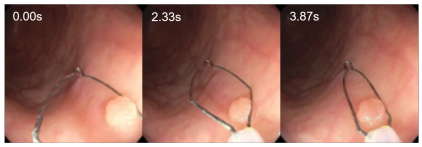


Q: Which treatment-related appearance is primarily shown in this video?
A: A dyed, lifted polyp
B: A dyed resection site or margin

Qwen3.7 Plus Answer: A. A dyed, lifted polyp
Conclusion: ✗
The visual evidence clearly displays an injection needle applying blue dye to a raised mucosal lesion, creating a lifted appearance characteristic of chromoendoscopy-assisted polypectomy preparation. No resection site or margin is visible. I choose A.

Human-corrected Answer: A. A dyed, lifted polyp
Conclusion: ✓
The visual evidence clearly displays a blue-tipped instrument adjacent to a raised mucosal lesion, with a lifted appearance compatible with a dyed, lifted polyp. Frames show blue-stained elevated mucosa protruding beside a blue instrument tip. I choose A.

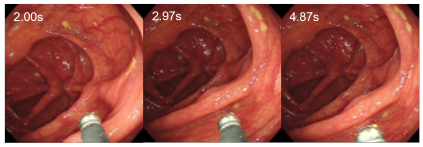
Figure 10: Additional visual-evidence corrections. Answers remain unchanged while revisions qualify mucosal visibility (top) and replace an unsupported injection claim with visible instrument and tissue descriptions (bottom). Red marks unsupported claims; green marks reviewed content.



Q: From the beginning to the end of this clip, what is the dominant apparent motion direction of the polyp region in the image plane?
A. Up and to the left.
B. To the left.
C. Down and to the left.
D. Upward.

Qwen3.7 Plus Answer: B. To the left.
Conclusion: ✗
The polyp remains relatively stationary in the image plane throughout the clip, with only slight movement due to the instrument's adjustment, but no clear dominant motion direction is evident. I choose B.

Human-corrected Answer (GT): B. To the left.
Conclusion: ✓
The polyp region starts in the lower right and translates to the left within the image plane by the end of the clip. The polyp shifts horizontally leftward from the right side toward the center across frames 0-23. I choose B.



Q: What is the most likely pathological category of the polyp shown in this colonoscopy video?
A. Sessile serrated lesion.
B. High-grade adenoma.
C. Low-grade adenoma.
D. Invasive carcinoma.
E. Traditional serrated adenoma.
F. Hyperplastic polyp.

MiniCPM-V-4.5 Answer: B. High-grade adenoma. ✗
Conclusion:
The polyp appears large, broad-based, and has a smooth surface with a central depression, which are features suggestive of a high-grade adenoma. The surrounding mucosa shows some inflammation and hyperemia, but no clear signs of invasive carcinoma or other specific serrated features are visible. I choose B.

Human-corrected Answer (GT): C. Low-grade adenoma. ✓
Conclusion:
No distinct polyp is visible; the mucosa appears inflamed with debris. Without a clear lesion, a low-grade adenoma is the most likely pathological category, with no clear malignant features. I choose C.

Figure 11: Additional hallucination examples. A correct motion answer accompanies an inconsistent stationary-motion claim (top), while a pathology response cites unsupported lesion features (bottom). Red marks erroneous claims; green marks reference answers and reviewed descriptions.