HEICO-FOCUS: A CLINICALLY GROUNDED DATASET FOR LONG-CONTEXT VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in Vision-Language Models (VLMs) have led to rapid progress in video understanding across a wide range of benchmark tasks. However, existing evaluations largely focus on short-term reasoning, failing to assess a critical capability: maintaining cumulative temporal consistency over extended time horizons. To close this evaluation gap, we introduce HeiCo-FOCUS, a clinically grounded dataset for evaluating long-context video understanding through the task of **F**oreign **O**bject **C**ontextual **U**nderstanding in **S**urgery. Built on a dataset of **Hei**delberg **Co**lorectal surgeries, this task requires models to continuously track multiple objects as they are inserted, manipulated, occluded, and removed over procedures lasting up to hours. HeiCo-FOCUS comprises 30,000 visual question answerings (VQAs) pairs covering five core capabilities: object recognition, temporal grounding, aggregation, event and procedural understanding, and complex reasoning. The dataset was constructed through a rigorous multi-stage annotation pipeline involving large-scale crowd annotation and 39 surgical domain experts to ensure high quality and clinical relevance. To systematically probe model behavior, we introduce a multi-track evaluation framework that progressively increases temporal and contextual demands from single frames to full procedures. Experiments with ten frontier VLMs show that HeiCo-FOCUS tasks are far from solved: only around half of the models clearly outperform a text-only baseline. Across the video tracks, models perform best on *event and procedural understanding* (mean Accuracy: 56.5% across all models), while *temporal grounding* remains particularly challenging for all evaluated models (mean Accuracy: 19.7%). We therefore expect HeiCo-FOCUS to serve as a catalyst for the development of models capable of reliable, temporally consistent reasoning over hours-long videos.

1 INTRODUCTION

Recent progress in general-domain Vision-Language Models (VLMs) has enabled temporal reasoning over extended video streams (Bai et al., 2025; Comanici et al., 2025). Emerging long-video benchmarks evaluate these capabilities, showing improvements in tasks such as visual question answering (VQA) and event understanding over longer temporal contexts (Wu et al., 2024; Chandrasegaran et al., 2024; Fu et al., 2025; Shen et al., 2024; Zohar et al., 2025; Wang et al., 2025b; Liu et al., 2024; Zhou et al., 2025; Hong et al., 2023). However, these benchmarks primarily assess local reasoning within extended segments and do not require persistent tracking of entities across long horizons or the aggregation of state over time.

Surgical videos represent a particularly important and challenging domain: procedures unfold over tens of minutes to hours, involve continuous scene changes, and require precise interpretation of fine-grained visual cues. Despite this, the surgical AI community has so far lacked a dedicated, standardized benchmark to assess whether emerging long-context video capabilities translate to real clinical workflows (see Table 1 in Appendix A). The HeiCo-FOCUS dataset addresses this unmet need by introducing a structured benchmark centered on a concrete patient safety problem: ensuring the retrieval of foreign objects at the end of surgery (see Figure 1). In minimally invasive surgical procedures, surgical sponges, sutures, and other foreign objects are inserted into the abdomen via trocar sites to carry out steps of the procedure. Unintentionally leaving foreign objects in the abdomen is a rare but consequential incident, as they can cause serious complications (Badiee et al., 2025; Weprin et al., 2021). Using foreign objects as the central theme, HeiCo-FOCUS offers com-

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

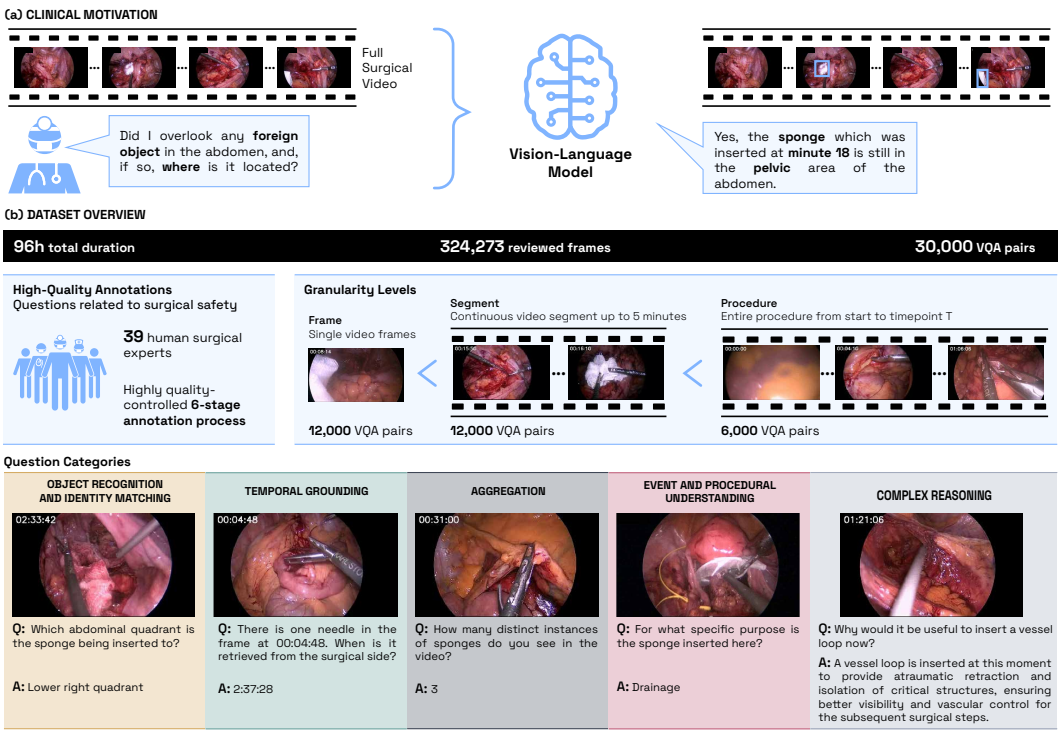


Figure 1: **Overview of the HeiCo-FOCUS benchmark. a) Clinical motivation.** Ensuring the retrieval of foreign objects at the end of a surgery is critical, as retained objects can lead to severe complications. **b) Dataset overview.** The dataset comprises 96 hours of surgical video annotated with 30,000 visual question answering (VQA) pairs spanning five core capabilities (see Figure 2).

prehensive capability coverage, spanning object recognition, temporal grounding, aggregation, event and procedural understanding, and complex reasoning, enabling fine-grained analysis of model capabilities (Figure 2). From a long video benchmarking perspective, keeping track of foreign objects is a highly attractive application because it inherently requires persistent memory, temporal consistency, and aggregation over extended time horizons.

The uniqueness of HeiCo-FOCUS is rooted in three fundamental design decisions:

1. **Quality-first approach:** While large-scale, large language model (LLM) curated datasets have significantly accelerated progress in AI, they primarily serve the purpose of training models and inevitably contain noise. In contrast, evaluation benchmarks require the highest possible level of reliability, as their credibility directly determines the validity of scientific comparisons. Recent work in medical AI has highlighted how insufficient data provenance can undermine research quality and even contribute to paper retractions (Gibson et al., 2026). **HeiCo-FOCUS directly addresses this critical gap by providing a fully trusted benchmark:** (1) building upon a trusted, clinically curated base dataset (HeiCo; Maier-Hein et al., 2021), and (2) involving human experts at every stage of the annotation pipeline to ensure high-quality, clinically meaningful annotations and evaluation. Notably, compared to more than 400 analyzed medical imaging AI benchmarks (median: 3 annotators) (Reinke et al., 2026), HeiCo-FOCUS involved more than an order of magnitude more domain experts, reflecting the substantial annotation effort required to produce a clinically reliable benchmark.
2. **Persistent object identities for long-context VQA:** Existing surgical VQA datasets largely generate question-answer pairs by converting existing vision annotations using large language models. This introduces a fundamental limitation: many clinically relevant questions are never covered. For example, questions related to retained foreign objects require persistent object identities over long time horizons. Existing datasets do not provide instance-level annotations, current models cannot reliably generate them automatically, and

manual annotation is prohibitively expensive as illustrated by the enormous cost of our dataset. Furthermore, automatically generated questions are inherently limited by the underlying annotations and often fail to capture clinically meaningful long-context reasoning. HeiCo-FOCUS closes this gap through a dedicated, expert-driven annotation pipeline that explicitly annotates object instances (Figure 3). **Without HeiCo-FOCUS, the community will lack a clinically grounded benchmark for systematically evaluating long-context understanding tasks that require persistent object identities.**

3. **Clinical relevance:** Many surgical AI datasets have been designed around procedures and tasks that are easiest to acquire and annotate, rather than those with the greatest clinical need. This is nicely illustrated by a recent scoping review (Carstens et al., 2025) that showed that laparoscopic cholecystectomy accounts for more than 50% of surgical AI studies and has therefore largely driven progress in the field. This is interesting because cholecystectomies are technically straightforward procedures with low complication rates ($\sim$2–3%; Duca et al., 2003), suggesting that their predominance is driven by available landmark datasets. As datasets shape research priorities, benchmarks should steer research toward clinically important problems. Uniquely, HeiCo-FOCUS addresses both: a huge technical challenge (long-context video understanding) and an actual globally important clinical problem. **Instead of further expanding cholecystectomy datasets, we deliberately chose colorectal procedures (complications in up to one third of patients; Tevis & Kennedy, 2016), which present an up to 15 times higher complication rate, 5-fold longer procedures, and on average three times more foreign object instances per video.**

Applying state-of-the-art (SOTA) VLMs to HeiCo-FOCUS highlights current limitations in long-context understanding, temporal consistency, and safety-critical reasoning. Notably, HeiCo-FOCUS pushes existing models to their limits, indicating that the benchmark is unlikely to be saturated in the near term (Figure 4).

2 Related Work

Long video benchmarks, such as (Wu et al., 2024; Chandrasegaran et al., 2024; Fu et al., 2025; Shen et al., 2024; Zohar et al., 2025; Wang et al., 2025b; Liu et al., 2024; Zhou et al., 2025; Hong et al., 2023; Mangalam et al., 2023; Wu et al., 2025; Yu et al., 2019) have been instrumental in advancing video understanding beyond short clips by introducing tasks that require reasoning over extended temporal contexts. However, they still typically rely on relatively short temporal windows—often limited to a few minutes.

Several groups have pioneered the development of surgical VQA datasets, with existing resources summarized in Table 1 (Appendix A). These datasets have significantly advanced the field by providing large-scale question–answer pairs and enabling the training of multimodal models in surgical domains. In particular, recent datasets have reached substantial scale in terms of both video volume and number of VQA pairs. However, despite these advances, several important limitations remain. First, existing datasets are restricted to frame-level or short video clips, with little to no support for long-context reasoning beyond a few minutes. As a result, they do not capture the temporal dependencies and cumulative reasoning required in real surgical workflows. Second, questions are often limited to perception and event understanding. We are not aware of a single surgical AI dataset that addresses temporal grounding and long-context reasoning. Third, many datasets rely either on rule-based categorical question generation (Seenivasan et al., 2022) or on AI-generated VQA pairs sourced from public platforms like YouTube, utilizing automated audio transcription and rephrasing (Li et al., 2026; Perez et al., 2026). Fourth, expert verification is largely absent, with only a small fraction of datasets reporting systematic review by medical professionals.

Together, these limitations highlight the need for a benchmark that combines long-context video understanding, clinically grounded question design, and rigorous expert validation.

3 HEICO-FOCUS

HeiCo-FOCUS is a clinically grounded benchmark for evaluating long-context video understanding. Based on all 30 **Hei**delberg **C**olorectal Surgeries of the HeiCo dataset (Maier-Hein et al., 2021), it contributes 30,000 VQA pairs addressing **F**oreign **O**bject **C**ontextual **U**nderstanding in **S**urgery.

Dataset overview and license. The original HeiCo dataset comprises a total of 30 colorectal surgical procedures, including 10 sigmoid resections, 10 rectal resections, and 10 proctocolectomies, annotated with surgical phases and instruments. We deliberately build upon this trusted resource to (1) leverage an established, ethically governed clinical dataset with transparent provenance, (2) focus on complex procedures with substantial complication rates and a broad diversity of foreign objects, and (3) enable researchers to combine HeiCo-FOCUS with the rich existing HeiCo annotations, facilitating complementary analyses of surgical phases, instruments, and long-context object dynamics on the same procedures. For each procedure, we generated 1,000 VQA pairs, comprising 400 questions for the FRAME track, 400 for the SEGMENT track, and 200 for the PROCEDURE track. The VQA generation pipeline (see Section 4) was designed to achieve approximately balanced coverage across all capability categories and temporal contexts. To support fine-grained evaluation and stratified analysis, each VQA pair is associated with structured metadata including the expected answer format (e.g., open ended, an integer number, a time estimate, etc.) and associated capabilities (see Section 4.1). Following the original HeiCo dataset split (Maier-Hein et al., 2021), we recommend using the sigmoid resection procedures as the held-out test set (4,000 FRAME, 4,000 SEGMENT, and 2,000 PROCEDURE questions). To comply with the original HeiCo license, HeiCo-FOCUS is released under CC BY-NC-SA 4.0. Usage and access details are presented in Appendix B.

The design of HeiCo-FOCUS was guided by the following principles:

Clinical relevance and real-world grounding. A central design objective was to ensure that the benchmark is both clinically meaningful and technically rigorous. To achieve this, surgical experts were closely involved throughout all stages of the benchmark development.

Fine-grained understanding of VLM capabilities. To enable a detailed analysis of model behavior, we developed a comprehensive capability taxonomy (see Figure 2) that spans multiple dimensions of visual and temporal understanding. Secondly, to allow for controlled analysis of how model performance evolves from instantaneous perception to long-horizon reasoning, the benchmark was structured into three complementary tracks: a **FRAME Track** to assess single-image understanding, focusing on perception and aggregation tasks, a **SEGMENT Track** to assess short-term video understanding (all capabilities) within video segments of up to 5 min and a **PROCEDURE Track** to assess long-context capabilities over extended video durations from the beginning up to a time point t (5–296 min).

Quality-controlled annotation. To ensure high annotation quality, we implemented a multi-stage annotation pipeline (see Figure 3, Section 4.2, and Table 2 in Appendix D) combining automated methods, large-scale crowd annotation, and expert annotation and verification. Overall, 39 domain experts (clinical and technical) were involved in the data annotation (7 expert surgeons, 18 medical students, 14 surgical AI researchers).

Bias-aware validation. To prevent questions from being answered based solely on prior knowledge or textual plausibility (Mayer et al., 2025; Asadi et al., 2026), all questions were filtered with a panel of text-only models (*blind filter*), as detailed below. Furthermore, the capability taxonomy enables stratified analyses across different capabilities, preventing overrepresentation of simpler tasks (e.g., object identification) from dominating performance estimates. Finally, the hierarchical structure of the data - where multiple questions are associated with the same video or segment - is explicitly accounted for in the statistical analysis, for example through hierarchical aggregation and bootstrapping.



Figure 2: **HeiCo-FOCUS taxonomy with sample questions**; definitions in Appendix C.

4 VQA Generation Pipeline

The HeiCo-FOCUS dataset was constructed through a multi-stage annotation pipeline designed to balance scalability with high clinical and technical quality.

4.1 Taxonomy

Our taxonomy for surgical VQA is depicted in Figure 2. For question categorization, each question is assigned to one primary capability category and may optionally be tagged with additional secondary sub-categories to reflect its multi-faceted nature. The primary category is determined by the dominant reasoning requirement of the question. For example, *"Where"* questions are assigned to *object recognition and identity matching* $\rightarrow$ *spatial localization*, *"How many"* questions to *aggregation*, and *"When"* questions to *temporal grounding* $\rightarrow$ *temporal localization*. This primary assignment enables the balancing of VQA pairs across capabilities while secondary labels allow fine-grained stratified evaluation. The complete taxonomy is provided in Appendix C.

4.2 Annotation Pipeline

Our annotation pipeline combines large-scale crowd annotation, automated processing, and expert-driven refinement to progressively enrich the data from raw video to clinically meaningful and verified VQA pairs, as illustrated in Figure 3 and detailed in Appendix D.

The process began with **Stage 1**, in which short video clips were screened to identify the presence of foreign objects. In **Stage 2**, frames sampled at 1 frame per second (fps) from relevant segments were annotated with bounding boxes and class labels. These initial annotations were subsequently refined in **Stage 3** with automated methods that correct and complete bounding box predictions.

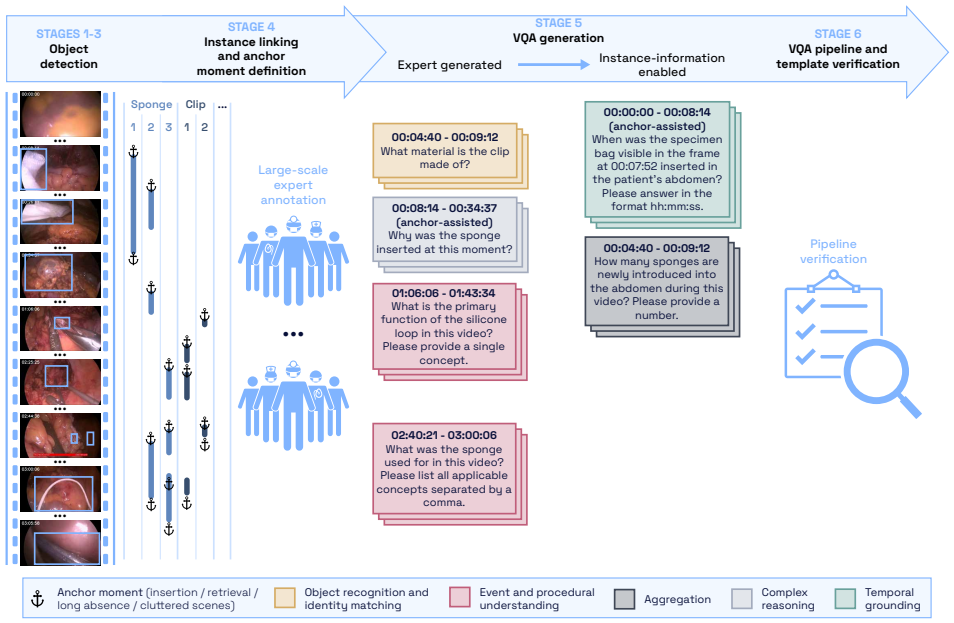


Figure 3: **Multi-stage annotation pipeline of HeiCo-FOCUS.** Raw surgical videos are progressively transformed into high-quality, clinically grounded visual question answering (VQA) pairs. Stages 1-3 perform foreign object detection, followed by Stage 4, where instances are linked across time. In Stage 5, VQA pairs are generated using a combination of expert-designed questions and instance-informed automatic templates. Stage 6 conducts end-to-end expert verification.

Stage 4 was by far the most resource-intensive one. Here, 18 medical students assigned consistent instance identities to individual objects across time, yielding a structured representation of foreign object trajectories. This required the generation of bounding boxes around foreign objects at 1 fps and for more than 300 k frames in total. Appendix I illustrates annotation challenges.

Building on the fine-grained instance-level annotations, **Stage 5** derived anchor moments, such as insertion and retrieval events, long absences, and cluttered scenes, which served as key reference points for VQA generation. VQA pairs were then created through a combination of automated template-based methods filtered by medical students for validity, and surgical expert-designed questions, ensuring coverage across capability categories and complexity levels. Finally, in **Stage 6**, the whole VQA generation pipeline was verified by the analysis of about 500 diverse sample questions, covering all capabilities and foreign objects.

Once the complete annotation pipeline had been validated, it was applied to generate a total of 30,000 VQA pairs for the HeiCo-FOCUS benchmark, as detailed in the following section. To further promote transparency, quality control, and continuous improvement, we additionally provide a community feedback mechanism that enables users to report potentially ambiguous or erroneous cases during a restricted post-release review period.

4.3 VQA Design

The VQA generation process builds upon Stage 4 of the annotation pipeline (see Figure 3), which resulted in 324,273 manually curated frames containing 79,682 verified foreign-object bounding boxes. Every class except clips carries instance identities across time (55,629 boxes); as it turned out to be infeasible to do instance linking for clips (even for expert surgeons), they were annotated for visibility only and excluded from questions that require following an individual object over time. These high-quality instance-level annotations provide a structured representation of object trajectories over time and enable automatic generation of questions for a large subset of capabilities, including temporal grounding, aggregation, and multi-step compositional reasoning. However, several sub-capabilities-particularly functional reasoning, causal reasoning, and anatomically grounded localization-required additional clinical expertise beyond what can be derived from instance an-

notations alone. To ensure both high clinical validity and broad capability coverage, the question generation therefore followed a multi-stage, expert-guided pipeline combining manual annotation, expert validation, and automated augmentation.

Candidate question and anchor moment generation. We first defined a catalog of candidate question templates covering all benchmark tracks and capability categories(e.g., "How many <foreign object>s were inserted during this video segment"). Each template was associated with metadata specifying its applicable track, primary and secondary capability labels, relevant foreign object classes, required visual evidence, and answer format. In parallel, we derived informative anchor moments from the instance-level annotations, including potential insertion and retrieval events, long absence intervals, and visually cluttered frames or segments. These anchors served as high-value time points for clinically meaningful, technically challenging questions.

Expert annotations. Based on question and anchor moment candidates, medical students validated automatically generated anchor moments, confirming insertion and retrieval events and prefiltered generated questions. These questions were either answered when confidence was high, discarded if ill-posed, or flagged for expert review when uncertainty remained. Next, expert surgeons reviewed these flagged cases, validated anchor moments, and provided authoritative answers. In addition, they enriched the dataset by contributing new questions that could be asked at any time during the procedures. In this step, experts were asked to target underrepresented capabilities such as reasoning, functional understanding, and anatomically grounded localization, for which questions cannot be automatically answered based on instance annotations.

Template-based question generation and balancing. To achieve a balanced benchmark, clinician-written questions were complemented with question templates systematically instantiated both at clinically informative anchor moments and at other eligible time points to ensure broad coverage across capabilities, foreign-object classes, answer values, and temporal contexts. For sampling, the taxonomy was represented as a tree whose leaves contained 85 question templates, and questions were iteratively drawn from the least represented capability, sub-capability, template, and template instance. Within templates, sampling was additionally balanced across answer values and distributed over time.

After end-to-end verification of the complete annotation process, we followed the described steps to generate 1,000 VQA pairs per video. To remove questions answerable from text alone, three models (Mistral Small 2603, Gemini 2.5 Flash Lite, and DeepSeek V4 Flash) answered every candidate question without the image or video, with the option to answer "unsure"; a question was discarded if all three answered correctly.

5 Experiments and Results

Metrics and ranking. We evaluated model performance primarily using accuracy. Answers were parsed according to their expected format (Table 3). Open-ended, multiple-choice, and matching answers were graded for semantic equivalence by a Large Language Model (LLM) judge (gpt-oss-120b, validated against two human raters in Appendix F).

Inference and model selection. All evaluations were conducted via cloud-based Application Programming Interfaces (APIs) to ensure a standardized and reproducible evaluation environment. The benchmark includes a representative selection of current frontier VLMs, namely GPT-5.6 Sol, Claude Sonnet 5, Gemini 3.8 Flash, Gemini 3 Flash, Gemini 3.5 Flash Lite, Meta Muse Spark 1.3, SpaceXAI Grok-4.3, DeepSeek V4 Flash Vision, Amazon Nova 2 Lite, and Xiaomi MiMo-V 2.5. Details on the prompts provided to the VLMs can be found in Appendix E.

Data preprocessing. The input pipeline was adapted to the native capabilities of each model. Depending on model support, inputs were provided either as native video or as chronologically ordered frame sequences. To ensure consistency across tracks while balancing temporal granularity and computational efficiency, SEGMENT windows were sampled at 1 fps, and PROCEDURE windows were represented by 300 uniformly spaced frames. Visual inputs were standardized to a spatial resolution of 210×360 pixels.

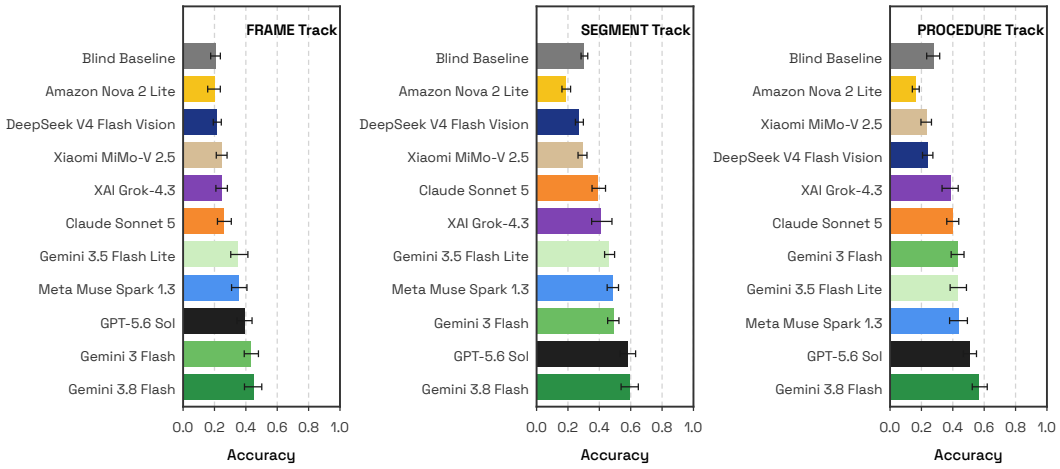


Figure 4: **Overall performance for all models and tracks.** Accuracy is the mean of per-video Accuracy. Uncertainty is depicted using bootstrapped 95% confidence intervals (CIs).

Baseline. As baseline, we used a no-image setting in which GPT-5.6 Sol received only the textual prompt and question, but no visual input. This baseline assessed the extent to which questions can be answered from textual priors, dataset biases, or linguistic plausibility alone, without visual or temporal grounding.

Results. In the human baseline, three raters (two surgical residents and one medical student) answered 300 questions on the test videos, drawn from an earlier version of the question pool. They reached 62.0%, 69.8%, and 57.4% Accuracy on the FRAME, SEGMENT, and PROCEDURE track, respectively (Appendix G).

The performance of the evaluated VLMs on the test data is summarized in Figure 4. On the FRAME track, the highest overall Accuracy was obtained by Gemini 3.8 Flash (44.8% [39.1, 50.1] 95% confidence interval (CI)), followed by Gemini 3 Flash (43.2% [39.0, 48.0]) and GPT-5.6 Sol (38.9% [34.3, 44.0]), improving on the text-only baseline (20.7%) by 24.1, 22.5, and 18.2 percentage points (pp), respectively. Only half of the models clearly outperformed the baseline: Muse Spark 1.3 and Gemini 3.5 Flash Lite followed with gains of 15.0 and 14.3 pp, whereas the other five models (Claude Sonnet 5, Grok 4.3, MiMo-V 2.5, DeepSeek V4 Flash Vision, and Nova 2 Lite) stayed within 6 pp of it. *Object recognition and identity matching* proved more challenging than *aggregation* (see Figure 7 in Appendix H), with *object attributes* yielding the lowest scores (22.7% averaged over all models; see Figure 8 in Appendix H).

Across the video tracks, *event and procedural understanding* obtained the highest scores (average 56.5% across all models), while none of the models achieved satisfactory performance on *temporal grounding* tasks. With an average Accuracy of 41.7% over all ten models, performance on the SEGMENT track exceeded the text-only baseline (30.4%) by 11.3 pp on average. The highest overall Accuracy was obtained by Gemini 3.8 Flash (59.4% [53.9, 64.9]), followed by GPT-5.6 Sol (58.3% [53.2, 63.1]) and Gemini 3 Flash (49.1% [45.3, 52.5]), improving on the baseline by 29.0, 27.9, and 18.7 pp, respectively. *Temporal grounding* was the most difficult capability (see Figure 7 in Appendix H), with a mean performance of 26.9% averaged over all models.

On the PROCEDURE track, Accuracy averaged 37.9% over the ten models (text-only baseline: 27.8%). The highest overall Accuracy was obtained by Gemini 3.8 Flash (56.7% [52.4, 62.0]), followed by GPT-5.6 Sol (51.0% [46.7, 55.1]) and Muse Spark 1.3 (43.8% [38.0, 49.3]). These three models improved on the baseline by 28.9, 23.2, and 16.0 pp, respectively, whereas DeepSeek V4 Flash Vision, MiMo-V 2.5, and Nova 2 Lite scored below it. *Temporal grounding* dropped to 12.6% averaged over all models, less than half of its SEGMENT value.

On the same 300 questions as the human raters, the best model on each track scored 12.0, 12.1, and 6.9 pp lower on the FRAME, SEGMENT, and PROCEDURE track, respectively (Table 4 in Appendix G).

6 DISCUSSION

HeiCo-FOCUS uniquely evaluates long-context video understanding in a real-world, safety-critical setting. Unlike existing datasets that focus on short clips or isolated perception tasks, it requires models to maintain consistent representations of multiple objects over procedures lasting up to hours, where errors accumulate and cannot be corrected from local context. By anchoring evaluation in the concrete task of foreign object tracking and retrieval - an objective, verifiable, and clinically critical problem - the benchmark directly links technical performance to real-world utility.

Importantly, our human baseline highlighted the difficulty of the long-context reasoning targeted by HeiCo-FOCUS. Even the human raters reached only 35% on *event aggregation* and 40% on *duration estimation* (20 questions each), both of which require following foreign objects over long time spans. This observation retrospectively supports our decision to establish persistent object identities through a dedicated, quality-controlled annotation process rather than asking clinicians to formulate and answer long-context questions directly from the videos. By grounding question generation in these curated instance trajectories, we reduce errors arising from ad-hoc manual reconstruction of object histories while retaining clinical expertise for question types that require higher-level medical reasoning.

To our knowledge, HeiCo-FOCUS is the first surgical VQA benchmark to include questions spanning full procedures exceeding 20 minutes while explicitly evaluating persistent object tracking, temporal consistency, and long-horizon aggregation in surgery. However, constructing such a benchmark posed several challenges. First, defining consistent annotation rules for different foreign objects required careful clinical and technical consideration. For example, certain object types, such as surgical staples, can be too small and numerous to be reliably annotated individually by human annotators. Similarly, object-specific decisions, such as how to handle suture threads attached to needles, required detailed annotation guidelines, as documented in the appendix. Second, video artifacts, as well as limited temporal and spatial resolution, posed challenges even for expert annotators (see Appendix I). Third, designing a capability taxonomy that was simultaneously clinically meaningful, technically informative, and sufficiently disentangled for systematic evaluation required extensive iterative refinement between clinical and technical experts. We addressed many of these challenges through a rigorous transdisciplinary and quality-controlled annotation process involving crowd workers, medical students, surgeons, and surgical AI researchers. Nevertheless, we acknowledge that some annotation and taxonomy decisions may remain debatable. To promote transparency and continuous improvement, we therefore include a community feedback process that allows users to report potentially problematic or ambiguous cases.

Several results deserve further discussion. Accuracy did not follow price: the most accurate model, Gemini 3.8 Flash, cost about a fifth of GPT-5.6 Sol per question on the video tracks (Figure 9). While we captured statistical uncertainty via bootstrapping, we did not systematically investigate prompt sensitivity or variability across repeated model runs. Note that, in this context, evaluating long-context VLMs on extended surgical videos incurs substantial monetary and environmental costs (see Figure 9), highlighting the need to prioritize efficient evaluation protocols. Finally, the dataset was derived from a limited number of colorectal procedures, which may introduce biases related to surgical style, instrumentation, patient population, or recording conditions. Consequently, benchmark performance should not be interpreted as evidence of safe generalization across procedures, institutions, or clinical settings. Future work should expand the diversity of procedures, institutions, and surgical settings represented in the benchmark, while placing particular emphasis on increasing the number and diversity of expert-generated reasoning questions.

In conclusion, through its clinically grounded design, capability-centric evaluation, and rigorous multi-stage annotation process, HeiCo-FOCUS establishes a fundamentally new testbed for assessing whether modern VLMs can move beyond short-term pattern recognition toward reliable long-horizon reasoning. By exposing failures in cumulative temporal consistency, it provides a foundation for developing and evaluating models that can operate reliably in complex, real-world settings.

ETHICS STATEMENT

HeiCo-FOCUS re-annotates the publicly released HeiCo corpus, which is fully anonymized. We add no patient data.

REPRODUCIBILITY STATEMENT

The supplementary material contains an anonymized copy of the evaluation package (dataset loaders, answer-format parsing, rule-based scoring and the LLM judge). The question–answer pairs are released through Hugging Face (Appendix B; link withheld for anonymity). Model prompts are given verbatim in Appendix E, the scoring rules in Appendix E.4, and the judge setup in Appendix F.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic data, implement methods, and clean and reformat the dataset (the text-only filter and the answer judge are described in Section 4.3 and Appendix F). We have not used generative AI tools to propose or refine hypotheses, design or give feedback on the research methodology or experiments, support qualitative or thematic data analysis, or interpret results; developing theoretical models or conceptual frameworks, formulating or proving mathematical claims, and translation are not applicable to this work. Additionally, we used generative AI tools to create figures, write and edit code, edit this paper for readability, format references, and discover research topics or identify gaps. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Fardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage the illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

Barzin Badiee, Saad Mallick, Konmal Ali, Melissa Justo, Sona Mahrokhi, and Peyman Benharash. Retained foreign bodies after major operations: Trends, risk factors, and associated outcomes. *Surgery*, 185:109513, 2025.

Long Bai, Mobarakol Islam, Lalithkumar Seenivasan, and Hongliang Ren. Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. *arXiv preprint arXiv:2305.11692*, 2023.

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

Matthias Carstens, Shubha Vasisht, Zheyuan Zhang, Iulia Barbur, Annika Reinke, Lena Maier-Hein, Daniel A. Hashimoto, and Fiona R. Kolbinger. Artificial intelligence for surgical scene understanding: a systematic review and reporting quality meta-analysis. *npj Digital Medicine*, 9(1), 2025. doi: 10.1038/s41746-025-02227-4.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

S. Duca, O. Bălă, N. Al-Hajjar, C. Iancu, I. C. Puia, D. Munteanu, and F. Graur. Laparoscopic cholecystectomy: incidents and complications. A retrospective analysis of 9542 consecutive laparoscopic operations. *HPB*, 5(3):152–158, 2003. doi: 10.1080/13651820310015293.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

Alexander D Gibson, Nicole M White, Gary S Collins, and Adrian G Barnett. Evidence of unreliable data and poor data provenance in clinical prediction model research and clinical practice. *BMC Medicine*, 24(1), 2026. doi: 10.1186/s12916-026-04981-y.

Runlong He, Mengya Xu, Adrito Das, Danyal Z Khan, Sophia Bano, Hani J Marcus, Danail Stoyanov, Matthew J Clarkson, and Mobarakol Islam. Pitvqa: Image-grounded text embedding llm for visual question answering in pituitary surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 488–498. Springer, 2024.

Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13480–13492, 2023.

Quentin Lhoest, Albert Villanova del Moral, Patrick von Platen, Thomas Wolf, Mario Šaško, Yacine Jernite, Abhishek Thakur, Lewis Tunstall, Suraj Patil, Mariama Drame, Julien Chaumond, Julien Plu, Joe Davison, Simon Brandeis, Victor Sanh, Teven Le Scao, Kevin Canwen Xu, Nicolas Patry, Steven Liu, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Nathan Raw, Sylvain Lesage, Anton Lozhkov, Matthew Carrigan, Théo Matussière, Leandro von Werra, Lysandre Debut, Stas Bekman, and Clément Delangue. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184. Association for Computational Linguistics, November 2021. URL https://aclanthology.org/2021.emnlp-demo.21.

Yaoqian Li, Xikai Yang, Dunyuan Xu, Yang YU, Litao Zhao, Xiaowei Hu, Jinpeng Li, and Pheng-Ann Heng. Surgpub-video: A comprehensive surgical video framework for enhanced surgical intelligence in vision-language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6628–6635, 2026.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, et al. Heidelberg colorectal data set for surgical data science in the sensor operating room. *Scientific data*, 8(1): 101, 2021.

Karttikeya Mangalam, Raiymbek Akshkulakov, and Jitendra Malik. Egoschema: a diagnostic benchmark for very long-form video language understanding. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Leon Mayer, Piotr Kalinowski, Caroline Ebersbach, Marcel Knopp, Tim Rädsch, Evangelia Christodoulou, Annika Reinke, Fiona R Kolbinger, and Lena Maier-Hein. 6 fingers, 1 kidney: Natural adversarial medical images reveal critical weaknesses of vision-language models. *arXiv preprint arXiv:2512.04238*, 2025.

Alejandra Perez, Chinedu Nwoye, Ramtin Raji Kermani, Omid Mohareri, and Muhammad Abdullah Jamal. Surglavi: Large-scale hierarchical dataset for surgical vision–language representation learning. *Medical Image Analysis*, pp. 103982, 2026.

Annika Reinke, Evangelia Christodoulou, Sthuthi Sadananda, A. Emre Kavur, Khrystyna Faryna, Daan Schouten, Bennett A. Landman, Carole Sudre, Olivier Colliot, Nick Heller, et al. Medical imaging AI competitions lack fairness. *arXiv preprint arXiv:2512.17581*, 2026.

Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. Surgical-vqa: Visual question answering in surgical scenes using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 33–43. Springer, 2022.

Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. Surgicalgpt: end-to-end language-vision gpt for visual question answering in surgery. In *International conference on medical image computing and computer-assisted intervention*, pp. 281–290. Springer, 2023.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Sarah E. Tevis and Gregory D. Kennedy. Postoperative complications: Looking forward to a safer future. *Clinics in Colon and Rectal Surgery*, 29(3):246–252, 2016. doi: 10.1055/s-0036-1584501.

Guankun Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, et al. Endochat: Grounded multimodal large language model for endoscopic surgery. *Medical Image Analysis*, pp. 103789, 2025a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025b.

S. Weprin, F. Crocerossa, D. Meyer, K. Maddra, D. Valancy, R. Osardu, H. S. Kang, R. H. Moore, U. Carbonara, F. J. Kim, and R. Autorino. Risk factors and preventive strategies for unintentionally retained surgical sharps: a systematic review. *Patient Safety in Surgery*, 15(1):24, 2021. doi: 10.1186/s13037-021-00297-3.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. Moviebench: A hierarchical movie level dataset for long video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28984–28994, 2025.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9127–9134, 2019.

Kun Yuan, Manasi Kattel, Joël L Lavanchy, Nassir Navab, Vinkle Srivastav, and Nicolas Padoy. Advancing surgical vqa with scene graph knowledge. *International journal of computer assisted radiology and surgery*, 19(7):1409–1417, 2024.

Zhitao Zeng, Zhu Zhuo, Xiaojun Jia, Erli Zhang, Junde Wu, Jiaan Zhang, Yuxuan Wang, Chang Han Low, Jian Jiang, Zilong Zheng, et al. Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. *arXiv preprint arXiv:2506.02555*, 2025.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18891–18901, 2025.

A Comparison with Existing Surgical VQA Datasets

Table 1: Overview of publicly released surgical visual question answering (VQA) datasets. Class labels refers to VQAs generated from classification labels extracted from available vision datasets. *Full dataset not released yet.

Dataset	#Videos	#VQAs	Context Length			Source
			Frame	Segment (30 sec≤ t≤5 min)	Procedure t>5 min	
Cholec80-VQA (Seenivasan et al., 2022)	40	43,182	43,182	0	0	Class. labels
EndoVis-18-VQA (Seenivasan et al., 2022)	14	11,783	11,783	0	0	Class. labels
EndoVis-VQLA (Bai et al., 2023)	24	12,255	12,255	0	0	Class. labels
PSI-AVA-VQA (Seenivasan et al., 2023)	8	10,291	10,291	0	0	Class. labels
PiTVQA (He et al., 2024)	25	884,242	884,242	0	0	Class. labels
SSG-VQA (Yuan et al., 2024)	50	960,000	960,000	0	0	Class. labels
Surg-396K (Wang et al., 2025a)	109	396,000	396,000	0	0	Class. labels
SurgVLM-DB (Zeng et al., 2025)	849	7,798,400	7,798,400	0	0	Class. labels
SurgPub-Video (Li et al., 2026)	3,538	48,520	0	48,520	0	Audio transcripts
Surg-LaVI (Perez et al., 2026)	5,317	240,000	77*	117,061*	320*	Audio transcripts
Ours	**30**	**30,000**	**12,000**	**12,000**	**6,000**	**Expert annotations**

B Data Access and Usage

The dataset is hosted through Hugging Face datasets (Lhoest et al., 2021), and publicly available at *[link removed for anonymity]*. We stay compliant with the datasets API and a single call allows access to all question answer pairs. For seamless integration with the loading and processing of videos for each of the tracks, we also provide a Python utilities package *[link removed for anonymity, provided in the supplementary material instead]*. The usage of the package is shown in Figure 5, with more involved data loading classes and additional examples provided in the repository.

```python
from focus import (FocusDataset, DatasetSplit, Track,
                   FocusConfig, set_config, download,
                   Evaluator, Response)

# configure local data root and download video files
set_config(FocusConfig(root_dir="/data/focus"))
download("heico")

# QA annotations are fetched automatically from HuggingFace
ds = FocusDataset("heico", DatasetSplit.TEST, Track.SEGMENT)

responses = [
    Response(qID=req.qID, content=model(req))
    for req, _ in ds
]

results, summary = Evaluator().run(
    requests=ds.requests,
    references=ds.references,
    responses=responses,
)
```

Figure 5: Loading and evaluating a model on HeiCo-FOCUS using the accompanying Python package. Video files are downloaded once via `download()`, and question–answer annotations are fetched automatically from Hugging Face. The `Evaluator` handles answer-format parsing and routes open-ended and multiple-choice answers to an LLM judge.

C TAXONOMY

Object recognition and identity matching: Recognition of object instances, their semantic type, attributes, spatial context, and identity consistency across time. Subcategories:

- Object identification: semantic classification of visible object instances
- Object instance identity matching: correspondence of object instances across time
- Object attributes and state: recognition of observable object properties or handling states
- Object spatial localization (camera): object localization relative to the image plane
- Object spatial localization (situs): object localization relative to anatomical structures

Temporal grounding: Localization of events or object occurrences within the video timeline, including their temporal position and duration. Subcategories:

- Temporal localization: identification of the time point or interval at which an event or object occurrence takes place
- Duration estimation: estimation of the temporal extent for which an event or object occurrence persists

Aggregation: Combination of information on foreign objects and events across multiple objects, instances, and/or time points. Subcategories:

- Object aggregation: aggregation of information across multiple foreign object instances and/or categories
- Event aggregation: aggregation of events that involve foreign objects across time

Event and procedural understanding: Recognition and interpretation of actions, events, and their procedural structure over time. Subcategories:

- Foreign object interaction recognition: identification of actions or manipulations involving a foreign object
- Foreign object usage purpose: identification of the general procedural role of a foreign object at a given time. Answers are grounded in what is visually demonstrated; no inference about hypothetical outcomes or hidden intent
- Temporal ordering: determination of the relative order of events or procedural steps, including the identification of the n-th event in a series of events

Complex reasoning: Inference of functional, causal, relational, or outcome-related information beyond direct observation, and reasoning over interactions, dependencies, comparative patterns, and temporally extended evidence across objects and events. Subcategories:

- Functional reasoning: inference of the intended function or role of an object or action beyond what is directly shown
- Causal and consequence reasoning: inference of cause–effect relationships between actions, events, or object states and inference of potential or actual outcomes resulting from actions or events
- Multi-step compositional reasoning: inference that requires the execution of a sequence of interdependent reasoning steps across objects, space, and time, involving the decomposition of a query into multiple sub-tasks to derive an answer that cannot be attributed to a single dominant capability

See Figure 2 for examples.

D VQA GENERATION PROCESS

Table 2: Overview of the six-stage annotation pipeline used to construct the HeiCo-FOCUS benchmark. Raw surgical videos are progressively transformed into clinically grounded visual question answering (VQA) pairs through foreign object screening, frame-level localization, automated refinement, instance-level tracking with anchor moment identification, and multi-stage question generation and verification. The table summarizes inputs, outputs, annotation scale across 30 videos, and the combination of crowd, expert, and automated methods used at each stage.

Stage	**Purpose**	**Input**	**Output**	**#annotations (total across 30 videos)**	**Annotators / Method**
Stage 1 Foreign object presence screening	Identification of video segments with foreign objects	5 s long videos	Presence flag for each class of foreign objects on each clip	69,469 annotated 5 s videos; 16% contained a foreign object	Quality Match GmbH, Häusser-straße 36, 69115 Heidelberg, Germany
Stage 2 Frame-level localization (bounding boxes at 1 fps)	Identification and localization of foreign objects in individual frames	Frames (sampled at 1 fps) corresponding to clips with foreign objects	Frames (sampled at 1 fps) with bounding boxes and class assignment for all foreign objects	53,001 bounding boxes on a total of 54,528 frames	Quality Match GmbH, Häusser-straße 36, 69115 Heidelberg, Germany; iMerit, 160 West Santa Clara St., Suite 600, San Jose, CA 95113, USA
Stage 3 Automated bounding box completion / correction	Automatic refinement of crowd-based foreign object identification and localization	Frames (sampled at 1 fps) with bounding boxes and class assignment for all foreign objects	Automatically corrected frames (sampled at 1 fps) with bounding boxes and class assignment for all foreign objects	4,832 corrected frames	Automated method

Stage	Purpose	Input	Output	#annotations (total across 30 videos)	Annotators / Method
Stage 4 Instance assignment	Instance assignment to individual objects (clips only first three frames), final correction of bounding boxes, and class assignments.	Frames (sampled at 1 fps) with bounding boxes and class assignment for all foreign objects	Frames (sampled at 1 fps) with class and instance assignments and bounding boxes around all foreign objects	79,682 bounding boxes (instance identities for all classes except clips); 324,273 curated frames	18 Medical students
Stage 5 VQA generation and clinical question design	Generation of questions and answers based on foreign object information (auto-generated + expert-added)	Frames (sampled at 1 fps) with class and instance assignments, bounding boxes around all foreign objects, and anchor moments	VQA pairs balanced across the 5 core capabilities, including 15 sub-capabilities	30,000 VQA pairs balanced across the 5 capability groups. FRAME track: 12,000 VQA pairs SEGMENT track: 12,000 VQA pairs PROCEDURE track: 6,000 VQA pairs	18 Medical students and 7 expert surgeons; automatic method based on instance information and expert annotations
Stage 6 End-to-end VQA verification	Verification of the question generation pipeline	A spreadsheet (via link) containing question IDs, corresponding timestamps, and the associated surgical procedure video	The same spreadsheet completed with answers, a confidence score (1–10), optional comments, and, where necessary, a "Needs revision" flag	500 verified questions	14 Surgical AI researchers + 3 Medical students + 5 expert surgeons

Labeling instructions were provided separately for all stages that required human involvement.

Figure 6 shows the Stage 4 annotations of one video per procedure type.

E Model Prompts

We provide details on the prompt design used for all experiments. Each prompt is composed of four components: (i) a context prefix describing the surgical setting, (ii) a foreign object definition (see Appendix E.2), (iii) the task-specific question, and (iv) format specifications that enforce structured outputs. Appendix E.5 shows a complete prompt for one SEGMENT question.

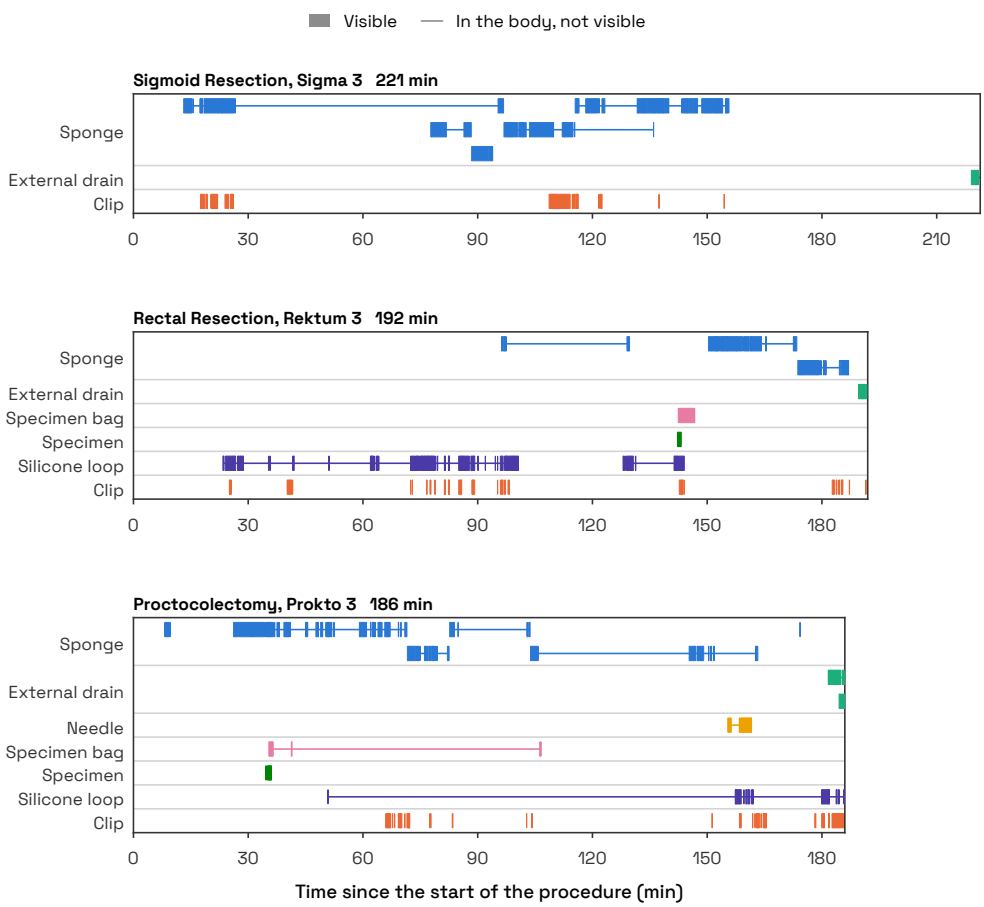


Figure 6: **Foreign-object timelines from Stage 4.** One video per procedure type on a common time axis. Bars mark the seconds in which an instance is visible; thin lines join its first and last sighting. Instances of one class share a row when they are at least three minutes apart. Clips carry no instance identities and form one visibility row.

E.1 Context Prefix

FRAME

> "You are assisting in the analysis of laparoscopic videos from minimally invasive surgeries. Specifically, you see a frame from a {procedure_type} at timepoint {timestamp}. Analyze the footage and answer the question based on the visual evidence. Be as concise as possible by following the output format instructions."

SEGMENT

> "You are assisting in the analysis of laparoscopic videos from minimally invasive surgeries. Specifically, you see a video from a {procedure_type} from timepoint {timestamp_start} to {timestamp_end}. The video is sampled at 1 fps. Analyze the footage and answer the question based on the visual evidence. Be precise and concise."

PROCEDURE

> "You are assisting in the analysis of laparoscopic videos from minimally invasive surgeries. Specifically, you see a video from a {procedure_type} from timepoint

{timestamp_start} to {timestamp_end}. Analyze the footage and answer the question based on the visual evidence. Be precise and concise."

E.2 Foreign Object Definitions

```
=== FOREIGN OBJECT DEFINITION ===
```

A foreign object is any object fully introduced into the patient's body cavity during surgery that must be retrieved or accounted for. Importantly, standard surgical instruments that remain connected to the external environment (e.g., graspers, scissors, trocars, staplers, cameras) are not considered foreign objects. Furthermore, we exclude detachable parts of surgical instruments, particularly anvil components of staplers.

```
=== FOREIGN OBJECT CLASSES ===
```

The following are the only classes to consider. Do not infer or name any class outside this list, such as staples.

- **Sponge**: A soft, absorbent material used to soak up fluids. They are typically white when fresh and can become reddish-brown when saturated with blood.
- **External drain**: A clear or fluid-filled tube used to evacuate fluids from the surgical site to the outside of the body. Its tip may be temporarily visible within the surgical field during placement or adjustment.
- **Needle**: A sharp, pointed metal instrument, straight or curved, used for placing sutures. A needle is considered visible only if the metallic needle itself is seen. The suture thread alone does not qualify as needle visibility.
- **Clip**: A small metal or polymer device used to seal vessels or ducts. May potentially remain in the body. Clips only count as foreign objects once placed in the abdomen. They do not count as foreign objects while loaded within the clip applier instrument.
- **Specimen bag**: A sterile pouch used to collect and retrieve resected tissue or organs from the body cavity. Only consider the pouch itself as a foreign object and ignore the string attached to it.
- **Silicone loop**: A soft and flexible, typically white band used to encircle and control blood vessels or structures for isolation and traction.
- **Specimen**: Excised biological tissue (e.g., appendix, resected bowel segment) that must be retrieved from the body cavity before completion of the procedure. Tissue is considered a foreign object once it is fully detached from surrounding anatomy. This excludes fat and blood. Specimens contained within a specimen bag are considered retrieved; specimen bags are annotated separately.
- **Absorbable hemostatic agent**: A soft, absorbable material (for example oxidized cellulose or a gelatin sponge) applied to a bleeding surface to promote clotting. It may be intentionally left inside the body cavity.

```
=== TEMPORAL AND VISIBILITY RULES ===
```

- A foreign object is considered inserted / introduced at the moment it first becomes partially visible inside the body cavity. In case of clips, it is also required to leave the clip applier. Class-specific visibility exceptions (e.g., needle string) take precedence.
- A foreign object is considered present as long as it is inside the body cavity, even during periods of occlusion or when temporarily off-screen.
- A foreign object is considered removed / retrieved at the moment it fully exits the body cavity or is fully inserted into a specimen bag. A specimen fully inside a specimen bag is considered retrieved; the specimen bag itself remains present until it exits the body cavity.
- If the same physical object reappears after occlusion, it is the same instance — do not count it again. When multiple instances of the same class are present, reappearing objects should be attributed to already-known instances before assuming a new one.
- All time references are given relative to the start of the intervention (00:00:00).

E.3 Format Instructions

The following output type-specific instructions are added to the prompt:

- **Time point**: Return the answer only in hh:mm:ss format without any additional text and with respect to the start of the procedure (not necessarily the start of the video clip). Example: 01:37:12
- **Duration**: Return the answer only in hh:mm:ss format without any additional text. Example: 00:02:13
- **Foreign object class**: Return only labels from the foreign objects list. Example: Sponge
- **Number**: Return only a single integer value. Example: 3
- **Percentage**: Return only a single percentage value between 0 and 100 without the percent sign. Example: 75
- **Binary**: Return only yes or no. Example: yes
- **Multiple Choice**: Return only one of the provided options.
- **Open Ended**: Answer the question directly using only the most relevant details. Do not add lengthy explanations or rephrasings of the question.
- **All (postfix)**: Strictly adhere to these output format specifications. Do not add any extra text.

E.4 Scoring

Each answer is parsed according to its expected format and scored with the rule in Table 3.

Table 3: Scoring rule per answer format. d is the length of the input window in seconds. Answers that do not parse in the expected format count as incorrect.

Format	Counted as correct if
Number	the integer equals the reference
Yes/no	the answer equals the reference (case-insensitive)
Foreign-object class	the set of named classes equals the reference set (order- and case-insensitive; "none" only on its own)
Percentage	within ± 5 pp of the reference
Time point, duration	within $\tau = \min(5,\, 1 + 4d/360)$ s of the reference: 1.3, 2.3, and 4.3 s for 30 s, 2 min, and 5 min windows; 5 s for windows of 6 min or longer
Open-ended, multiple choice, matching	judged semantically equivalent to the reference by gpt-oss-120b; answers longer than 300 characters count as incorrect

E.5 Example Prompt

```
You are assisting in the analysis of laparoscopic
videos from minimally invasive surgeries.
Specifically, you see a video from a Sigmoid Resection
from timepoint 00:14:40 to 00:19:21.  The video is
sampled at 1 fps.

Analyze the footage and answer the question based on
the visual evidence.  Be precise and concise.

=== FOREIGN OBJECT DEFINITION ===
A foreign object is any object fully introduced into
the patient's body cavity during surgery that must be
retrieved or accounted for.  Importantly, standard
surgical instruments that remain connected to the
external environment (e.g., graspers, scissors,
```

```
trocars, staplers, cameras) are not considered foreign
objects.  Furthermore, we exclude detachable parts of
surgical instruments, particularly anvil components of
staplers.
=== FOREIGN OBJECT CLASSES ===
The following are the only classes to consider.  Do
not infer or name any class outside this list, such as
staples.
Sponge:  A soft, absorbent material used to soak up
fluids.  They are typically white when fresh and can
become reddish-brown when saturated with blood.
External drain:  A clear or fluid-filled tube used
to evacuate fluids from the surgical site to the
outside of the body.  Its tip may be temporarily
visible within the surgical field during placement
or adjustment.
Needle:  A sharp, pointed metal instrument, straight
or curved, used for placing sutures.  A needle is
considered visible only if the metallic needle itself
is seen.  The suture thread alone does not qualify as
needle visibility.
Clip:  A small metal or polymer device used to seal
vessels or ducts.  May potentially remain in the body.
Clips only count as foreign objects once placed in the
abdomen.  They do not count as foreign objects while
loaded within the clip applier instrument.
Specimen bag:  A sterile pouch used to collect and
retrieve resected tissue or organs from the body
cavity.  Only consider the pouch itself as foreign
object and ignore the string attached to it.
Silicone loop:  A soft and flexible, typically white
band used to encircle and control blood vessels or
structures for isolation and traction.
Specimen:  Excised biological tissue (e.g., appendix,
resected bowel segment) that must be retrieved from
the body cavity before completion of the procedure.
Tissue is considered a foreign object once it is fully
detached from surrounding anatomy.  This excludes fat
and blood.  Specimens contained within a specimen bag
are considered retrieved; specimen bags are annotated
separately.
Absorbable hemostatic agent:  A soft, absorbable
material (for example oxidized cellulose or a gelatin
sponge) applied to a bleeding surface to promote
clotting.  It may be intentionally left inside the
body cavity.
=== TEMPORAL AND VISIBILITY RULES ===
A foreign object is considered inserted / introduced
at the moment it first becomes partially visible
inside the body cavity.  In case of clips, it is also
required to leave the clip applier.  Class-specific
visibility exceptions (e.g., needle string) take
precedence.
A foreign object is considered present as long as
it is inside the body cavity, even during periods of
occlusion or when temporarily off-screen.
```

```
A foreign object is considered removed / retrieved at
the moment it fully exits the body cavity or is fully
inserted into a specimen bag.  A specimen fully inside
a specimen bag is considered retrieved; the specimen
bag itself remains present until it exits the body
cavity.
If the same physical object reappears after occlusion,
it is the same instance - do not count it again.
When multiple instances of the same class are
present, reappearing objects should be attributed to
already-known instances before assuming a new one.
All time references are given relative to the start of
the intervention (00:00:00).

After the Sponge was inserted in the abdomen, which
different foreign object classes were inserted or
created (in case of a specimen) afterwards, if any?

Please provide the class name(s) or answer none.
Return only labels from the foreign objects list
above.  Example:  Sponge.  Strictly adhere to these
output format specifications.  Do not add any extra
text.
```

F VALIDATION OF THE LLM JUDGE

Open-ended, multiple-choice, and matching answers are graded by gpt-oss-120b. To test the judge against human grading, we drew 164 open-ended answers (76 FRAME, 76 SEGMENT, and 12 PROCEDURE), stratified by track and by the verdict of a judge model (Llama 4 Maverick) , with equal numbers of accepted and rejected answers. Two raters, a surgical resident and a surgical data science researcher, graded each answer independently. Like the judge, they did not see the judge's verdict. Weighted to the population the answers were drawn from, gpt-oss-120b agrees with the resident on 90.7% of answers ($\kappa = 0.74$) and with the second rater on 92.7% ($\kappa = 0.78$), while the two raters agree with each other on 91.8% ($\kappa = 0.78$). The judge is the stricter grader in 13 of its 15 disagreements with the resident and in 7 of 13 with the second rater, so grading by the raters would, if anything, raise the reported accuracies.

G HUMAN BASELINE

Table 4: Accuracy (%) of human raters and models on the same questions. Three raters (two surgical residents and one medical student) answered 300 test questions in total. Their answers were graded identically to the models.

	FRAME	SEGMENT	PROCEDURE
Human raters	**62.0**	**69.8**	**57.4**
Gemini 3.8 Flash	46.0	57.7	50.5
GPT-5.6 Sol	40.0	55.0	45.5
Gemini 3 Flash	50.0	49.0	39.6
Meta Muse Spark 1.3	30.0	49.0	42.6
Gemini 3.5 Flash Lite	44.0	46.3	42.6
SpaceXAI Grok-4.3	28.0	38.9	33.7
Claude Sonnet 5	38.0	38.3	29.7
DeepSeek V4 Flash Vision	34.0	28.9	16.8
Xiaomi MiMo-V 2.5	24.0	28.2	18.8
Amazon Nova 2 Lite	12.0	26.2	15.8
Text-only baseline	24.0	35.6	25.7

H FURTHER RESULTS

Figure 8 gives Accuracy per sub-capability, and Figure 9 plots Accuracy against the API cost of each run.

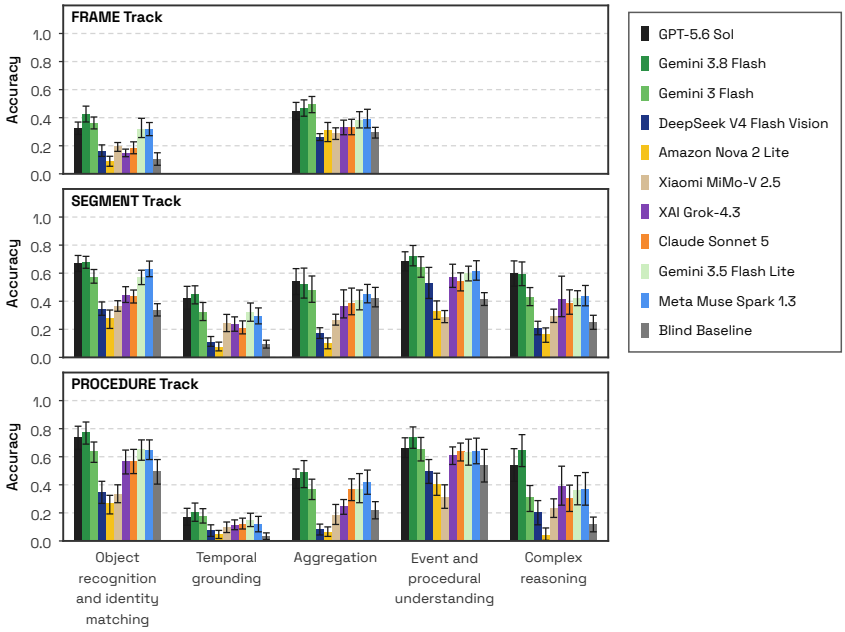


Figure 7: **Core results of the FRAME, SEGMENT, and PROCEDURE tracks.** Performance of models stratified by capability group. Points are means of per-video Accuracy over the test videos; 95% CIs from a two-level bootstrap over videos and questions (1,000 resamples).



Figure 8: **Accuracy per sub-capability.** Mean of per-video Accuracy (%) for each model and sub-capability; blank cells have no questions of that sub-capability in the track. Blind Baseline is the text-only baseline.

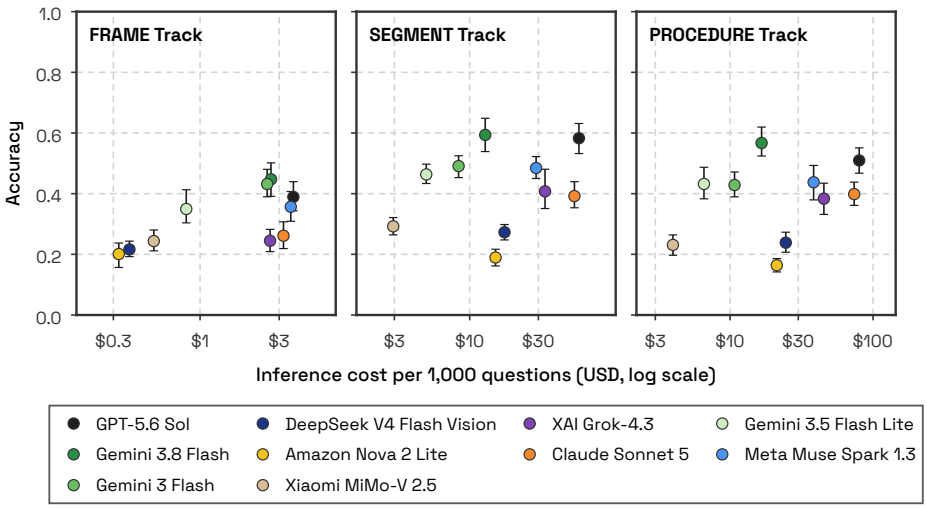
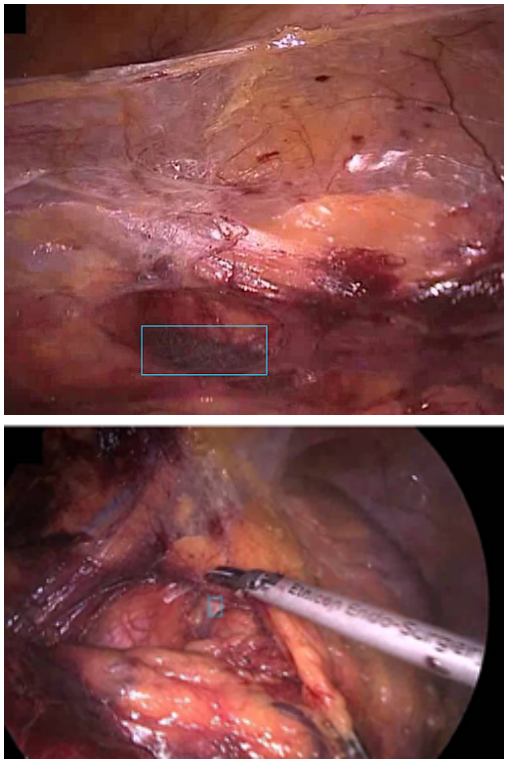
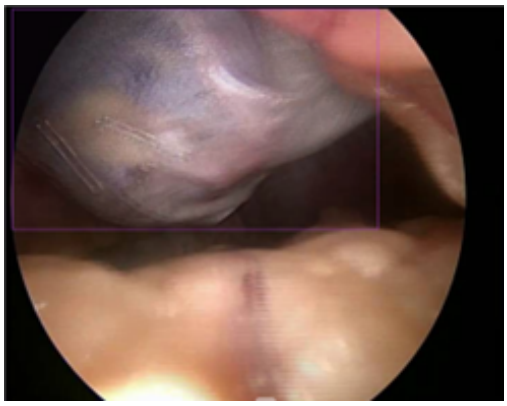
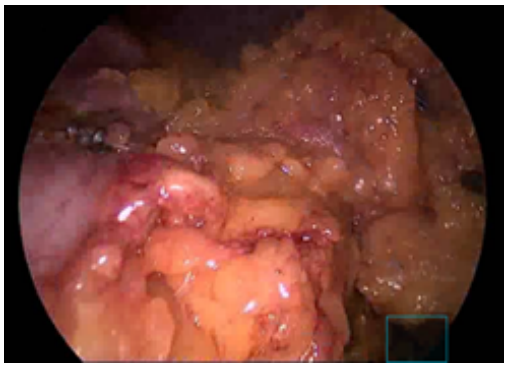
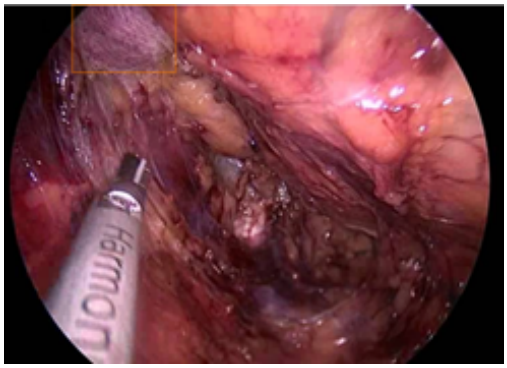
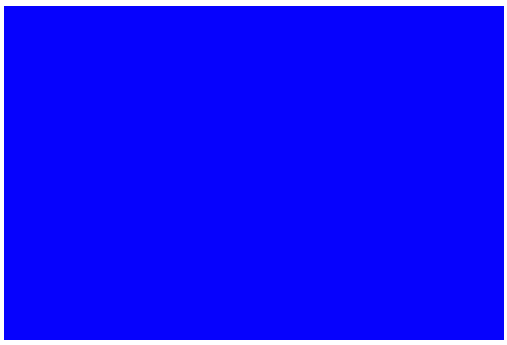


Figure 9: **Accuracy and inference cost.** Accuracy with 95% CIs, as in Figure 4, against the API cost per 1,000 test questions (log scale), for the models that see the video. Cost is the sum of the API charges for the answers, without judge calls.

I CHALLENGING EXAMPLES

Table 5: Sample images illustrating some of the annotation challenges the annotators faced when generating the instance annotations as the basis for the visual question answering (VQA) generation pipeline.

Annotation issue/challenge	**Sample image**
Sponges tucked deep into the surgical site are challenging to identify and link.	
Staples can be confused with clips and are infeasible to count.	

Annotation issue/challenge	Sample image
Specimen and clips attached to them should not be annotated as foreign objects as soon as they are placed in the specimen bag.	
Due to inadequate lighting at the image borders, distinguishing between shadows and blood-soaked sponges is challenging.	
For several sponges, it was difficult to decide whether they had been removed from the surgical site or remained in situ only to reappear later, particularly when the video was interrupted by a prolonged blue screen.	 

Annotation issue/challenge	Sample image
Due to the dense local arrangement of clips, perspective shifts sometimes led to ambiguities in identity association, making it challenging to distinguish one clip from another.	