ORENA – SAVE FOCUS CHALLENGE- PROCEDURE TRACK

The PROCEDURE Track challenges models with long videos (full procedures from beginning up to a certain point reflecting real intraoperative and postoperative query scenarios). Questions may require recalling when an object was last seen, aggregating object counts across long intervals, or determining whether all introduced foreign objects requiring retrieval have been removed. This track targets the key technical frontier of long-context VLMs and directly reflects real-world intraoperative quality assurance needs, where critical safety questions depend on understanding events that occurred much earlier in the operation.

Title

Use the title to convey the essential information on the challenge mission.

ORena - SAVE FOCUS challenge: Foreign Object Contextual Understanding for Safe Surgical AI

PROCEDURE Track

Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

ORena – FOCUS PROCEDURE Track

CHALLENGE ORGANIZATION

Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Recent progress in general-domain Vision Language Models (VLMs) has enabled increasingly strong temporal reasoning over extended video streams. However, the surgical AI community has so far lacked a dedicated, standardized challenge that evaluates whether these emerging capabilities translate to real clinical workflows, where critical events unfold over tens of minutes to hours. This gap is important: many clinically meaningful questions in minimally invasive surgery require persistent memory, temporal consistency, and reasoning across long time horizons.

The ORena SAVE FOCUS challenge addresses this unmet need by providing a structured benchmark that targets an urgent patient-safety problem in minimally invasive procedures: ensuring the retrieval of foreign objects, such as sponges and needles, from the abdomen at the end of the operation. Unintentionally leaving foreign objects in the abdomen is a rare but consequential incident in minimally invasive surgery, as they can cause serious complications.

Specifically, FOCUS aims to generate scientific progress by tackling two fundamental research questions (RQs):

RQ1 (Clinical utility): Can VLMs generate clinically meaningful and safety-relevant information about surgical foreign objects? This question targets the clinical value of VLMs for intraoperative quality assurance, focusing on actionable insights related to foreign objects such as sponges, needles, and clips.

RQ2 (Technical limits): What are the current limitations of VLMs in surgical scene reasoning? To answer this question, FOCUS encompasses three tracks that progressively increase the temporal and contextual demands on the model: a FRAME Track (single-image understanding) to assess foundational visual perception and surgical-domain interpretation; a SEGMENT Track (short video segments) to evaluate short-term temporal reasoning, local tracking, and action understanding; and a PROCEDURE Track (long-context up to full procedures reflecting real intraoperative and postoperative query scenarios) to probe long-horizon memory, persistent object tracking across occlusions and scene

changes, aggregation over time (e.g., counting and retrieval status), and global reasoning across events. Together, these tracks enable a systematic characterization of where current VLMs succeed and fail as task complexity transitions from instantaneous perception to long-context intraoperative reasoning.

Critically, FOCUS is enabled by a unique dataset that, to the best of our knowledge, is the first challenge resource to provide full-length laparoscopic videos with fine-grained foreign-object annotations at scale. Importantly, the multi-center dataset includes instance-consistent labels, making it possible to evaluate models not only on short-term detection, but also on long-horizon tracking, counting, and retrieval verification across extended procedures. With a total of over 100,000 Visual Question Answering (VQA) pairs obtained from 400 surgical videos acquired from all over the world, FOCUS establishes a standardized benchmark that simultaneously (i) probes the technical limits of long-context VLMs in real-world surgical video and (ii) addresses a concrete quality-assurance objective with direct relevance to intraoperative patient safety.

FOCUS is hosted within ORena, a new umbrella framework for surgical AI competitions inspired by the “Arena” paradigm, but adapted to the specific constraints and opportunities of the Operating Room (OR).

Keywords

List the primary keywords that characterize the challenge. (Separate your inputs with comma like Keyword 1, Keyword 2)

Surgical AI, Surgical Data Science, Vision-Language Model, Long-context Reasoning, Surgical Foreign Objects, Laparoscopy, Endoscopy, Surgical Video Understanding

Year

Please indicate the year of the challenge. If you are applying for next year’s conference, please write the year of that conference.

2026

Novelty of the challenge

Briefly describe the novelty of the challenge.

FOCUS is the first surgical VLM challenge to explicitly benchmark long-context, understanding of full-length laparoscopic procedures, rather than only short, curated segments. It introduces a structured evaluation framework that jointly measures (i) clinically relevant foreign-object safety questions and (ii) technical limits of modern VLMs. By aligning long-context video reasoning with a concrete patient-safety use case FOCUS provides the first standardized testbed that connects emerging long-context VLM capabilities to real intraoperative quality assurance needs.

Task description and application scenarios

Briefly describe the application scenarios for the tasks in the challenge.

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. These objects frequently remain intra-abdominally for extended periods of the procedure, as objects, such as sponges, may contribute to exposure of the surgical field or may be reused. Unintentionally leaving foreign objects in the abdomen is a rare but consequential incident in minimally invasive surgery, as they can cause serious complications [1]. To avoid such incidents, counting of all foreign objects at the beginning and the end of every procedure has been established as a standard quality control measure to ensure that all used objects are removed from the abdomen by the end of the procedure. However, due to the continuously changing and closed-up view in laparoscopic surgery, localizing remaining objects intra-abdominally at the end of surgical procedures can be time-consuming and represents a practical nuisance that surgeons routinely encounter.

The challenge tasks reflect intraoperative quality-assurance scenarios in minimally invasive surgery where surgeons must ensure foreign objects are safely managed. Example applications include continuous counting and tracking of sponges, needles, clips, sutures, and specimen bags, real-time localization of the last known position of an object that is unaccounted for, and verification that all introduced items requiring removal have been retrieved before closure.

FOCUS benchmarks vision–language models on clinically relevant Visual Question Answering (VQA) tasks for foreign object understanding in minimally invasive surgery, with the goal of advancing AI methods that can support intraoperative quality assurance and patient safety. In this context, VQA refers to the task of answering structured, clinically grounded questions based on surgical video content and associated metadata.

In the context of this challenge, a foreign object refers to an object that is fully introduced into the patient's body cavity during a surgical procedure and is expected to be retrieved or accounted for. This includes items such as surgical sponges, compresses, needles, clips, drains, specimen bags, elastic straps, and similar objects that are temporarily placed intra-abdominally.

In addition, isolated biological material that is detached during the procedure and must be retrieved - such as an excised appendix, a gallstone, or resected tissue fragments - are also considered foreign objects under this definition, as they require removal from the body cavity before completion of the operation.

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.

FOCUS is organized in three tasks, here referred to as tracks.

The FRAME Track evaluates a model's ability to answer clinically relevant questions from a single image. This track targets core surgical scene understanding skills such as foreign object identification, attribute recognition, and spatial localization within a single moment in time. It serves as an accessible entry point while establishing a strong baseline for visual perception and surgical-domain image interpretation.

The SEGMENT Track focuses on short video segments (up to 5 min), requiring models to incorporate local temporal context to answer questions about foreign objects and their interactions with anatomy and instruments. Tasks emphasize motion-aware perception, short-term tracking, and understanding of brief action sequences, such as insertion, manipulation, or removal of sponges, needles, or clips. This track bridges static recognition and longer-horizon reasoning by testing whether models can integrate events across time within a bounded context.

The PROCEDURE Track challenges models with long videos (procedures from beginning up to a certain point reflecting real intraoperative and postoperative query scenarios), to assess their capacity for long-term memory, persistent tracking, and global reasoning. Questions may require recalling when an object was last seen, aggregating object counts across long intervals, or determining whether all introduced foreign objects requiring retrieval have been removed. This track targets the key technical frontier of long-context VLMs and directly reflects real-world intraoperative quality assurance needs, where critical safety questions depend on understanding events that occurred much earlier in the operation.

[1]: Badiee, Barzin, et al. "Retained foreign bodies after major operations: Trends, risk factors, and associated outcomes." Surgery 2025

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

If the challenge is part of a workshop, please indicate the workshop.

N/A

Expected number of participants

Please explain the basis of your estimate (e.g., numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

As this challenge is unprecedented, reliable estimates based on prior participation numbers are not available. We therefore estimate 50–100 participating teams based on the following considerations:

1. The challenge targets a timely and high-impact frontier - long-context vision-language understanding in surgical video - which is of strong interest not only to the surgical AI community but also to the broader computer vision and multimodal learning communities.
2. The challenge offers $50k+ in prize money, providing a substantial incentive for participation (unprecedented in the field of surgical AI). Comparable high-incentive initiatives, such as the RSNA Breast Cancer Detection Challenge (> 1,000 participants), demonstrate that significant prize funding can attract large and diverse participation beyond the core medical imaging community [1].

[1]: https://www.kaggle.com/competitions/rsna-breast-cancer-detection/leaderboard

Duration

How long does the challenge take? Possible values: half day, full day, 2 hours, etc.

Half day

Longer duration explanation

In case you selected half or full day, please explain why you need a long slot for your challenge.

- Unprecedented prize money and expected attention: FOCUS offers unusually high prize incentives for MICCAI challenges, which is expected to draw broader participation and visibility, including teams from outside traditional medical imaging. A longer slot ensures we can accommodate the increased interest, highlight top contributions, and provide appropriate recognition and discussion.
- Multiple tracks and audiences: We intentionally include both short-context ("entry") tracks and long-context ("frontier") tracks to be inclusive while still pushing technical limits. This creates more results to present and compare, and it attracts both clinical and technical stakeholders who benefit from dedicated discussion time.

Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

Yes. We aim to submit the challenge to Nature Biomedical Engineering.

MICCAI LNCS proceedings

Indicate if you want to offer MICCAI Springer LNCS proceedings to the participants. Publishing a proceedings volume is optional and at the discretion of each challenge's organizers. At a minimum, organizers must ensure that a description of each participant's submission is publicly available. Organizers who wish to publish MICCAI Springer LNCS proceedings must adhere to the MICCAI Satellite events publication process.

No

Space and hardware requirements

Please describe the platform used for any online challenge. For on-site challenges, indicate how you plan to provide a fair computing environment. Please list any technical equipment or support needed (e.g., projectors, computers, monitors, loud speakers, microphones).

We will use a derived version of the open source grand-challenges platform for online validation. For the on-site part we will need standard conference equipment (projector, loud speakers, microphones, WiFi).

PROCEDURE TRACK (LONG-CONTEXT VIDEO UNDERSTANDING)

Summary

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

FOCUS benchmarks vision–language models on clinically relevant question answering for foreign object understanding in minimally invasive surgery, with the goal of advancing AI methods that can support intraoperative quality assurance and patient safety. Across all tracks, tasks evaluate a model's ability to interpret laparoscopic scenes and provide actionable information about foreign objects, such as detection, identification, localization, tracking, counting, and retrieval status, at increasing levels of temporal and contextual complexity.

The PROCEDURE Track challenges models with long videos (full procedures from beginning up to a certain point reflecting real intraoperative and postoperative query scenarios). Questions may require recalling when an object was last seen, aggregating object counts across long intervals, or determining whether all introduced foreign objects requiring retrieval have been removed. This track targets the key technical frontier of long-context VLMs and directly reflects real-world intraoperative quality assurance needs, where critical safety questions depend on understanding events that occurred much earlier in the operation.

Keywords

List the primary keywords that characterize the challenge. (Separate your inputs with comma like Keyword 1, Keyword 2)

Surgical AI, Surgical Data Science, Vision-Language Model, Long-context Reasoning, Surgical Foreign Objects, Laparoscopy, Endoscopy, Surgical Video Understanding

ORGANIZATION

Organizing team

Provide information on the organizing team (names and affiliations).

Coordinators:

Lena Maier-Hein (German Cancer Research Center (DKFZ))
Thomas G. Weiser (Wellcome Leap, Stanford University)

Clinical Chairs:

Daniel Hashimoto (University of Pennsylvania)
Fiona Kolbinger (Purdue University)
Thomas Pausch (University of Heidelberg)
Thomas G. Weiser (Wellcome Leap, Stanford University)

Technical Chairs:

Salman Khan (Mohamed bin Zayed University of Artificial Intelligence)
Lena Maier-Hein (DKFZ)
Stefanie Speidel (National Center for Tumor Diseases)
Danail Stoyanov (University College London)

Executive committee:

Lucas Luttner (DKFZ) - PhD lead
Patrick Godau (DKFZ) - Postdoc lead
Jule Brandt (DKFZ)
Evangelia Christodoulou (DKFZ)
Janne Heinecke (DKFZ)
Doreen Heckmann-Nötzel (DKFZ)
Niklas Holzwarth (DKFZ)
Michelle Karadeema (Wellcome Leap)
Marcel Knopp (DKFZ)
Leon Mayer (DKFZ)
Annika Reinke (DKFZ)

Contact Person

Provide information on the primary contact person.

Prof. Dr. Lena Maier-Hein
German Cancer Research Center (DKFZ)
l.maier-hein@dkfz.de

Are clinicians part of the organizing team?

Indicate whether clinicians are part of the organizing team. If yes, describe their role.

Yes, the following members of the organization team are clinicians:

Jule Brandt (DKFZ)
Thomas G. Weiser (Wellcome Leap, Stanford University)
Fiona Kolbinger (Purdue University)

Daniel Hashimoto (University of Pennsylvania)
Janne Heinecke (DKFZ)
Thomas Pausch (University of Heidelberg)

Thomas G. Weiser is a surgeon and the program director of the SAVE program at Wellcome Leap. Wellcome Leap funded parts of the work that led to the organization of the challenge.

The clinicians were involved in designing the challenge, adjusting annotations protocols, as well as conducting and reviewing annotations on surgical videos.

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place. Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline
- Repeated event as open call challenge

FOCUS is hosted within ORena, a new umbrella framework for surgical AI competitions inspired by the “Arena” paradigm, but adapted to the specific constraints and opportunities of the Operating Room (OR). The inaugural ORena competition FOCUS, sponsored by Wellcome Leap SAVE, Helmholtz Imaging and the DKFZ Division of Intelligent Medical Systems (IMSY), is dedicated to foreign object understanding in minimally invasive surgery. As the first edition under the ORena framework, FOCUS sets a foundation for future challenges that expand beyond foreign-object safety to broader long-context surgical scene understanding, procedure monitoring, and next-generation decision support in the operating room.

Event

Report the event (e.g., conference) that is associated with the challenge (if any).

International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2026

Report the platform used to run the challenge

Report the platform (e.g., grand-challenge.org, synapse, Kaggle, …) used to run the challenge.

We will host a customized clone of https://grand-challenge.org/, which is an open source Apache licensed project (https://github.com/DIAGNijmegen/rse-grand-challenge) and tailor it to the needs of our challenge.

Do you agree that your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer will not impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g., filling their submission request).

yes

Website

Provide the URL for the challenge website (if any).

https://or-arena.org/

Allowed user interaction

Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

Examples:

- No user interaction is allowed at any step
- User interaction is allowed for curating training data (i.e. excluding some training samples).

Algorithms should be fully automatic

Training data policy

Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

The PROCEDURE Track is designed as an open track to maximize innovation in long-context surgical video understanding. Participants may use any training data, including the challenge-provided dataset, publicly available datasets, proprietary/private datasets, and any pre-trained models (open or closed). There is no restriction on whether external data or models must have been publicly available at the time of challenge launch, and teams may create additional (private) annotations on any external data they use.

Organizer policy

Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

Examples:

- May not participate.
- May participate but not eligible for awards and not listed in leaderboard.

Members of the organizers' labs are not eligible for awards. The organizers will provide baseline models that serve as reference points for participants and will be clearly identified as such on the leaderboard.

Award policy

Define the award policy. In particular, provide details with respect to challenge prizes.

We have already secured $50k prize money and are negotiating with several companies to raise $100k+.

About 40% of the total prize money will be allocated to the PROCEDURE Track, with this amount distributed approximately equally between the Technical and Clinical leaderboards. Within each leaderboard, the prize money will be awarded to the top three teams in a 50% / 30% / 20% split. Adjustments may be made to ensure rounded and practical prize amounts and to deal with ties.

The prize money will be distributed to the top three teams that meet the following criteria:

1) The submission outperforms the baselines.
2) Team members are eligible to receive prize money (see below).
3) The submission description provides sufficient methodological details to allow for a meaningful interpretation.

Two baseline models will be provided:

- A state-of-the-art frontier VLM (e.g., ChatGPT, Gemini), selected based on the best performance on the validation dataset and applied without task-specific fine-tuning, given that a model with such a long context video will be available at that time.
- A state-of-the-art open-source VLM, fine-tuned by the organizers to provide a strong and accessible baseline.

Due to funding restrictions imposed by the U.S.-based funding agency, teams from countries subject to U.S. restrictions (e.g., Russia, North Korea, Iran) are not permitted to participate. All participating teams, including those ineligible for awards, will be listed on the public leaderboard to ensure transparency and fair benchmarking.

Result announcement policy

Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

The top 3 teams that beat the baseline will be announced publicly.

Publication policy

Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

All teams that have beaten the baselines will be invited as co-authors on the planned publication. Each team can name up to three co-authors. Exceptions may be made upon reasonable request.

Submission method

Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions:
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.
- Algorithm container submission (type 2) on Grand Challenge.

All submissions must be made through the challenge website, which will be based on grand-challenge. Algorithm submissions will be via Docker containers. Precise instructions will be provided alongside a template repository before the pre-evaluation phase starts. Docker containers must work without internet access. Inference will be limited to a single GPU and must be completed within a maximum time budget per question. If the answering of a question takes longer than the budget, the respective question will be treated as answered incorrectly. The budget will be depending on the track:

FRAME track: 5 seconds* on a 48GB VRAM GPU

SEGMENT track: 15 seconds* on an 80GB VRAM GPU

PROCEDURE track: 30 seconds* on an 80 GB VRAM GPU

*Note: Although we are confident with the provided numbers, constraints with the compute costs for the inference might imply slight adaptations for the exact inference resources. Updates will be made public on the official challenge website.

Teams must also submit a submission description with sufficient methodological details to allow for a meaningful interpretation of the results and to be eligible to receive the prize money. Details on the exact format will be made available at least four weeks before the final submission deadline.

Pre-evaluation

Provide information on the possibility for participating teams to evaluate their evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

A leaderboard will be set up to which each team can make up to 10 submissions*. Evaluation will be performed on a sample of 20 representative videos. Only teams that beat both baselines on at least one of the leaderboards (technical and clinical) will proceed to the final test stage.

*Note: Although we are confident with the numbers provided, constraints with the compute costs for the inference might imply slight adaptations for the exact number of submissions if we encounter a very large or very small number of participants. Updates will be made public on the official challenge website.

Creating multiple accounts per person or multiple teams per institution/lab to exploit the submission budget during pre-evaluation or final submission may result in a permanent exclusion from the challenge.

Schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any) URL
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)

- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Until 15 May: challenge website launches, registration opens, release of training data (first batch), release of code for data loading and validation

15 June: release of remaining training data (second batch), release of minimal working example to dockerize and submit

15 July: launch of pre-evaluation phase with public leaderboard

15 August: final submissions are possible for teams that beat the baselines on the validation leaderboard

September 1st: registration and pre-evaluation closes

September 8th: final submission deadline (docker + method description)

October 4th or 8th: result announcements at MICCAI

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

All videos have been deidentified using automated and manual deidentification techniques, and as they only display intraabdominal views, do not include information identifying the patient. Ethical approval for using the videos as part of the FOCUS Challenge will not be required. However, all participants in the challenge must agree to terms of use, in particular to use the videos as intended for the purposes of competing in this Challenge, and to not attempt to re-identify any information within the videos in order to determine its provenance or clinicians/facilities involved.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied (click here for more information).

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

The 30 HEICO videos have already been publicly released under a CC BY-NC-SA license [1, 2].

The remaining 170 training videos will be available under the following data usage agreement:

By registering for the FOCUS Challenge, each team agrees:

(1) to use the data and videos provided only within the scope of the challenge and for the purposes of participating in the challenge,

(2) to neither pass it on to a third party nor share it beyond members of the team,

(3) to not publish the data or underlying annotations, or otherwise make them publicly available, and

(4) to refrain from any attempt to reidentify information that has been deidentified, such as individual surgeons or facilities, or the provenance of the video,

(5) to maintain the data within a protected/secure environment compliant with HIPAA, GDPR, or similar regulation, and ensure access to the videos are restricted to members of the challenge team only.

Algorithms and models generated by individual teams may be used for noncommercial or commercial purposes.

All question answer pairs for the training data will be publicly available under a CC BY license, except for the additionally annotated questions and answers on the HEICO data, which will be released under the CC BY-NC-SA license to comply with the license of the original video data.

[1] Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P. M., ... & Müller-Stich, B. P. (2021). Heidelberg colorectal data set for surgical data science in the sensor operating room. Scientific data, 8(1), 1-11.

[2] Roß, T., Reinke, A., Full, P. M., Wagner, M., Kenngott, H., Apitz, M., ... & Maier-Hein, L. (2021). Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge. Medical image analysis, 70, 101920.

Code availability of the organizers

Provide information on the accessibility of the organizers' evaluation software (e.g., code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

The evaluation code will be made publicly available via a public GitHub repository upon launch of the challenge, though certain details, including the exact judge models for evaluation will remain undisclosed until the end of the challenge, to prevent tuning towards the LLM judges. The code to produce rankings based on metric results will be released along with the training data.

Code availability of the participating teams

In an analogous manner, provide information on the accessibility of the participating teams' code.

Model publication for the PROCEDURE Track is not mandatory to be eligible for awards (see above). Participants are, however, strongly encouraged to make their training scripts, inference code, and trained models publicly available to promote transparency, reproducibility, and scientific exchange within the community. Teams choosing not to release their code are still required to provide sufficient methodological details in their submission descriptions to allow for a meaningful interpretation of the results and to be eligible to receive the prize money. This balanced policy aims to foster open research while maintaining inclusivity for industrial and applied contributions.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Access to the test labels: All annotations are hosted at DKFZ. Clinicians of the organizing team have access to all labels. Involved annotation companies had access to raw videos and the bounding boxes they created.

The FOCUS Challenge is sponsored in large part by the Wellcome Leap SAVE program, led by program director Thomas G. Weiser. Wellcome Leap is a 501c3 charitable organization based in the United States with a mission to accelerate health breakthroughs on a global scale.

It is further sponsored by the ORena initiative of the DKFZ Intelligent Medical Systems (IMSY) lab as well as Helmholtz Imaging.

Funding for compute infrastructure will be provided by Amazon Web Services (AWS).

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Intervention assistance

Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection

- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Visual Question Answering (VQA)

Temporal Reasoning

Target cohort

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort)

Describe the target cohort of task, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Patients undergoing minimally-invasive body cavity (thorax, abdomen, pelvis, retroperitoneum) surgery involving foreign objects anywhere in the world.

Challenge cohort

Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

The challenge cohort consists of patients that underwent minimally invasive surgical intervention at one of multiple sites across the globe and encompasses a diverse range of surgical procedures and patient populations. Part of the training videos (30 videos) were taken from the existing HeiCo dataset [1] and annotated with questions and answers. Most of the remaining data were acquired as part of the SAVE project funded by Wellcome Leap [2] and are not yet publicly available.

Training data include the following laparoscopic surgical procedures:

- 170 cholecystectomies
- 10 proctocolectomies
- 10 rectal resections
- 10 sigmoid resections

Test data will include 200 videos from a broad range of procedures (cholecystectomies and additional procedure types not to be conveyed to the participants). Leaderboard validation data will include 20 videos representative of the test data.

[1]: Maier-Hein, L., Wagner, M., Ross, T. et al. Heidelberg colorectal data set for surgical data science in the sensor operating room. Sci Data 8, 101 (2021). https://doi.org/10.1038/s41597-021-00882-2

[2]: https://wellcomeleap.org/save/

Imaging technique(s)

Specify the imaging technique(s) applied in the challenge.

Surgical endoscopy; Diverse systems applied

Context information: Image data

Provide additional information given along with the images. The information may correspond directly to the imaging data (e.g., tumor volume).

We will provide the procedure name along with the video as well as the time-point of the last frame representing the time at which a question is asked during a procedure.

Context information: Patient

Provide additional information given along with the images. The information may correspond to the patient in general (e.g., gender, medical history).

none

Data origin

Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g., brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Internal anatomy during minimally-invasive surgery

Algorithm Target

Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g., tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Foreign objects in relation to patient anatomy

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (parameter 26), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter 26):

Accuracy, Applicability, Complexity, Consistency, Ergonomics, Feasibility, Hardware requirements, Interaction, Integration in workflow, Precision, Reliability, Robustness, Runtime, Sensitivity, Specificity, Usability, User satisfaction

Accuracy, Reliability, Robustness

DATA SETS

Data acquisition device(s)

Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g., manufacturer) as well as information on additional devices used for performance assessment (e.g., tracking system used in a surgical setting).

The 30 HeiCo videos were acquired in the integrated operating room KARL STORZ OR1 FUSION® (KARL STORZ SE & Co KG, Tuttlingen, Germany). The remaining cases were recorded with diverse surgical endoscopy systems.

Data acquisition details

Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g., image acquisition protocol(s)).

N/A - different protocols from systems used around the globe

Center(s)/institute(s)

Specify the center(s)/institute(s) in which the data was acquired and/or the ata providing platform/source (e.g., previous challenge). If this information is not provided (e.g., for deidentification reasons), specify why.

The 30 HeiCo videos were recorded at Heidelberg University Hospital. The test data was acquired from more than 5 centers not represented in the training data.

Characteristics of the subjects

Describe relevant characteristics (e.g., level of expertise) of the subjects (e.g., surgeon)/objects (e.g., robot) involved in the data acquisition process (if any).

N/A for regulatory reasons

Case definition

State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (parameter 21) and may include context information (parameter 18). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

In this track, one case corresponds to a single VQA instance derived for a laparoscopic surgical video. A case consists of: (1) a full procedure from beginning up to a certain point, (2) a question, (3) categorization of the question according to the challenge taxonomy (shared with the participants only for the training data) (4) meta information about the procedure type and (5) a reference answer provided by an expert surgeon or extracted automatically from the expert annotations.

Number of cases

State individually total number of training, validation and test cases.

Training: 20,000 questions obtained from 200 videos

Validation: 2,000 questions obtained from 20 videos

Test: 20,000 questions obtained from 200 videos

Quantity of data which is already annotated

How much of the data are already annotated (stratified by train test in percentage)?

The annotation process comprises 5 stages (see below “Instructions given to the annotators”). We have completed stage 1-3 for all the data and are in the process of completing stage 4.

Explanation of data proportion

Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

The test data was selected to rigorously stress-test robustness across diverse procedure types, foreign object classes, and question formulations. We reserved 20 videos for the online leaderboard and split the remaining videos 50/50 in a training and test set. Since participants are encouraged to leverage data beyond the challenge dataset for training, we can afford this splitting without restricting participants too much.

Further important characteristics of the cases

Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

The majority of the videos (390) and all labels are new. The test data will not be released.

Further important characteristics of the cases

Mention further important characteristics of the training, validation and test cases (e.g., class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

N/A

Method for determining the reference annotation

Describe the method for determining the reference annotation i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

The scope for the questions in FOCUS ranges from perception-level questions in single frames (FRAME Track; this one) to long-horizon reasoning across entire procedures (PROCEDURE Track), while maintaining a consistent focus on foreign objects. Questions in the PROCEDURE Track are constructed such that they require analysis of the provided video and that they can be unambiguously answered based on the video as well as surgical textbook knowledge. This ensures that the benchmark evaluates multimodal reasoning grounded in visual evidence and procedural context rather than external medical knowledge alone. Questions unrelated to foreign objects or requiring speculative clinical judgment beyond the observable context are deliberately excluded.

The annotations were performed in five stages. For each Stage, a detailed annotation protocol was provided to the annotators.

STAGE 1 - Foreign object presence screening (5s clips).

Purpose: Identification of video segments with foreign objects

Input: 5 sec clips

Output: Presence flag for each class of foreign objects on each clip

Annotators: QualityMatch Crowdsourcing company [1]

STAGE 2 - Frame-level localization (bounding boxes at 1 fps).

Purpose: Identification and localization of foreign objects in individual frames

Input: Frames (sampled at 1 fps) corresponding to clips with foreign objects

Output: Frames (sampled at 1 fps) with bounding boxes and class assignment for all foreign objects

Annotators: QualityMatch and iMerit Crowdsourcing companies [1, 2]

STAGE 3 - Automated bounding box completion/correction.

Purpose: Automatic refinement of crowd-based foreign object identification and localization

Input: Frames (sampled at 1 fps) with bounding boxes and class assignment for all foreign objects.

Output: Automatically corrected frames (sampled at 1 fps) with bounding boxes and class assignment for all foreign objects

STAGE 4 - Manual refinement, instance assignment, and expert verification.

Purpose: Instance assignment to individual objects and final correction of bounding boxes and class assignments

Input: Frames (sampled at 1 fps) with bounding boxes and class assignment for all foreign objects.

Output: Frames (sampled at 1 fps) with class and instance assignments and bounding boxes around all foreign objects

Annotators: Initial annotation generation by a team of medical students. Verification and correction by a team of four expert surgeons and two highly experienced medical students. In this stage, missing objects can still be added.

STAGE 5 - VQA generation and clinical question design.

Purpose: Generation of questions and answers based on foreign object information. Part of the questions are automatically generated based on the available instance information. Further questions are added by a team of expert surgeons

Input: Videos with class and instance assignments and bounding boxes around all foreign objects at 1 fps

Output: N questions per video Annotators: A team of > 10 expert surgeons and medical students.

For track 3, we classify questions as follows (potentially with multiple labels):

- Object recognition and identity matching (Which object (type)? In which state? Where?): Recognition of object instances, their semantic type, attributes, spatial context, and identity consistency across time
- Temporal grounding (When? How long?): Localization of events or object occurrences within the video timeline, including their exact temporal position and duration
- Aggregation (How many?): Combination of information on foreign objects and events across multiple objects, instances, and/or time points
- Event and procedural understanding (Which action?): Recognition and interpretation of actions, events, and their procedural structure over time
- Complex reasoning (Why? What happens if?): Inference of functional, causal, or outcome-related information beyond direct observation

The complete taxonomy will be released with the first batch of training data.

[1] https://www.quality-match.com/product

[2] https://imerit.net/domains/medical-ai/

Instructions given to the annotators

Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

The annotation process was executed in five stages, utilizing specific software frameworks and training protocols for each:

- Stages 1 & 2 (Foreign object presence screening and frame-level localization): These stages utilized the QualityMatch [1] and iMerit [2] annotation frameworks. All crowd-sourced annotators underwent rigorous vendor-led training specific to the selected tools and the annotation tasks prior to commencing work.
- Stage 3 (Automated bounding box completion/correction): A semi-automated quality control phase was implemented using a Python-based YOLO object detection model. This model performed cross-validation on the data generated in the previous stages to identify and flag inconsistencies.
- Stage 4 (Manual refinement and instance assignment): A customized version of the CVAT (Computer Vision Annotation Tool) [3] was adapted specifically for the FOCUS challenge. For this critical stage, annotators received direct training from the challenge executive committee regarding both the adapted tool and the specific refinement objectives.
- Stage 5 (Annotation enrichment): This stage involved the creation of a VQA dataset centered on expert clinical reasoning and the identification of "moments of importance" (e.g., specific points of foreign object insertion and retrieval) from the temporal data generated in Stage 4. Prior to annotation, clinicians, who were actively involved in the co-design of the enrichment protocol, underwent a specialized training on CVAT. The training provided instructions for two primary tasks: the expert verification of existing instance annotations from previous stages and the manual enrichment with complex question-answer pairs. Following the established protocol, clinicians were guided to derive both open- and closed-ended questions that addressed medical reasoning and semantic surgical understanding specifically tailored to the type of foreign object identified.

Documentation: Comprehensive annotation instructions were provided to annotators at every stage. These detailed protocols will be made available alongside the dataset.

[1] https://www.quality-match.com/product

[2] https://imerit.net/domains/medical-ai/

[3] https://github.com/cvat-ai/cvat

Details on the subject(s)/algorithm(s) that annotated the cases

Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g., information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Stage 1-2: Crowd workers with expertise with the tool and basic knowledge with medical tasks from previous annotation projects.

Stage 3: Object detection models (YOLO) have been used to enrich the crowd annotations with further bounding boxes by identifying possible missing objects.

Stage 4: Medical students (11 at the time of the submission of the challenge) with up to 3 years of annotation experience. These are all students from the University of Heidelberg with expertise within the surgical domain.

Stage 5: 4 Expert surgeons including the clinical PIs of the project with more than 10 years of experience in surgical video analysis and highly experienced medical students.

Method(s) used to merge multiple annotations

Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Multi-stage annotation process with multiple verification steps as detailed above.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

The HeiCo dataset [1] was used as is, without further pre-processing. New videos have been de-identified by out-of-body detection and blurring respective parts. These videos have also been trimmed accordingly.

[1]: Maier-Hein, L., Wagner, M., Ross, T. et al. Heidelberg colorectal data set for surgical data science in the sensor operating room. Sci Data 8, 101 (2021). https://doi.org/10.1038/s41597-021-00882-2

Sources of error related to the image annotation

Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter- and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

There are multiple sources of error related to the annotation process. To ensure high-quality annotations, we conducted a multi-stage annotation process with the following correction/verification steps:

1) Correction of crowd object detection annotations with an automatic method to minimize missed foreign object instances
2) Correction of automatic annotations by medical students to ensure a high level of consistency, set up as a two stage process, based on a four-eyes principle and assisted by automatic flagging
3) Verification (and possible correction) of student annotations by senior surgeons

Other sources of error

In an analogous manner, describe and quantify other relevant sources of error.

N/A

ASSESSMENT METHODS

Metric(s)

Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (parameter 20). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

We navigated the tradeoff between interpretability of the ranking and methodological precision by separating the ranking metric from the detailed performance analyses. For detailed performance analyses, we apply targeted metrics for each question type, such as set-based metrics for questions asking about visible foreign-object classes, distance-based measures for temporal localization questions and Expected Cost for ordinal outputs (e.g., counting questions). To simplify the ranking scheme, we work with Accuracy as the primary metric and compensate for the main disadvantages - specifically the prevalence dependency and threshold sensitivity - with appropriate data splits and hierarchical and stratified aggregation.

Accuracy is computed as the proportion of correctly answered questions. For open-ended questions, semantic correctness is assessed by an LLM-as-a-judge approach. Up to three LLMs decide via majority vote whether the proposed answer matches the reference. We will make sure that any attempt to manipulate the LLM-as-a-judge through adversarial prompting, "jailbreaking", or other malicious techniques designed to unfairly influence the evaluation will result in immediate disqualification of the team. In cases where annotation uncertainty exists, for example with respect to a specific temporal reference, an application-dependent variant of Accuracy is employed [1]. In this setting, predictions are considered correct (TP) if they fall within a specific tolerance range (e.g., a temporal window). Tolerance ranges are defined based on inter-rater variability and/or clinical input.

[1] Dergachyova, O., Bouget, D., Huaulmé, A., Morandi, X., & Jannin, P. (2016). Automatic data-driven real-time segmentation and recognition of surgical workflow. International journal of computer assisted radiology and surgery, 11(6), 1081-1089.

Justification of metric(s)

Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

For detailed performance analyses, metrics are chosen in a customized manner depending on the specific question, as suggested by the Metrics Reloaded framework [1]. For the ranking scheme, Accuracy was selected to reflect our primary assessment goal of correctness, i.e., whether a model provides correct answers to clinically meaningful questions. Accordingly, Accuracy, as it directly operationalizes correctness in a transparent manner and represents the state-of-the-art standard for validating VQA tasks. It further enables consistent comparison across categorical and open-ended questions.

For open-ended questions, an LLM-as-a-judge paradigm is employed, which has become a common and effective approach in modern VQA and VLM benchmarks to assess semantic correctness while allowing for clinically meaningful linguistic variability. Preliminary experiments demonstrated only very minor discrepancies when using different state-of-the-art large language models as judges.

Reliability aspects are addressed through the overall challenge design rather than through additional metrics. Specifically, the test set comprises unseen procedures, unseen centers and unseen questions compared to the training data. A stratification approach enables the testing of generalization capabilities.

[1] Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., ... & Jäger, P. F. (2024). Metrics reloaded: recommendations for image analysis validation. Nature methods, 21(2), 195-212.

Method used to compute a performance rank

Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking. Ideally, provide the ranking scheme as a concrete pseudo code.

Key design goals

- Interpretability: while we take a complex approach to the detailed performance analysis of methods (see statistics section below), our aim was to have an easy-to-understand ranking scheme. For this reason, we based the ranking on a single metric.
- Fairness across categories: avoid bias from uneven question counts (e.g., many identification questions)
- Noise resistance: irrelevant performance differences within a category should not be rewarded
- Robustness-first: emphasize generalization, e.g., procedure types that have not been present in the training data
- Theoretical guarantees: although no merging mechanism can check all boxes of desirable properties (“Arrow's impossibility theorem”), we prefer one with overall more desirable properties
- Reproducibility: deterministic and transparent aggregation rules

Overview

The ranking happens in three steps:

1. Capability-level ranking: For different capabilities individual rankings are created. Irrelevant performance differences result in identical ranks.
2. Merging through Copeland method: The individual rankings are combined to determine the overall ranking. We use the Copeland method to do so [1].

3. Resolving ties: For the top 3 places we will resolve ties by looking at the mean win rate of tied models using bootstrapping within all capabilities.

Proposed ranking scheme

Each test case is annotated with the following meta information:

i. Robustness level: In-distribution (ID) vs Out-of-distribution (OOD) tag with respect to procedure type and question
ii. Clinical relevance: Binary variable indicating whether a question is clinically relevant (e.g., “Where is the sponge that was inserted at time hh:mm:ss?”) or only of technical interest (e.g., “When was the second needle inserted?”)
iii. Primary capability: Given the primary question intent, the test case is mapped to one primary capability using the proposed taxonomy (Main capabilities: Object Recognition and Instance Matching, Temporal Grounding, Aggregation, Event and Scene Understanding, Reasoning)

The following will be performed separately for (i) all questions and (ii) only the clinically relevant questions.

Using these attributes we split all questions into buckets: For each model, and each primary category, model scores are computed for the ID and OOD questions. They represent the mean Accuracy over all questions belonging to the respective primary category and distribution split.

The initial ranking for each bucket is based on this Accuracy score. Then we will perform significance tests between pairs of models to determine whether a performance delta truly justifies different ranks and adjust the ranking so that non-significant differences result in the same rank position.

The merged ranking is then computed from the Copeland rule over all rankings. We say model A dominates model B in case A is ranked higher than B more often than B is ranked higher than A. The score of each model is given by the number of models it dominates minus the number of models it is dominated by. The higher the score, the better.

In case of ties within the first three positions, we will resolve them by directly comparing all tied models. Within each bucket we perform bootstrapping and compute the respective Accuracies of any bootstrap sample. The win rate is the fraction of bootstrap samples that result in a model to have the highest Accuracy of all models. Tied models are then sorted by their mean win rate over all buckets.

To outperform a baseline model during the pre-evaluation phase it suffices to have a higher mean Accuracy over the buckets. We will also make mean accuracy performances of the final phase public.

Two leaderboards will be compiled:

- Technical leaderboard: Measures general foreign object understanding and is computed from the macro-average of all scores corresponding to a model using all VQA pairs in the test set.
- Clinical leaderboard: Measures the model’s ability to answer clinically relevant questions and is computed from the macro-average of all scores corresponding to a model using only clinically relevant VQA pairs in the test set.

[1]: Rofin, Mark, et al. "Vote’n’rank: Revision of benchmarking with social choice theory." Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023.

Submissions with missing results

Describe the method(s) used to manage submissions with missing results on test cases.

Test cases with missing response data will be handled like false answers.

Justification of ranking

Justify why the described ranking scheme(s) was/were used.

We designed the ranking scheme according to the above-summarized requirements. To reward robustness, we give equal weight to ID and OOD questions. We evaluate each primary capability separately before combining them with equal weight. Before combining the scores across all buckets, we use the significance testing to ensure only meaningful performance deltas result in different ranks. We favor Copeland over other ranking schemes (for example the linear ranking, also known as the Borda rule [1]) because it has more desirable properties [1], particularly it is more robust against irrelevant alternatives. As a measure to resolve ties between models after the Copeland method, we use the

head-to-head win rate with bootstrapping to determine a better model. The final ranking scheme focuses purely on the question, "What is the best model?" rather than assessing the extent to which a problem is solved.

[1]: Rofin, Mark, et al. "Vote'n'rank: Revision of benchmarking with social choice theory." Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023.

Statistics - Overview

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

While we keep the ranking scheme relatively simple for the purpose of interpretability, the purpose of our statistical analyses is to provide a nuanced understanding of model capabilities. It is based on the following principles:

- Taking non-independency of answers present in the data into account: Given the dependencies in our data (e.g., frames and segments originating from the same video, or answers retrieved from questions referring to the same frames, cannot be considered independent), we will account for this clustering in all inferential analyses. Specifically, model performance variability (i.e., standard deviation, confidence intervals) will be estimated using bootstrapping while taking data clustering into account.
- Multi-label handling: As many questions are assigned multiple capability labels (e.g., temporal grounding and spatial localization), analyses will explicitly account for label overlap and explore its effect on model performance. We will report both marginal performance per category and conditional performance within key co-occurrence strata. Additionally, adjusted category effects will be estimated using multivariable logistic regression models.
- Stratified analyses: We will stratify analyses along multiple dimensions, including sub-capabilities, context length (short, medium, long), clinical impact levels, and robustness setting (ID/OOD). This enables systematic characterization of where performance degrades as task complexity increases and where specific capability combinations pose challenges.
- Ranking uncertainty quantification: Following our proposed ranking scheme (see above), we will in the end report measures of variability of the resulting rankings, such as the standard deviation and confidence intervals.

Statistics - Precision of the performance estimates

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g., confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

To evaluate the precision of performance estimates, we will be implementing bootstrapping to compute confidence intervals of performance metrics on the test set. Moreover, performance reporting will be stratified per task categories as described in [1] and the precision estimates will be reported alongside confidence intervals.

[1] Kurpath, M. I., Kaithakkodan, J. M., Zhou, J., Mullappilly, S. S., Almansoori, M., Ahsan, N., ... & Cholakkal, H. (2025). A Benchmark and Agentic Framework for Omni-Modal Reasoning and Tool Use in Long Videos.

Statistics - Performance variability across cases

Provide a description of how variability of the performance of individual algorithms across test cases is assessed (e.g., SD across test cases, IQR, graphs, reporting outliers...).

To describe performance variability across test cases, we will report the standard deviation of the performance metrics and complement this with boxplots that visually illustrate variability for each combination of task and participating algorithm. Furthermore, We will stratify analyses along multiple dimensions, including sub-capabilities, clinical impact levels, and robustness setting (ID/OOD).

Statistics - Rankings variability

Provide a description of how variability of rankings is assessed.

To assess and visualize ranking variability, the challengeR software [1] will be used. In addition, we will report confidence intervals for the differences in performance between participating algorithms for each performance metric considered.

[1] Wiesenfarth, M., Reinke, A., Landman, B. A., Eisenmann, M., Saiz, L. A., Cardoso, M. J., ... & Kopp-Schneider, A. (2021). Methods and open-source toolkit for analyzing and visualizing challenge results. Scientific reports, 11(1), 2369.

Statistics - Tests for significance

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

Our approach for determining model rankings incorporates model performance uncertainty by means of bootstrapping and thus offers stronger evidence for outperformance and subsequently ranking stability compared to traditional significance testing.

Statistics - Missing data handling

Provide a description of the missing data handling.

Any question for which results data is missing will be treated as an incorrect answer and be assigned the worst possible score (depending on the metric).

Statistics - Software

Indicate any software product that is used for all data analysis methods.

Our statistical analyses will be conducted in Python using relevant libraries such as scipy.stats and statsmodels.

Further analyses

Present further analyses to be performed (if applicable), e.g., related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

Additional analyses will include subgroup performance breakdowns across technical (context length, capability) and clinical axes (decision impact, required knowledge), robustness gap analysis between ID and OOD, and qualitative failure mode analysis. We will also explore post-hoc ensembling of compatible submissions to quantify potential upper bounds.