# The SAGES Critical View of Safety Challenge: Structured description of the challenge design

## CHALLENGE ORGANIZATION

### Title

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

The SAGES Critical View of Safety Challenge

### Challenge acronym

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

CVS Challenge

### 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.

The application of Computer Vision (CV) and Machine Learning (ML) to minimally invasive surgery promises objective assessment of visual features in surgical video that contribute to surgical decision-making. Future prospects for surgical risk mitigation include enhanced supervision for surgeons and augmented teaching opportunities. Currently, surgical Artificial Intelligence (AI) is limited to research, waiting to be translated into clinical practice. There is a lack of widely validated results, calibration of uncertainty, robustness to domain shifts, and a high computational barrier to enable wide deployment that limits the application of AI to life-saving intraoperative use.

Laparoscopic cholecystectomy, a standardized operation for gallbladder removal, is one of the most frequently performed minimally invasive procedures worldwide. While it has become a benchmark procedure for computational exploration of intraabdominal video data, there is no clinical translation, partly due to the limited generalizability of AI architectures developed almost entirely on localized, homogenous datasets.

The Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) Critical View of Safety (CVS) Challenge is the first international biomedical data challenge from a surgical society, offering a unique infrastructure for global data collection and leveraging multidisciplinary expertise for the standardized assessment of the CVS. The aim of the challenge is to computationally address the detection of CVS, a routinely performed surgical safety measure crucial for minimizing bile duct injuries during cholecystectomy. Despite the high frequency of cases and standardized operative approach for laparoscopic cholecystectomy, there is a risk of significant intra- and post-operative complications, such as common bile duct injuries. Assisting surgeons in achieving and recognizing the CVS will help to improve the safety of laparoscopic cholecystectomy worldwide.

The challenge offers a global and diverse dataset of 1000 laparoscopic cholecystectomy videos, provided by 67 surgeons (data donors) from 53 countries and 6 continents, alongside clinically relevant metadata. This dataset encompasses a wide diversity of patient demographics and procedural quality to reflect the worldwide diversity in patients and surgeons. The data will be released to the global community with the aim of developing models capable of reliably and consistently classifying the CVS with adequate generalizability to the global patient population. By incorporating the perspectives of clinicians, computer scientists, and industry through structured multidisciplinary Advisory Committees (AC), the challenge offers the opportunity for the development of AI suitable for high-stakes surgical settings.

Data acquisition was designed to provide consistent and reliable deidentification of out-of-body images and pseudonymization of metadata. The dataset has been meticulously curated based on standardized protocols composed through expert consensus of the multidisciplinary ACs. The data has been indexed according to demographics -- source location, performing surgeons` experience level, surgical indication, and clinical characteristics in the video (e.g., fluorescence, robotics, intraoperative cholangiogram). Each video was annotated with the CVS and its subcomponents to ensure consistency and reliability in the data. The structured annotation pipeline, rooted in a consensus annotation protocol revised by clinical experts in hepatobiliary surgery, includes proficiency-based training of annotators from multiple countries. The annotation task was the classification of the three CVS Criteria on a video and frame basis.
The execution of the CVS Challenge is governed by metrics selected by the multidisciplinary AC to evaluate AI models` performance in identifying the achievement of the subcomponents and overall CVS. Participants can work with various data splits, enabling them to test the robustness of their algorithms across a heterogeneous dataset rich in clinical and demographic variability.

The CVS Challenge sets a new benchmark in collaborative efforts between surgeons and computer scientists, encompassing consensus-based guidelines governing a global data acquisition and curation framework. Additionally, structured annotation curricula based on clinical expert consensus, ensure a homogenous, clinically meaningful ground truth for model training. The primary aim of the challenge is to ensure uniformity in the computational assessment of the CVS for high clinical value, with the goal to improve safety and outcomes of laparoscopic cholecystectomy. The extensive dataset, rigorous curation protocol, and strategic execution of the challenge, backed by a strong advisory framework, pave the way for transformative developments in surgical practice and patient care as well as opening the possibility for future challenge iterations.

### Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Surgical Safety, Surgical AI, Minimally Invasive Surgery, Laparoscopic Cholecystectomy, Classification

### Year

The challenge will take place in 2024

## FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

### Workshop

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

The challenge is not part of a workshop.

### Duration

How long does the challenge take?

Full day.

### 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.

100

### Publication and future plans

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

The joint publication will be coordinated by the challenge organizers. The organizers will publish a manuscript summarizing the challenge results within six months after completion of the challenge. This paper will publish at least the top three winning methods (1st, 2nd, 3rd).

Annotators, who successfully completed the Annotation School and all videos assigned to them by the challenge organizers (minimum 100 videos) will be listed as co-authors in the final manuscript. Data Donors, who contributed at least 50 unique laparoscopic cholecystectomy videos and metadata through the designated Video Acquisition Portal, and attended at least one official CVS Challenge summit, will be acknowledged in a `Data Donor Consortium`. Advisory Committee members who attended at least one official CVS Challenge summit and actively participated in the composition of the CVS Challenge Annotation Protocol, will be acknowledged in an Advisory Committee Consortium. In addition to the listed authorship guidelines specific to the individual co-author`s role, all authors and consortium members must assist in the drafting and revision of the manuscript and critically assess important intellectual content. Furthermore, all co-authors and consortium members must give final approval of the manuscript version to be published. The final order of authorship will be determined by the challenge organizers.

Future iterations of the CVS Challenge (e.g. data challenges centered around different use cases in laparoscopic cholecystectomy / different procedures) can use the established pipeline for data acquisition and annotation. Following this publication, the participating teams are welcome to publish their individual results under citation of the official challenge manuscripts and all previous manuscripts summarizing the CVS Challenge.

### Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

This is dependent on the specific sub-challenges. Additionally, computing resources are addressed in the proposed subchallenge C.

# TASK 1: Critical View of Safety Achievement

## 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.

The application of Computer Vision (CV) and Machine Learning (ML) to minimally invasive surgery promises objective assessment of visual features in surgical video that contribute to surgical decision-making. Future prospects for surgical risk mitigation include enhanced supervision for surgeons and augmented teaching opportunities. Currently, surgical Artificial Intelligence (AI) is limited to research, waiting to be translated into clinical practice. There is a lack of widely validated results, calibration of uncertainty, robustness to domain shifts, and a high computational barrier to enable wide deployment that limits the application of AI to life-saving intraoperative use.

Laparoscopic cholecystectomy, a standardized operation for gallbladder removal, is one of the most frequently performed minimally invasive procedures worldwide. While it has become a benchmark procedure for computational exploration of intraabdominal video data, there is no clinical translation, partly due to the limited generalizability of AI architectures developed almost entirely on localized, homogenous datasets.

The Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) Critical View of Safety (CVS) Challenge is the first international biomedical data challenge from a surgical society, offering a unique infrastructure for global data collection and leveraging multidisciplinary expertise for the standardized assessment of the CVS. The aim of the challenge is to computationally address the detection of CVS, a routinely performed surgical safety measure crucial for minimizing bile duct injuries during cholecystectomy. Despite the high frequency of cases and standardized operative approach for laparoscopic cholecystectomy, there is a risk of significant intra- and post-operative complications, such as common bile duct injuries. Assisting surgeons in achieving and recognizing the CVS will help to improve the safety of laparoscopic cholecystectomy worldwide.

The challenge offers a global and diverse dataset of 1000 laparoscopic cholecystectomy videos, provided by 67 surgeons (data donors) from 53 countries and 6 continents, alongside clinically relevant metadata. This dataset encompasses a wide diversity of patient demographics and procedural quality to reflect the worldwide diversity in patients and surgeons. The data will be released to the global community with the aim of developing models capable of reliably and consistently classifying the CVS with adequate generalizability to the global patient population. By incorporating the perspectives of clinicians, computer scientists, and industry through structured multidisciplinary Advisory Committees (AC), the challenge offers the opportunity for the development of AI suitable for high-stakes surgical settings.

Data acquisition was designed to provide consistent and reliable deidentification of out-of-body images and pseudonymization of metadata. The dataset has been meticulously curated based on standardized protocols composed through expert consensus of the multidisciplinary ACs. The data has been indexed according to demographics -- source location, performing surgeons` experience level, surgical indication and clinical

characteristics in the video (e.g., fluorescence, robotics, intraoperative cholangiogram). Each video was annotated with the CVS and its subcomponents to ensure consistency and reliability in the data. The structured annotation pipeline, rooted in a consensus annotation protocol revised by clinical experts in hepatobiliary surgery, includes proficiency-based training of annotators from multiple countries. The annotation task was the classification of the three CVS Criteria on a video and frame basis.
The execution of the CVS Challenge is governed by metrics selected by the multidisciplinary AC to evaluate AI models` performance in identifying the achievement of the subcomponents and overall CVS. Participants can work with various data splits, enabling them to test the robustness of their algorithms across a heterogeneous dataset rich in clinical and demographic variability.

The CVS Challenge sets a new benchmark in collaborative efforts between surgeons and computer scientists, encompassing consensus-based guidelines governing a global data acquisition and curation framework. Additionally, structured annotation curricula based on clinical expert consensus, ensure a homogenous, clinically-meaningful ground truth for model training. The primary aim of the challenge is to ensure uniformity in the computational assessment of the CVS for high clinical value, with the goal to improve safety and outcomes of laparoscopic cholecystectomy. The extensive dataset, rigorous curation protocol, and strategic execution of the challenge, backed by a strong advisory framework, pave the way for transformative developments in surgical practice and patient care as well as opening the possibility for future challenge iterations.

### Keywords

List the primary keywords that characterize the task.

Surgical Safety, Surgical AI, Minimally Invasive Surgery, Laparoscopic Cholecystectomy, Classification

## ORGANIZATION

### Organizers

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

Ozanan Meireles, MD (1,7), Nicolas Padoy, PhD (2), Daniel Hashimoto, MD (3), Jennifer Eckhoff, MD (1,4), Deepak Alapatt, MSc (2), Guy Rosman, PHD (1), Pietro Mascagni, MD, PhD (2), Jean-Paul Mazallier, PhD (2), Xiang Li, PhD (5), Zhiliang Lyu, MSc (5), Sarah Choksi, MD (6), Filippo Filicori, MD (6), Yutong Ban, PhD (1)

1 - Surgical Artificial Intelligence and Innovation Laboratory (SAIIL), Massachusetts General Hospital (MGH), 55 Fruit Street, Wang Ambulatory Care Center 3-339, Boston, USA
2 - CAMMA, IHU Strasbourg, 1 Place d`Hopital, Strasbourg, France
3 - Penn Computer Assisted Surgery and Outcomes Laboratory (PCASO), University of Pennsylvania, 3400 Spruce Street, Philadelphia, USA
4 - University Hospital Cologne, Kerpenerstraße 66, 50937 Köln, Germany
5 - Center for Advanced Medical Computing and Analysis (CAMCA), MGH, 55 Fruit Street, Boston, USA
6 - Northwell Health, 186 East 76th Street, New York, USA
7 - SAIIL, Duke University, 2301 Edwin Rd, Durham, USA

b) Provide information on the primary contact person.

Jennifer Eckhoff, MD
Email: jennifer.eckhoff@uk-koeln.de
Phone: +491714720837 / +1 (617) 686 8019

### 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

One-time event with a fixed submission deadline.

### Challenge venue and platform

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

International Conference on Medical Image Computing and Computer Assisted Intervention 2024 (MICCAI)

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

https://cvs-challenge.grand-challenge.org/

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

https://cvschallenge.org

### Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully automatic.

b) 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 to publicly available data including (open) pre-trained nets.

Training data should be obtained from the provided SAGES CVS Dataset, as well as from publicly available datasets. No privately curated or annotated datasets are permitted to be used in training.

c) 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.

Members of the organizers` institutes may participate, but cannot be directly working with the organizers (i.e. same team) and will not be eligible for awards.

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

The first, second, and third team will receive awards. The awards stipend is to be determined.

e) 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.

Top three performing methods will be announced publicly on the challenge website. These teams will be listed as co-authors in the final manuscript.

f) 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).

The joint publication will be coordinated by the challenge organizers. The organizers will publish a manuscript summarizing the challenge results within six months after completion of the challenge. This paper will publish at least the top three winning methods (1st, 2nd, 3rd).

Annotators, who successfully completed the Annotation School and all videos assigned to them by the challenge organizers (minimum 100 videos) will be listed as co-authors in the final manuscript. Data Donors, who contributed at least 50 unique laparoscopic cholecystectomy videos and metadata through the designated Video Acquisition Portal, and attended at least one official CVS Challenge summit, will be acknowledged in a `Data Donor Consortium`. Advisory Committee members who attended at least one official CVS Challenge summit and actively participated in the composition of the CVS Challenge Annotation Protocol, will be acknowledged in an `Advisory Committee Consortium`. In addition to the listed authorship guidelines specific to the individual co-authors' role, all authors and consortium members must assist in the drafting and revision of the manuscript and critically assess important intellectual content. Furthermore, all co-authors and consortium members must give final approval of the manuscript version to be published. The final order of authorship will be determined by the challenge organizers.

Future iterations of the CVS Challenge (e.g. data challenges centered around different use cases in laparoscopic cholecystectomy / different procedures) can use the established pipeline for data acquisition and annotation. Following this publication, the participating teams are welcome to publish their individual results under citation of the official challenge manuscripts and all previous manuscripts summarizing the CVS Challenge.

### Submission method

a) 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: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

The submission instructions will be published on the CVS Challenge website as well as on grand-challenge.org. Results will be submitted via a docker container. The model output format will be published along with the

instructions. There should only be one final submission per team per sub-challenge. No evaluation of algorithms before submission of the final results will be conducted by the challenge organizer. Participants will be free to self-evaluate using internal validation sets. To note, all submissions will require inference to be possible on a single Nvidia GPU (<12 GB).

b) Provide information on the possibility for participating teams to 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.

Evaluation metrics will be provided for participants to self-evaluate their metrics on a self-selected validation set. During the validation phase, participants will be given instructions to self-evaluate their dockers as a sanity check before the final submission. Only the last submission by each team will be considered for participation in the challenge.

### Challenge schedule

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

- the release date(s) of the training cases (if any)
- 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

The training data release will be conducted in three intervals: 1) the first batch of the training data will be released mid April 2) the second batch will be released mid June 2024, 2) the third and final batch of the training data will be released mid July. Participants can join the challenge until 10 days prior to the deadline. The submission deadline is the 15th of September 2024. Results will be released during the workshop-challenge day.

Release of training data:
15th April 2024 (1.Batch = 250 videos)
15th June 2024 (2. Batch = 250 videos)
15th July 2024 (3. Batch = 200 videos)

Registration deadline: 10 days before deadline
Submission deadline: 15th September 2024

### 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).

Data donation was conducted under the condition of an established regional IRB at each data donor`s institution. Data Donors agreed to this in the terms and conditions on the data donation platform. All data is de-identified through the SAGES Video Acquisition Portal, which was developed by Surgical Safety Technologies. No separate

IRB is required for the execution of the CVS Challenge.

### 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.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY-NC-ND (Attribution-NonCommercial-NoDerivs).

### Code availability

a) 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.

We will provide the evaluation code to challenge participants alongside the final batch of the training data. This release is planned for mid-July.

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

Participants are required to encapsulate their inference algorithms and associated model weights into a Docker container. The challenge organizers will ensure that the participants' source code and model weights remain confidential and will not be disclosed publicly during the evaluation phase. After the challenge is finished, code and model weights from selected participating teams with leading scores will be made publicly available.

### 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.

The CVS Challenge is organized by the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) with financial support from the SAGES Foundation, Intuitive Surgical, Olympus, Medtronic Inc, and Surgical Safety Technologies. The industrial sponsors of the challenge will not have access to the test data for the duration of the challenge but may access the data once the full dataset is released to the public upon completion of the challenge.

## 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

Main: Research
Secondary: Intervention Assistance, Intervention Planning, Training

### Task category(ies)

State the task category(ies)

Examples:

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

Classification / Prediction

### Cohorts

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).

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

Both the challenge and target cohorts are patients undergoing laparoscopic cholecystectomy, a routinely performed minimally invasive surgical procedure to remove the gallbladder due to inflammation or other pathologies. A well established surgical safety measure, the Critical View of Safety (CVS) (Strasberg et al. J Am Coll Surg. 2010) targets the prevention of common bile duct injuries, which is a feared complication with significant impact to patients` morbidity and mortality. The CVS consists of three criteria which entail the dissection and isolation of three anatomic landmarks, which can be visually distinguished in the intraabdominal video obtained from the laparoscope. The video data is acquired using a variety of laparoscopes from different manufacturers, which adds to the diversity of the data and ensures adequate representation of the real-world population. Classification of these criteria is targeted in the CVS Challenge challenge.

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

The challenge cohort is identical to the target cohort.

### Imaging modality(ies)

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

Intra Abdominal endoscopic video data.

### Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

The videos and corresponding ground truth annotations of the CVS criteria (per video and per selected frame classification) will be provided to challenge participants.

b) ... to the patient in general (e.g. sex, medical history).

Additional information corresponding to the patient will be restricted to the procedure type (laparoscopic cholecystectomy). No additional information is provided to challenge participants besides intraabdominal endoscopic video data. The obtained demographics associated with the donated video data, which is provided at the discretion of the individual data donors, is solely accessible by challenge organizers and will be for data splits. Patients` sex or gender is not acquired and is irrelevant to the proposed question.

### Target entity(ies)

a) 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.

The data consists of intraabdominal video footage obtained through a rigid laparoscope.

b) 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.

The algorithm targets classification probabilities of the three criteria composing the CVS. These criteria, defined in the consensus-based Annotation Protocol, are C1) Two and only two tubular structures are seen connected to the gallbladder, referring to the Cystic Duct and Cystic Artery, C2) The hepatocystic triangle is cleared from fat and/or connective tissue so that unimpeded view is obtained and C3) The lower part of the gallbladder is dissected off the liver bed to expose the lower 1/3 of the cystic plate.

### 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 (see below, parameter metric(s)), 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 metric(s)).

Within the CVS Challenge we conduct four (4) subchallenges. Beyond base algorithm accuracy, these subchallenges will cater to known and crucial limitations of AI in general and AI for CVS estimation. We detail the relevant metrics for each subchallenge in the metrics section of this proposal.

Subchallenge A - CVS Achievement:
Subchallenge A will focus on the algorithmic performance in assessing the overall achievement of the CVS and the achievement three (C1+C2+C3) subcriteria (C1, C2, C3).

Subchallenge B - Uncertainty Quantification:
Subchallenge B will measure algorithmic uncertainty and confidence calibration by comparing the model confidence to the annotator confidence estimated by interrater agreement.

Subchallenge C - AI Efficiency:
Subchallenge C will test CVS performance but be limited to teams that attest to low computational needs for inference. We will selectively run the top candidate teams` algorithms to verify the low compute ('code of honor' / random testing approach). Thus the efficiency of the proposed AI models will be assessed, and judged by computational effort, testing, and training time. Efficiency will be parameterized according to memory, VRAM, and GPU flops constraints.

Subchallenge D - Domain shift:
Subchallenge C will measure robustness to domain shifts in different deployment conditions (demographics and characteristics displayed in the data). We will use a resampling approach to create variant test sets with carefully curated shifts in the distribution of the data along selected axes, along with a rebalanced training set. Performance changes will be measured according to the same criteria as Subchallenge A (the main challenge). The overall robustness measure will be set according to a 90-quantile over the performance measures achieved over a set of variant test sets. The selected variables to shift data distributions include: Laparoscope type, origin, fluorescent light (Indocyanine Green) use yes/no, robotic surgery yes/no, intraoperative cholangiogram yes/no, subjective video annotation difficulty as rated by annotators.

# DATA SETS

## Data source(s)

a) 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 data is collected from several different laparoscopic imaging systems self-reported by video uploaders. Major manufacturers (Stryker, Storz, Olympus, Medtronic, Intuitive, Conmed) are well represented in this dataset (>50 cases each). A subset of the videos was collected through a device reported as `unknown`.

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

All raw videos have gone through pre-processing to standardize the image encoding (h.264) and format (mp4).

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

The data was collected from a diverse set of sources (>60 institutions) spanning over 50 countries and 6 continents. Specific locations of the source data cannot be revealed due to privacy concerns but currently, approximately 30% of the collected data comes from Low to Middle-income countries per the last OECD listing.

d) 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).

This diverse, global dataset of laparoscopic cholecystectomy videos spans multiple techniques (robotic/manual) and different imaging modalities (e.g. white light, ICG), as well as intraoperative diagnostic measures (cholangiogram). The dataset exhibits negative as well as positive examples of the CVS and its individual subcriteria being achieved or not achieved.

## Training and test case characteristics

a) 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) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Every case corresponds to one specific laparoscopic cholecystectomy procedure. Each case is composed of a 90-second video segment, with the last frame of this clip corresponding to the instant when either the cystic duct or artery was clipped (or tied). With each case, frame-specific annotations of the three criteria defining the CVS will be provided as binary assessments every five seconds within this 90-second video segment.

b) State the total number of training, validation and test cases.

The total proportion of training to test videos will be approximately 700 and 300, respectively, following a standard 70-30 split. Participants will be free to use a proportion of the training set to create internal validation sets.

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

These proportions were selected to (a) provide a sizable training set to the participant to potentially push the state-of-the-art for this clinically meaningful endpoint (b) to keep the test set large enough to be further subsampled to test different aspects of robustness to distribution shifts.

d) 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.

The challenge organizers will ensure a balanced and adequate representation of the different demographics and characteristics (devices, imaging modalities etc.) in the training and test set. We will use a resampling approach to create variant test sets with carefully curated shifts in the distribution of the data along selected axes, along with a rebalanced training set. Performance changes for all challenges will be measured according to the same criteria as Subchallenge A (see metrics section of proposal).

### Annotation characteristics

a) 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.

Critical View of Safety annotation is inherently subjective. To account for the effect of annotation bias and the inherent subjectivity of the task, each included frame is annotated independently by three separate annotators. The resulting interrater agreement will be used as reference annotation. Annotators are assigned to a specific case at random.

b) 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.

Prior to the annotation process, the dataset was preannotated by the organizing team, with the purpose of (1) excluding videos, which did not meet the inclusion criteria (no alternative operative approaches e.g. Fundus first cholecystectomy, no incomplete or aborted procedure, no conversion to open surgical, no out of body images in the selected 90 second segment, and the first clip applied to either cystic duct or artery, and (2) to index the data according to clinical characteristics (i.e. robotic procedures vs manual, use of fluorescence / Indocyanine Green and execution of an intraoperative Cholangiogram), which were used for data splits for the later described subchallenges.

Annotators underwent a structured curriculum during which they were trained to assess the CVS and subcriteria using an annotation protocol. The CVS annotation protocol was refined from (https://arxiv.org/abs/2106.10916) in a three-round modified Delphi consensus among Key Opinion Leaders defining guidelines for Safe Cholecystectomy and members of the CVS Advisory Committee. For illustration, the protocol was supplemented with flashcards, depicting clinical examples, as well as instructional videos. The Delphi participants subsequently

annotated 60 videos each, the resulting interrater agreement served as the reference annotation for the `CVS Annotation School dataset`. Following a virtual onboarding session of 45 minutes, annotators were required to annotate the 60 videos in the `CVS Annotation School dataset` in a proficiency-based progression, until 70% interrater agreement was achieved with the predefined reference annotation.

c) 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.

All annotators are surgical residents. Annotators were recruited through an open call, screened based on participation in regularly scheduled calls, trained using the described annotation curriculum, and filtered based on performance on a pre-defined set of videos annotated following the annotation protocol.

d) 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.

The annotations from multiple annotators will not be explicitly merged in the training set to provide an additional supervisory signal to participants. Ground-truth annotations to compute metrics on the test set will be calculated using a majority vote of annotators except in the uncertainty subchallenge.

### 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.

All raw videos have gone through pre-processing to standardize the image encoding (h.264) and format (mp4). No other specific preprocessing will be performed before providing data to the participants.

### Sources of error

a) 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.

Annotator fatigue is particularly relevant to the generation of this dataset as the annotation period spans several months. This is being mitigated through gradual distribution of batches for annotation. Furthermore the subjective nature of CVS assessment can be regarded as a potential error source, which is mitigated by providing the reference annotation through interrater agreement of 3 annotators per datapoint. Other sources of errors include variability in the dataset for rare combinations of underlying factors and labels, imperfect estimates of the underlying distributions for the subchallenges B (uncertainty) and D (robustness). We will monitor these via resampling of the datasets.

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

See above.

## ASSESSMENT METHODS

### Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). 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)

Subchallenge A - CVS Achievement
Primary metric: Mean Average Precision (mAP)
Secondary Metrics: Per-class Counting Metric: Positive Likelihood Ratio (LR+) and Multi-class Counting Metric: Balanced Accuracy (BA)

Subchallenge B - Uncertainty Quantification
Primary Metric: Brier Score (BS)
Secondary Metrics: Regression accuracy on confidence and Negative Log Likelihood (NLL)

Subchallenge C - AI Efficiency
Primary Metric and Secondary Metrics: same as subchallenge A, with the main difference of a limited inference compute but methods that are self-certified to meet the eligibility criteria (determined based on GFLOPS/VRAM) will be considered.

Subchallenge D - Domain shift
Primary Metric and Secondary Metrics: same as subchallenge A.

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

Subchallenge A - CVS Achievement
We chose a multi-threshold metric (mAP) as our primary metric as different thresholds serve different clinical applications of these models (e.g. for reporting vs. for guidance). Secondary metrics were chosen to compensate for dataset imbalance.

Subchallenge B - Uncertainty Quantification
Brier Score (BS) addresses both intra-annotator agreement and average difficulty rating. The metrics Brier Score were chosen to measure uncertainty calibration. In case of additional emitted confidence, using regression accuracy will allow the assessment of model self-awareness of uncertainty.

Subchallenge C - AI Efficiency
See Justification Subchallenge A.

Subchallenge D - Domain shift
The same base metrics as for subchallenge A will be used, with the exception of using a norm that is close to L norm in order to check distributional robustness. The 90% quantile will be used to reduce outlier effects, and yet keep the maximum norm behavior needed for robustness reasoning.

## Ranking method(s)

a) 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.

Ranking is performed as customary in similar classification problems, and given the metrics leveraged on the subchallenges. All of our metrics result in fully-ordered, scalar-valued, evaluations, and we will rank accordingly, with respect to the primary metric in each subchallenge.

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

Incomplete submissions or submissions missing results will be disqualified to avoid the introduction of bias to the overall results.

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

For subchallenges A, B, and C, we calculate all the metrics at a video (case) level and then average across all considered cases. Ranking will be according to the average score. For subchallenge D, we will look at the quantile across variation subsets to gauge the maximum effect due to distribution shift.

### Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

We will assess the variation of the proposed methods to test statistical significance. We will compare between-class variability across selected variation subsets (across different cases difficulty, demographics, etc.). Benchmark analysis will be done using Python.

b) Justify why the described statistical method(s) was/were used.

A standard approach for benchmarking validation, with some testing of robustness due to domain shifts as sampled across variations in main factors for which we presume there`s enough data to characterize distributions.

### 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.

Depending on the number of submissions, we will verify whether ensembling proposed methods can help improve results, and inspect test cases that either exhibited poor overall performance across candidates, or high variability between candidate methods.

## ADDITIONAL POINTS

## References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

N/A

## Further comments

Further comments from the organizers.

N/A