Endoscopic Vision Challenge 2026: Structured description of the challenge design

Remark: This challenge has been slightly modified. All changes are highlighted in red.

CHALLENGE ORGANIZATION

Title

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

Endoscopic Vision Challenge 2026

Challenge acronym

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

EndoVis26

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.

With the advent of artificial intelligence as key technology in modern medicine, surgical data science (SDS) promises to improve the quality and value of the particular domain of interventional healthcare through capturing, organization, analysis, and modeling of data, thus creating benefit for both patients and medical staff. Holistic SDS concepts span the topics of context-aware perception in and beyond the operating room, data interpretation and real-time assistance or decision support. At the same time, minimally invasive surgery using cameras to observe the internal anatomy has become the state-of-the-art approach to many surgical procedures. Contributing to the key aspect of perception, endoscopic vision thus constitutes a central component of SDS and computer-assisted interventions. From this arises the necessity for high-quality common datasets that allow the scientific community to perform comparative benchmarking and validation of endoscopic vision algorithms. EndoVis (http://endovis.org/) organizes highprofile international challenges for the comparative validation of endoscopic vision algorithms that focus on different problems each year at MICCAI, comprising various computer vision tasks (classification, segmentation, detection, localization, etc) and subdisciplines ranging from laparoscopy to coloscopy and surgical training. It acts as an umbrella for several sub-challenges in this field, for MICCAI 2026 we propose 6 different sub-challenges within EndoVis.

Challenge keywords

List the primary keywords that characterize the challenge.

Surgical Vision, Endoscopy, Classification, Detection, Segmentation, SLAM

Year

2026

Novelty of the challenge

Briefly describe the novelty of the challenge.

EndoVis consists of different sub-challenges, while three of them are an extension of sub-challenges that took place last year (SurgVu, STIR, RARE) with novel tasks and data, three novel sub-challenges are proposed as well (TIGER SQ-AI, CLiMB, iMED).

Task description and application scenarios

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

Application scenarios within EndoVis2026 are ranging from detecting early-stage cancer in Barrett's Esophagus (RARE sub-challenge), tool localization as well as video question answering in robot-assisted surgery (SurgVu), tracking and reconstruction of surgical tissues in robot-assisted surgical videos (STIR), semantic segmentation for surgical quality assessment in laparoscopic esophagectomy (TIGER-SQ-AI), SLAM in coloscopy (CLiMB) and camera pose estimation as well as view synthesis in minimally invasive surgery (iMED).

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

NA

Duration

How long does the challenge take?

Full day

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

EndoVis acts as an umbrella and consists of several sub-challenges (6 in total this year). Each sub-challenge has specific challenge tasks itself and needs a least one hour to present the challenge.

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.

60-80 (based on the number of previous EndoVis challenges)

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 particular sub-challenge organizers.

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

Collaboration with European Society of Radiology (ESR)

In collaboration with European Society of Radiology (ESR), we announce special clinical interest topics with associated clinicians who can help with the preparation of the proposals; the best 3 challenge proposals on these topics will get the opportunity to present their challenges at the European Congress of Radiology (ECR) 2027 in a special session. If you want to organize a challenge in collaboration with ESR on one of these topics, please reach out to the MICCAI Challenges Team (miccai-challenges-2026@dkfz-heidelberg.de) and we will put you in contact with the corresponding clinician.

Challenge in collaboration with ESR. Ticking 'Yes' implies that the challenge has been prepared in collaboration with the clinical contact point.

No

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.

depends on the specific sub-challenges, e.g. Synapse/GrandChallenge platform for example, no on-site challenge, normal conference infrastructure on the challenge day (beamer, loud speaker, ...)

TASK 1: RARE2026: Recognition of Anomalies in low-pREvalance cancer

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.

Developing computer-aided detection (CADe) systems for cancer detection in low-prevalence scenarios presents a significant challenge. In clinical practice, early-stage cancers are often rare events, vastly outnumbered by normal or non-pathological findings. This inherent class imbalance makes it difficult to train models that are both sensitive to rare anomalies and robust to the overwhelming majority of normal cases. The subtle nature of early cancerous changes further adds to this challenge, as these anomalies are easily overlooked or misclassified. As a result, models trained on artificially balanced datasets often fail to generalize to real-world clinical conditions, where accurate detection of rare cases is critical for timely intervention and improved patient outcomes.
The challenge centers on developing a classification system that can accurately identify early-stage cancers while maintaining a balance between sensitivity and specificity. Improper evaluation of CADe systems during development can have significant consequences during clinical deployment. For instance, systems that are overly sensitive to rare anomalies may generate an excessive number of false positives. Conversely, systems that are not sensitive enough risk missing early cancer cases, delaying crucial interventions and adversely affecting patient outcomes. Striking the right balance between sensitivity and specificity is critical, particularly in low-prevalence settings, where false positives are more likely to dominate unless rigorously controlled during evaluation.
Detecting early-stage cancer in Barrett's Esophagus (BE) exemplifies these challenges. Subtle neoplastic changes in BE often go unnoticed during routine endoscopic surveillance, yet early detection is vital. Timely identification enables curative treatment through endoscopic mucosal resection, with long-term remission rates exceeding 90%. In contrast, missed lesions that progress to advanced cancer result in dire prognoses, with a five-year survival rate of approximately 15%. Despite these high stakes, the prevalence of early neoplasia during surveillance is exceptionally low, complicating data collection and model training. This gap underscores the need for CADe systems that effectively address extreme class imbalance while being rigorously evaluated.
As a reiteration of last year's RARE25 challenge, this edition extends the scientific scope by granting participants access to a large-scale unlabeled gastrointestinal dataset containing over 5 million images. This addition opens new research directions for tackling rare-event detection, including large-scale representation learning, self-supervised pretraining, generative augmentation of rare findings, and data mining for informative samples, thereby enabling investigation of how data scale and learning strategies influence performance in clinically realistic low-prevalence settings.

Keywords

List the primary keywords that characterize the task.

Computer-aided detection, Low-prevalence, Recognition, Early-stage cancer detection

ORGANIZATION

Organizers

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

Tim Jaspers, Cris Claessens, Francisco Caetano, Fons van der Sommen
(Eindhoven University of Technology, Department of Electrical Engineering, ARIA LAB)
Martijn Jong, Rixta an Eijck van Heslinga, Floor Slooter, Jeroen de Groof, Jacques Bergman (University of Amsterdam, Amsterdam University Medical Centers, Department of Gastroenterology and Hepatology)

b) Provide information on the primary contact person.

Tim Jaspers: t.j.m.jaspers@tue.nl

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

Yes, clinicians are part of the organizing team. Their roles included overseeing data collection and curation, as well as providing guidance on the selection of the most relevant performance metrics and the overall evaluation setup.

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 with annual fixed conference submission deadline

Challenge venue and platform

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

EndoVis26@MICCAI26

b) Report the platform used to run the challenge.

grand-challenge.org

c) Do you agree that the 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 won't 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

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

Part of EndoVis Challenge (https://endovis.org). Sub-challenge web still TBD.

Participation policies

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

Users are only allowed to curate their training data (including other public datasets e.g.)

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 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.

Data usage for training is not limited to the provided dataset and participants may also use open-source data and/or pre-trained networks, as long as these are accessible to all teams.

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.

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

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

We will provide several awards depending on the availability of sponsoring.

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.

A live leaderboard will be available throughout the challenge. The top three methods will be publicly announced, and the organization team will also recognize innovative solutions.

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).

Team member will be asked to participate in a shared publication and will be listed as author with sufficient contributions. Participating teams may publish their own results after the publication of the shared publication.

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.

Algorithm container submission (type 2) on grand-challenge. Submission instructions will be made available on the challenge website. Submissions should also include a short methodology report.

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.

A smaller validation set is provided, each team member is allowed to submit up to 20 times on this set. On the final test set each teams can only submit their final solution.

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

Release of training data: already released
Start Open Development Phase: 1st of April 2026
Final Test phase: Starting 1th of September 2025 (2 weeks to submit final model on the test set)
Challenge Day: Day of Endovis 2026

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 data is acquired under Dutch Trial Register number NL8411. Data collection from each center involved review of the data collection plan by the local institutional review boards or medical ethics committee. To maintain privacy and comply with data protection regulations, strict anonymization of images was carried out as described in the previous paragraphs. This process ensured that the data was no longer considered 'personal data' as defined by the General Data Protection Regulation (GDPR).

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)

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.

CC BY-NC (Attribution-NonCommercial)

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.

All evaluation code will be open-sourced on the challenge GitHub page

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

All code of the participating teams need to be shared with the organizers together with their methodology report.

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.

There are no commercial sponsors or funding involved in the challenge.
The availability of the test set labels will be limited to the organizing team from the Eindhoven University of Technology and the Amsterdam Medical Center.

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

Diagnosis,Screening

Task category(ies)

State the task category(ies)

Examples:

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

Classification

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.

Barrett's esophagus surveillance patients.

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

Endoscopist performing Barrett's esophagus surveillance.

Imaging modality(ies)

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

Endoscopy

Context information

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

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

class label (neoplasic lesion, non-dysplasic lesion)

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

No additional patient information will be provided

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.

Endoscopic image of Barrett's esophagus surveillance patients

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.

Neoplastic lesions in Barrett's esophagus patients

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)).

Neoplastic lesions in Barrett's esophagus patients

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).

Training and test data: Exera III (Olympus Corp., Tokyo, Japan)

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

Pretraining data:
The data was acquired retrospectively using an automated query from EndoBase, designed as a broad retrieval process to collect all stored images within the system. All images where first semi automatically anonymized and afterward a manual curation step was conducted. For further details on the acquisition, please refer to Jong et al. (2025)
Training data:
The data was acquired retrospectively using an automated query from EndoBase, designed as a broad retrieval process to collect all stored images within the system. From this dataset, all Barrett’s esophagus-related images were manually identified and extracted. No standardized data acquisition protocol was applied during the original imaging process, meaning that imaging parameters, techniques, and device settings may have varied depending

on clinical practices and operator preferences at the time of acquisition.
Test data:
The test data was acquired through two methods. A portion of the data was retrospectively retrieved from archives (EndoBase) using patient lists, which also included pathology results (PA) for each patient. Images were then manually reviewed to determine whether abnormalities were present in each case. The other portion of the test data was collected prospectively, adhering to our standardized prospective acquisition protocol. For further details on the test data acquisition process, please refer to Fockens et al. (2023).
Fockens KN, Jong MR, Jukema JB, et al. A deep learning system for detection of early Barrett's neoplasia: a model development and validation study. Lancet Digital Health. 2023;5(12):e905-e916. doi:10.1016/S2589-7500(23)00199-1.
Jong MR, Boers TGW, Fockens KN, Jukema JB, Kusters CHJ, Jaspers TJM, et al. GastroNet-5M: A multicenter dataset for developing foundation models in gastrointestinal endoscopy. Gastroenterology. 2025. doi:10.1053/j.gastro.2025.07.030.

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.

Pretraining data: 8 Dutch hospitals acquired between 2012-2020 (Jong et al. 2025).

The training set is constructed with data from: Medical Center Spectrum Twente, Medical Center Spaarne,

The test set, on the other hand, includes data from 12 medical centers that are part of the BONSAI consortium (Fockens et al. (2023)).

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).

The training data was acquired during routine endoscopic surveillance in two medical centers (Medical Center Spaarne, Medical Center Spectrum Twente) . No specific group of experts was designated for the data acquisition process, as it involved standard clinical practice. Consequently, we cannot provide detailed information regarding the expertise level of the medical professionals involved in acquiring the data.

The test data is acquired by expert endoscopist specifically focused on Barrett’s esophagus screening.

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.

A case in this challenge represents a single endoscopic image with the respective label nondysplastic or neoplastic for the training, validation and test set.

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

• Unlabeled training set: 4,820653 endoscopic images of +- 500,000 procedures throughout the whole Gastrointestinal track
• Training set: 2,937 nondysplastic images and 158 neoplasia images, representing a 5% neoplasia ratio.
• Validation set: 1000 nondysplastic images and 100 neoplasia images. This small set is designed solely for participants to verify that their submission process is working. It is not intended to serve as a reliable indicator of model performance. The evaluation methodology remains consistent with the test set.
• Test set: 23.216 nondysplastic images and 3200 neoplasia images.

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

100%

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

The training set contains predominantly non-dysplastic lesions due to their availability and prevalence, reflecting real-world scenarios. The test set mimics real-world prevalence but includes enough neoplastic images to evaluate model performance adequately. The validation set has the same ratio of neoplastic to NDBE cases as the test set but is smaller in size.

e) 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 class balance in the training set reflects the real-world distribution. However, the test set contains a higher number of neoplastic lesions to better evaluate the models' ability to differentiate between non-dysplastic and neoplastic lesions.
However for evaluation, we aim to simulate a real-world class distribution using a 1:100 neoplasia-to-nondysplastic ratio. To achieve this, we will:

• Select non dysplastic cases for each evaluation iteration.
• Sample 100 as few neoplastic cases with replacement to match the target ratio.
• Repeat this process 1,000 times, ensuring that the full extent of the test set is utilized.
• Calculate the median PPV@RECALL=0.9

This approach maintains the test set's full resolution while effectively reflecting a real-world deployment scenario.

f) 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.

For this challenge, the training dataset is identical to that used in the previous edition. At the time of the last challenge, this dataset consisted entirely of new, previously unseen, and unpublished cases. No additional data have been added since then.

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.

For training data, class label ground truth was based on expert assessment. Each image was evaluated by at least 1 assessor. In case of doubt, a second assessor was consulted. Imagery was labeled as neoplastic if there was a visible lesion suspicious for early neoplasia, or as NDBE if there was no visible lesion.

For test data, class label ground truth was two-fold. First, similarly to the training data by visual inspection. In addition, histopathological assessment was included. Imagery was labeled as neoplastic if there was a visible lesion and histology of the endoscopic resection specimen revealed high-grade dysplasia or adenocarcinoma, or as NDBE if there was no visible lesion and all random biopsies were negative for neoplasia. All included patients were treatment-naïve for Barrett's esophagus.

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.

All annotators had extensive prior experience in data annotation through their involvement with a research consortium dedicated to developing AI for Barrett's esophagus patients for over a decade. Consequently, no additional instructions were deemed necessary.

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 have extensive clinical experience (4+ years) in the field of Barrett's esophagus endoscopy. Edge cases were evaluated by an expert endoscopist with over 20 years of experience with the diagnosis and treatment of early Barrett's neoplasia.

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.

N/A

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 raw training data undergoes minimal pre-processing before being provided to the participating teams. Specifically, black edges are cropped from the images to remove extraneous borders. No additional pre-processing techniques are applied to the data. If necessary, the same approach is consistently applied across training, validation, and test cases to ensure uniformity.

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.

In the training set class labels are based on visual examination by experience Barrett's esophagus researchers. Edge cases are evaluated by expert endoscopist with over 20 years of experience with the diagnosis and treatment of early Barrett's neoplasia. In the test and validation set all class labels are based on pathological assessment and no errors are expected.

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

N/A

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)

The primary and sole metric used for ranking is the PPV@RECALL=0.9 .

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

The primary goal of the system is to detect neoplastic lesions in images, where sensitivity (recall) is critical to avoid missing tumors, which could lead to undiagnosed or untreated cancers. Focusing on PPV at recall = 0.9 ensures high sensitivity while maintaining precision, reducing the risk of false positives and unnecessary tests.

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. Ideally, provide the ranking scheme as a concrete pseudo code.

Average PPV@RECALL=0.9 value over 1000 Bootstrap samples.

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

Any missing results will lead to the metric being set to 0.

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

Since the test set is still relatively small we opted to use a bootstrapping approach for generating the rankings across the methods.

Statistical analyses

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.

The primary statistical approach used in the challenge analysis is non-parametric bootstrapping for uncertainty quantification of performance metrics. Bootstrapping was chosen because it does not rely on strong

distributional assumptions and is well suited for evaluating model performance on finite and potentially imbalanced datasets

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).

confidence interval of the mean PPV@90RECALL on the test set computed using the 95th percentile bootstrap

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

Variability in algorithm performance across test cases is assessed using graphs, showing the confidence intervals.

Provide a description of how variability of rankings is assessed.

Ranking plots, generally, ranking stability will be analyzed by following the methods suggested in https://www.nature.com/articles/s41598-021-82017-6, including bootstrap based ranking variability and variability related to different ranking schemes.

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

Statistical significance of performance differences between submissions is assessed using paired Wilcoxon signed-rank tests on case-level performance metrics. This non-parametric test was chosen because it accounts for paired evaluations on the same test cases and does not assume normality of performance differences. Bootstrapping is used exclusively for uncertainty quantification and not for hypothesis testing.

Provide a description of the missing data handling.

Submissions containing missing data are not accepted by the evaluation system and automatically result in a failed submission. Participants are required to correct the issue and upload a complete resubmission; incomplete or failed submissions are excluded from all analyses and are not considered in the final evaluation.

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

Python (SciPy)

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.

Additionally, a novel analysis could evaluate the ability of models to localize anomalies, though this would require designing models explicitly for anomaly identification, which may be a future challenge focus.

TASK 2: SurgVu2026: Surgical Visual 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.

Growing at an exponential rate with well over a million cases performed each year, robotic assisted surgery (RAS) promises to transform surgical intervention. Both the nature of the data these cases produce, and the sheer amount of it, present altogether new possibilities for study. Pursuits such as the quantification of surgical performance, efficiency and tool choreography, OR resource planning, AI-guided surgical planning, and surgical data science in general can all exploit this new source of clinical data. Not surprisingly, machine learning techniques that can extract meaning from these vast amounts of data seem poised to play an integral role. With this goal in mind we invite the surgical data science community to take part in two challenges, utilizing the largest publicly available dataset released to date (840+ hours of data). The first sub challenge requires participants to build a model that localizes tools and their corresponding key-points in videos, using only tool presence data. The second sub challenge invites participants to build a video question and answering model, that can answer questions about the tools, actions and surgical steps contained with short video clips. By design, both sub challenges are repeated from last year, with the first sub challenge now in its fourth consecutive year. This deliberate continuity mirrors the approach taken by other landmark challenges, employing large datasets and complex tasks to foster sustained progress and meaningful impact within the surgical data science community.

Keywords

List the primary keywords that characterize the task.

Surgical data science, Surgical tool detection, Surgical activity recognition, Video question and answering

ORGANIZATION

Organizers

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

Aneeq Zia (Intuitive Surgical), Max Berniker (Intuitive Surgical), Rogerio Nespolo (Intuitive Surgical), Anthony Jarc (Intuitive Surgical)

b) Provide information on the primary contact person.

Aneeq Zia: aneeq.zia@intusurg.com

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

no

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 with annual fixed conference submission deadline

Challenge venue and platform

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

EndoVis26@MICCAI26

b) Report the platform used to run the challenge.

grand-challenge.org

c) Do you agree that the 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 won't 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

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

Part of EndoVis Challenge (https://endovis.org). Sub-challenge: https://surgvu26.grand-challenge.org/

Participation policies

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

Users (participating team members) are allowed to append the supplied training dataset with their own data (e.g. new labels, augmented frames, etc.), but this data must be made publicly available at the closing of the challenge

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 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.

Publicly available data is allowed

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.

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

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

for each task: 3 monetary prizes for 1st, 2nd , and 3rd place – depending on the availability of sponsoring.

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.

Complete leaderboard will be visible publicly on the challenge website. The top 3 teams for each tasks will also be announced within the prizes sub-page of the challenge website

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).

Team members will be asked to participate in a shared publication and will be listed as author with sufficient contributions. Participating teams may publish their own results after the publication of the shared publication.

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.

Algorithm container (type 2) on grand-challenge. Submission instructions will be made available on the challenge website. Submissions should also include a short methodology report.

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.

There will be preliminary testing phases for each task where the teams will be allowed multiple (up to 10) algorithm submissions for testing. The dataset within these preliminary testing phases will only be 10-15% of the actual testing data set. The final rankings of the teams will only be based on their algorithm performances on the final testing phase submission. Please check out an example preliminary and final testing phase leaderboards from one of our previous challenges below:Prelim testing - https://surgvu25.grand-challenge.org/evaluation/challenge/leaderboard/Final testing - https://surgvu25.grand-challenge.org/evaluation/final-testing-phase/leaderboard

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

Release of training data: 1st April, 2026
Start of preliminary testing phase: 1st June, 2026
Start of final test phase: 3rd August, 2026
New registrations deadline: 16th August, 2026
End of preliminary testing phase: 31st August, 2026
End of final test phase: 6th September, 2026
Report submission deadline: 20th September, 2026
Challenge Day: Day of Endovis 2026

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).

An existing Western IRB will be used

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)

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.

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

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.

The evaluation container github repo will be made public

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

The participating teams will be required to make their algorithm container generating code public through a github repo maintained by the organizers, for examples see:
https://github.com/isi-challenges/surgvu2025-category1-submission
https://github.com/isi-challenges/surgvu2025-category2-submission

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.

Only the organizing team (all within Intuitive) will have access to test cases and labels, hence there will be no conflict of interest with any other institution. All awards will also be sponsored by Intuitive, while any team from within Intuitive wanting to participate will not be eligible for any award.

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

Research

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization

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

Detection,Localization,Tracking,Classification,NLP

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.

Surgical tasks performed on porcine model by trainees during robotic surgical training. Tasks include suturing of different styles (1-hand, 2-hand, running), and dissection performed on various anatomy (uterine horn, rectal vein/artery, etc.). Tools include (but are not limited to) graspers, needle drivers, scissors, staplers, clip appliers, and energy instruments

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

Surgical tasks performed on porcine model by trainees of various skill levels during robotic surgical training

Imaging modality(ies)

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

Single channel of endoscopic video

Context information

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

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

Surgical tool detection: ground truth tool presence labels.Surgical step recognition: the training videos will come with ground trutch tool presence labels, step labels, and descriptions of the steps, along with example question and answer pairs. The testing data will have verified correct natural language responses.

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

n/a

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 will be acquired from basic tasks being performed on a porcince model using a da Vinci Xi or Si system

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.

Surgical tool detection: Prediction of surgical tool bounding boxes.Video question and answering: Reasonable answers to questions posed on short video clips.

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)).

Tool detection: For bounding box detection, the assessment will be done using mean average precision over multiple intersection-over-union (IOU) values - this metric is standard for COCO dataset. Video question and answering: For question and answering assessment is done using an ensemble of BLEU, ROUGE, METEOR and BERT-based scores. These metrics are standard for similar problems.

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 Intuitive Data Recorder (IDR) will be used to capture video at 720p and 30fps from one channel of the endoscope on da Vinci Xi or Si system

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

The videos will be captured at 720p and 30fps from one channel of the endoscope on da Vinci Xi or Si system

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.

Data will be collected at Intuitive Surgical training labs

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).

Experience of study participants will mostly be beginners (early in their learning curve) with a few experts (practicing surgeons) if possible.

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.

A case in this challenge will comprise of a video of a surgical training being performed on a porcine model. These videos will be long and variable in length where multiple surgical training steps will be present in each video along with different surgical tools. The exact tools and steps present in the challenge dataset are given below (this information is also available in greater detail in our dataset publication and will be available to the teams):Tools: Needle driver, monopolar curved scissor, force bipolar, clip applier, cadiere forceps, bipolar forceps, vessel sealer, permanant cautery hook/spatula, prograsp forceps, stapler, grasping retractor, tip-up fenestered grasper.Steps: Suturing, Uterine horn, Suspensory ligaments, rectal artery/vein, skills application, range of motion, retraction and collision avoidance, other.VQA: Descriptions of surgical steps, including relevant anatomy, actions taken and tools potentially used.

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

We will have 200+ cases for training and 50+ for testing. We will ensure variability in the dataset through the variety of tasks completed on the porcine model on different anatomy. Each case has an average length of around 4 hrs.

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

All 280 video clips of the training dataset (≈840 hours, ~18M frames) are available for both Category 1 & 2 and include:

Surgical step labels
Tool presence labels
Surgical step descriptions

Due to the large scale of the dataset, the data is not split using fixed train/validation percentages. Instead, the full annotated dataset is made available for training, while small evaluation sets are used for final model assessment.

Category 1 — Tool Detection

Validation / Test Dataset (Hidden from Teams)

Explicitly annotated for evaluation and separate from the training data
Includes:
Tool bounding box annotations
Videos downsampled to 1 frame per second
8 of the 12 surgical tools
Covers 77 surgical tasks

Category 2 — Vision–Language (Q&A;)

Validation / Test Dataset (Hidden from Teams)

Consists of 101 video clips, each 30 seconds long
Includes ground-truth answers for Q&A; evaluation
Q&A; annotations are not available for all videos and are provided only for this dedicated evaluation set

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

We will try to ensure that the dataset has a balanced range of different tools and surgical steps within the training and testing set. We expect our dataset to have around 10+ unique tool labels with unequal distribution across classes (as some tools occur much more often than other e.g needle driver) while the surgical step distribution will be much more balanced across training and testing sets.

e) 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.

We will try to ensure that the dataset has a balanced range of different tools and surgical steps within the training and testing set. We expect our dataset to have around 10+ unique tool labels with unequal distribution across classes (as some tools occur much more often than other e.g needle driver) while the surgical step distribution will be much more balanced across training and testing sets.

f) 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.

Test data is unseen for all categories. More details in the answer above.

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.

We will use a crowd (5+ annotaters) to annotate tool bounding boxes for our testing set. The annotations will not be redundant as bounding box annotations are not that subjective. The surgical step annotations will be done by a team of domain knowledge experts for training and testing sets.

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.

For test set annotation, the crowd-sourced annotators were already trained and experienced in spatial annotation for surgical tools. Each frame will be annotated then reviewed by the annotation team to ensure quality. Bounding box labels will be placed around the surgical tools along with an object ID for object tracking. Additional tool classification label, such as left or right side will also be annotated. For surgical steps, the annotators will be provided by clear step starting and ending times for annotations.

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.

Annotators will have significant experience in labelling bounding boxes for surgical tools and surgical steps.

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.

n/a

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.

Raw video frames will not be altered. For Category 2, VQA, multiple sets of question and answer pairs were generated by hand using surgical description tasks and accompanying video clips.

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.

Image annotation will only be needed for the test set. Main sources of error would include the bounding box not being 'tight' around the tool. Its hard to estimate the error quantitatively but we don't expect it to be more than 5% For surgical steps, there can be some sources of error as this type of annoation is temporal in nature. However, as the robotic surgical training steps are very well defined (as opposed to steps in clinical procedures), we do not expect there to be any significant error in annotations (<5%). For VQA question and answer pairs were manually created and we do not expect any significant errors.

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

The tool presence labels will be generated using the events stream from the da Vinci system. There is a possibility of a dropped event that can cause error in the training tool presence labels. However, we do not expect this to happen frequently. The step labels would not have any other relevant source of error.

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)

Tool detection: Mean average precision (mAP) for different intersection over union (IoU) values 0.50:0.05:0.95 will be used to assess performance of tool bounding box prediction algorithms.Video question and answering: the BLEU, ROUGE and METEOR scores will be used to assess the accuracy of questions. In addition, a BERT-based score will be employed to add additional coverage in the event there is disagreemnt between the various metrics.

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

Tool detection: This is the standard metric used for bounding box prediction algorithms (and is also the COCO primary challenge metric). By varying the thresholds, this metric provides a more thorough evaluation of tool localization/keypoint detection accuracy. The surgical step category is a standard classification problem where average f1-score can accurately measure the performance of the models. This metric has proven to be a good measure of model performances in our previous challenges as well.Video question and answering : the BLEU, ROUGE and METEOR scores are standard ofbjective functions to assess natural language models. Using a BERT embedding for semantic encoding is also popular. These multiple approaches will be used to mitigate the known issues with the NLP objectives. They are generally accepted and used to compare model fits and goodness.

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. Ideally, provide the ranking scheme as a concrete pseudo code.

The performance rank will be based on the rank of the evaluation metric (e.g mAP IoU 0.5:0.05:0.95 for bounding box detection, e.g BLEU, ROUGE, METEOR, BERT scores for natural language responses) - the higher the value of this metric, the higher the ranking of that team will be.

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

Missing results will be penalized and no score will be given for those cases

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

Using the standard metrics being used within the object detection and question and answering research seems like the right way to rank teams. The spatial detection metric tests the algorithms for detection of objects of different sizes (for tool detection category) which will be useful in differentiating high and low performing teams. Similarly, BLEU, ROUGE, METEOR and BERT scores reward and penalize the predictions for correct/incorrect responses accordingly.

Statistical analyses

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.

For category 1 of the challenge, Tool detection: Mean average precision (mAP) for different intersection over union (IoU) values 0.50:0.05:0.95 will be used to assess performance of tool bounding box prediction algorithms. For category 2, Video question and answering: the BLEU, ROUGE and METEOR scores will be used to assess the accuracy of questions. In addition, a BERT-based score will be employed to add additional coverage in the event there is disagreement between the various metrics.

Algorithm performance is evaluated by computing the task-specific metric independently for each test case/video clip and then computing a mean. Measures of metric uncertainty for each algorithm, and statistical significance across algorithm comparisons can be obtained using the many sample metric values (which we can assume are independent).

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).

The precision of algorithm performance will be assessed using the variability of metric values across the fixed test set. For each algorithm, the mean performance the standard error will be computed. Approximate 95% confidence intervals for the mean are computed assuming independence between test cases.

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

As noted above, variability in algorithm performance is quantified by using the sample values obtained on the test set and computing their spread via the SEM and a 95% confidence intervals of the sample mean.

Provide a description of how variability of rankings is assessed.

Preliminary algorithms rankings will be made using respective means. The robustness of the ranking will be examined by comparing the statistical differences between neighboring algorithms within the ranking (e.g. the difference between algorithms ranked 3rd and 4th). This can be done with comparisons of the 95% confidence intervals, and t-tests.
Generally, ranking stability will be analyzed by following the methods suggested in https://www.nature.com/articles/s41598-021-82017-6, including bootstrap based ranking variability and variability related to different ranking schemes.

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

Two tests are at our disposal. Pairwise algorithm comparisons are assessed by computing differences and reporting the mean difference together with its 95% confidence interval. Differences whose confidence intervals exclude zero are interpreted as meaningful. Paired t-tests can also be performed.

Provide a description of the missing data handling.

If an algorithm fails to produce an output for a given test case, that case is assigned a score of zero for the corresponding metric. This policy is applied uniformly across all teams and reflects a failure to provide a valid prediction rather than missing data in the dataset.

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

All evaluations are performed using standard Python libraries including scikit-learn and NLTK (Natural Language Toolkit) libraries.

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 analysis will be performed on the results to check for ranking variability. On top of the mAP across multiple IoU values for the first category, we will test ranking of the teams when using individual IoU values (e.g 0.5, 0.6, etc) instead of averaging over all. In addition to this, we will evaluate per class metrics for both challenge categories as well.For Category 2, we will use additional combinations of our chosen objective funcitons to understand the rank variability of algorithm performance.

TASK 3: STIR2026: Surgical Tissue Tracking Using the STIR (Surgical Tattoos in Infrared) Dataset

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.

This challenge shall quantify efficient methods for tracking and reconstruction of surgical tissues in videos. This challenge will help enable robust quantification of tracking and mapping methods over many scenes. This is essential to verify methods for use in image guidance and automation. Datasets that have been developed thus far either use rigid environments (not general), visible markers (visible to algorithms), or require annotators to label salient points in videos after collection (costly).
To address this gap, we would like to use a withheld component of our dataset (Surgical Tattoos in Infrared (STIR), https://dx.doi.org/10.21227/w8g4-g548) for a tissue tracking challenge. This component is separate from the STIR 2024 withheld data.
STIR includes manually annotated point and visibility labels across semi-densely sampled frames of in-vivo videos with stereo correspondences.
Submissions including efficient 2D tracking, 3D tracking, and dynamic neural radiance fields are welcomed. Submissions shall take in input locations for a video and output their position and visibility estimations for the annotated frames in each video. Error will be calculated between the estimated positions, and that of the ground truth.

Keywords

List the primary keywords that characterize the task.

Tracking, Deformable, Reconstruction

ORGANIZATION

Organizers

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

Adam Schmidt (Intuitive Surgical), Mert Karaoglu (ImFusion), Alexander Ladikos(ImFusion), Omid Mohareri (Intuitive)

b) Provide information on the primary contact person.

Adam Schmidt: adam.schmidt@intusurg.com

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

no

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 with annual fixed conference submission deadline

Challenge venue and platform

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

EndoVis26@MICCAI26

b) Report the platform used to run the challenge.

synapse for the challenge, IEEE DataPort for dataset hosting

c) Do you agree that the 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 won't 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

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

Part of EndoVis Challenge (https://endovis.org). Sub-challenge: similar to last year: https://www.synapse.org/Synapse:syn54126082/wiki/626617

Participation policies

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

No user interaction is allowed at any step.

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 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.

Participants can use any publicly available datasets, or publicly available (with weights) pre-trained neural networks for training their algorithms.

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.

Organization members may participate, but they will not be eligible for awards. They will be exempt from ranking, although listed on the leaderboard without numbering.

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

We will provide several awards depending on the availability of sponsoring.

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.

The methods and results will be announced publicly at the EndoVis Challenge at MICCAI 2026

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).

Team members will be asked to participate in a shared publication and will be listed as author with sufficient contributions. Participating teams may publish their own results after the publication of the shared publication.

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.

Docker container via synapse platform. Submission instructions will be made available on the challenge website. Submissions should also include a short methodology report.

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.

Members can only assess their methods on the openly available validation set. This can be done through the synapse website with a live leaderboard. Docker instructions and an example script will be provided for both the results submission and the benchmarking submission.

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

Release of training data, registration: March 31st
Submission instructions: April 15th
Validation runs will be opened for submission on synapse: May 30th
Submission deadline: August 20th
Report Submission Deadline: August 27th
Challenge Day: Day of Endovis 2026

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).

Ethics approval is not necessary.

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)

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.

CC BY (Attribution)

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.

The evaluation code is publicly available on github along with an example tracker.

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

Participants must share their code and models with the challenge organizers. Participants must agree to publish their code online after the challenge.

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.

Sponsorship will be provided by the organizers. Only the organizers will have access to the test data. Participants will not be able to access said data. The authors have no conflicts to report.

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 planning,Research,Assistance

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction

- Registration
- Retrieval
- Segmentation
- Tracking

Tracking

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.

The target cohort is humans with a focus on applications in tracking mapping and diagnostics in surgery. Additional details for any of the following answers can be found at: https://dx.doi.org/10.21227/w8g4-g548

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

The challenge cohort comprises in vivo and ex vivo stereo footage from porcine labs with many different tissue types.

Imaging modality(ies)

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

Rectified stereo RGB videos from a da Vinci Xi endoscope are used for algorithm testing. Labelled infrared stereo pairs are used for quantification.

Context information

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

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

Each stereo video pair includes calibration, and tracking labels as described in https://dx.doi.org/10.21227/w8g4-g548

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

none

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 target is surgical tissue in the abdomen, with final applications being in minimally invasive surgery. The challenge cohort only includes in vivo and ex vivo animal tissues, while the target cohort is human.

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 challenge has been designed to target performance on deformable surgical tissue of any type.

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)).

We would like to assess robustness and accuracy of algorithms under many different difficult scenes which include discontinuity of movement and tissue occlusion. We additionally will emphasize the importance of computational efficiency.

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).

For the data provided, the stereo video is captured using a da Vinci Xi endoscope (see https://dx.doi.org/10.21227/w8g4-g548)

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

The data acquisition is detailed in depth at https://dx.doi.org/10.21227/w8g4-g548 A da Vinci Xi endoscope is used to collect calibrated stereo data along with infrared images for the ground truth labels.

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.

Intuitive Surgical

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).

Refer to 21a for camera characteristics. As for the level of expertise, this data is collected by clinician engineers with hundreds of hours of experience using da Vinci robots.

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.

For training a case is an ex-vivo or in-vivo video clip with ground truth infrared start and end points as described in details in https://dx.doi.org/10.21227/w8g4-g548. Training and test cases are the same in structure apart from being captured on different scenes. Test cases will only be usable through the grand challenge interface.

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

There are >1000 training cases, and ~10 test cases. Validation will have ~100 cases (STIR 2024 and 2025 Test sets).

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

The training and validation sets are annotated and respectively consist of >1000 and around 100 cases (STIR 2024 and 2025 Test sets). The test sets shall be around 10 cases.

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

The training and test cases are separated to provide a large amount for training while still leaving a reasonable size for robust evaluation in testing and validation. The total number of cases is limited by the number of labelling experiments we performed.

e) 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 class distributions of training and test are the same since they are sampled equally. We would like to quantify the performance of methods in environments like those they are trained for.

f) 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.

Although the training and validation sets will consist of previously published data, the test set will consist of unseen and unpublished data.

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.

For the training and validations set, no annotators were used per-se. One person is used for filtering outlier data (IR frames that did not include labels/etc).For the test set, annotations are done manually using a custom GUI

algorithm setup using ImFusion Suite. Each label is manually annotated and verified by multiple individuals from the organization team.

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.

For the training and validation sets, detailed instructions and protocol are in the dataset paper: https://dx.doi.org/10.21227/w8g4-g548.
For the test set, the annotator is instructed as follows:
1. Load the stereo video together with its camera calibration information. This step automatically rectifies each image.
2. In the starting frame, annotate points that are visible in the left image by selecting them on the left side of the image viewer using mouse clicks. Then, in the same order, annotate the corresponding points in the right image using the right side of the viewer. Ensure that all annotations are placed on tissue rather than surgical tools, and select points that are well distributed spatially and across different levels of motion. The algorithm automatically aligns epipolar lines to compensate for minor vertical misalignments, thereby preserving multi-view geometric consistency. If a mistake is made during this process, a point can be removed using the “alt” modifier.
3. Set the “Frame stride” configurator to the FPS of the input video so that navigating with the arrow buttons advances the video at one-second intervals.
4. For each frame accessed using the arrow buttons, including the final frame of the video, repeat the same annotation procedure. If a point becomes occluded in a given frame, use the “ctrl” modifier to mark it as invisible.
5. Use the table in the controller to monitor the annotation count for each frame and verify that no point labels are missing.
6. If it is necessary to restart the annotation process, use the “Clear” button to remove all annotations. To prevent accidental deletion, a confirmation pop-up is displayed.
7. Finally, use the controller to review all annotations and export them using the corresponding “Export” button.

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.

The data was filtered by an engineer who has worked with da Vinci systems for over 5 years.

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.

n/a

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 image filtering methodology that processes Infrared labels into center labels is described in https://dx.doi.org/10.21227/w8g4-g548

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.

The primary possible error is bulk tissue movement between the IR frame capture and visible light start, as mentioned in the paper at https://dx.doi.org/10.21227/w8g4-g548 . This possible error is the same for each split and should be <1mm.

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

No other error is expected.

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)

We will use <δ_avg^x average Jaccard (AJ) and occlusion accuracy (OA) as are introduced in the TAP-Vid Benchmark [3] which is used for evaluation of many modern point trackers [2, 4, 5]. This is chosen rather than multi-object-tracking metrics [1] which focus on multiple object regions.
OA measures the accuracy of the binary label of a point's visibility. <δ_avg^x measures the fraction of points closer than a specified threshold distance and is averaged over multiple thresholds. AJ considers both the tracking and occlusion accuracy together and is defined as the fraction of true positives (predicted points that lie within a specified threshold of the visible ground-truth point) divided by the sum of true positives, false positives (points predicted as visible while the corresponding ground-truth point is either occluded or farther than the threshold), and false negatives (ground-truth visible points that are predicted as occluded or whose predictions lie beyond the threshold). Similar to popular methods [2, 3, 4, 5], we choose thresholds of 4, 8, 16, 32, and 64 pixels. Median trajectory error [5] and mean trajectory error will also be reported but not used for ranking. 2D tracking methods will be evaluated in pixels, and 3D will be in millimetres. Each point's contribution to the metric is equal. The final metric is an average over the total number of points.

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

These were chosen as we are interested in physical accuracy for tissue tracking, so metrics such as IOU or association are not useful for points. We desire metrics that are commonly used and standard for point tracking. <δ_avg^xhas become the most used metric for this purpose from our survey of the point tracking literature.

[1] J. Luiten et al., "HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking," Int J Comput Vis, vol. 129, no. 2, pp. 548–578, Feb. 2021, doi: 10.1007/s11263-020-01375-2.
[2] M. Neoral, J. Šerých, and J. Matas, "MFT: Long-Term Tracking of Every Pixel," presented at the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 6837–6847
[3] C. Doersch et al., "TAP-Vid: A Benchmark for Tracking Any Point in a Video," in Advances in Neural Information Processing Systems, Dec. 2022, pp. 13610–13626.

[4] Q. Wang et al., “Tracking Everything Everywhere All at Once,” presented at the Proceedings of the IEEE International Conference on Computer Vision, 2023.
[5] Y. Zheng, A. W. Harley, B. Shen, G. Wetzstein, and L. J. Guibas, “PointOdyssey: A Large-Scale Synthetic Dataset for Long-Term Point Tracking,” presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 19855–19865.

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. Ideally, provide the ranking scheme as a concrete pseudo code.

Algorithms will be ranked based on their performance on average Jaccard. There will be separate rankings for best 2D and 3D algorithms. Aggregation will be performed over all points over all images. That means a video clip with 15 points will contribute 50% more to the metric than one with 10 points.

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

Submissions with missing results will be excluded from evaluation.

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

Ranking by the average Jaccard score jointly summarizes both the tracking accuracy and the visibility prediction performance of a tracker. These two components are essential properties of a tracking method, and this metric follows the standard evaluation protocol used in prior state-of-the-art benchmarks for similar tasks.

Statistical analyses

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.

The accuracy metrics (average Jaccard, occlusion accuracy) will be treated as per-sequence and per-method for further analysis of deviation and significance. The end ranking of the methods will be done using their average following prior benchmarks of the similar tasks in different fields.
The latency metric will be computed per-frame for each method. Warm up steps will be discarded and computed as the 95th percentile for outlier removal.

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).

Mean and median of the metrics will be computed for further analysis as a part of the follow up report and presentation.

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

The standard deviation, median and mean metrics for each method across different sequences will be reported.

Provide a description of how variability of rankings is assessed.

Bootstrap resampling will be computed and reported to assess how the ranking will change. This will demonstrate which test sequences are difficult.
Generally, ranking stability will be analyzed by following the methods suggested in https://www.nature.com/articles/s41598-021-82017-6, including bootstrap based ranking variability and variability related to different ranking schemes.

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

For the accuracy metrics, variation of each method's performance on each sequence will be computed and reported. Along with the ranking variability, this will showcase the significance of performance differences between each method.

Provide a description of the missing data handling.

The tracking accuracy metric (average Jaccard) considers tracking losses as a part of its thresholding based computation.

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

Statistics will be computed using Python and relevant packages such as SciPy and Matplotlib for visualization. ChallengeR may be used for post-challenge analysis.

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.

N/A

TASK 4: TIGER-SQ-AI2026: AI-based surgical quality assessment for thoracic lymphadenectomy in minimally invasive esophagectomy

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.

Survival outcomes after oncologic resection for esophageal cancer vary widely, with median survival ranging from approximately two to more than six years [1,2]. The extent and quality of lymphadenectomy represent key determinants of both accurate tumor staging and long-term patient prognosis [3-5]. However, the radical nature of lymph node dissection poses substantial risk due to the proximity of critical anatomical structures, potentially leading to complications such as major bleeding, lymphatic fistula, pneumonia, or chylothorax [6]. In clinical practice, considerable variability exists in how lymphadenectomy is performed and documented, limiting comparability across centers and hindering objective quality assessment. To address this gap, the multicenter TIGER study [7] has been initiated from UMC Amsterdam to establish a structured surgical quality assessment framework based on intraoperative video analysis. The primary endpoint of the TIGER study is the distribution of lymph node metastases in esophageal and esophago-gastric junction carcinoma specimens following transthoracic esophagectomy with at least 2-field lymphadenectomy, in relation to tumor histology, tumor location, invasion depth, number of lymph nodes and lymph node metastases, pre-operative diagnostics, neo-adjuvant therapy, and (disease free) survival. Lymph node stations are evaluated and categorized as resected, incompletely resected, not resected, or not assessable. Currently, this scoring relies on time-consuming manual review of the captured video by human experts, creating a bottleneck for large-scale data collection and analysis. Artificial intelligence (AI) and machine learning offer promising opportunities to automate this assessment process [8]. Automatic, vision-based detection of anatomical landmarks and classification of lymph node stations in surgical esophagectomy videos could lay the foundation for objective, reproducible, and scalable assessment of lymphadenectomy completeness. In the future, we may be able to correlate AI predictions with human expert SQA and clinical outcome data from the TIGER study to support clinical decision-making. We therefore introduce TIGER-AI, a fully pixel-wise segmented dataset of thoracic lymphadenectomy scenes from minimally invasive esophagectomies. This challenge aims to advance AI-based methods for medical image analysis, supporting more consistent surgical quality evaluation and ultimately improving patient outcomes.

Keywords

List the primary keywords that characterize the task.

Robotic Image Analysis, Laparoscopic Image Analysis, Full Scene Segmentation, Surgical Data Science, Surgical Quality Assessment

ORGANIZATION

Organizers

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

Sebastian Bodenstedt, Max Kirchner, Stefanie Speidel (NCT Dresden)
Johanna Brandenburg, Marie Daum, Jürgen Weitz, Martin Wagner (University Hospital Carl Gustav Carus, Dresden)
Suzanne Gisbertz, Dillen van der Aa, Sofie Henckens (UMC Amsterdam)

b) Provide information on the primary contact person.

Johanna Brandenburg: Johanna.Brandenburg@ukdd.de
Max Kirchner: Max.Kirchner@nct-dresden.de
Marie Daum (mail@mariedaum.de)

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

Yes, clinicians are part of the organizing team. Johanna Brandenburg and Marie Daum, both clinicians and surgical residents, contribute to project administration, data curation, conceptualization, investigation, and supervision of the participating teams, in close collaboration with the technical organizers Max Kirchner and Sebastian Bodenstedt (NCT Dresden). Martin Wagner and Jürgen Weitz, both clinicians and board-certified surgeons, together with Stefanie Speidel (technical, NCT Dresden), are supervising the project. In addition, Suzanne Gisbertz (esophageal surgeon), Dillen van der Aa, and Sofie Henckens are clinicians from the partner institution UMC Amsterdam and involved in data curation, conceptualization, and supervision.

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. The challenge will remain open until a sufficient number of submissions (at least 5 per task) is reached, after which a joined paper will be written and the data will be made publicly available as a dataset paper.

Challenge venue and platform

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

EndoVis26@MICCAI26

b) Report the platform used to run the challenge.

synapse.org

c) Do you agree that the 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 won't 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

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

Part of EndoVis Challenge (https://endovis.org). Sub-challenge web still TBD.

Participation policies

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

No user interaction is allowed at any step hence only automatic methods are allowed.

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 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.

Provided training data and publicly available data, including open, pre-trained networks, may be used. Private/non-public data (including additional annotations) or models pretrained on such are not permitted.

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.

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

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

We will provide several awards depending on the availability of sponsoring.

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.

Results of all teams will be first presented at the Endoscopic Vision Challenge meeting at MICCAI 2026. Afterwards, the information will be made available to all participating teams. The results will be made publically available in the form of a joined journal paper.

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).

Team members will be asked to participate in a shared publication and will be listed as author with sufficient contributions. Participating teams may publish their own results after the publication of the shared publication.

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.

Docker container via synapse platform. Submission instructions will be made available on the challenge website. Submissions should also include a short methodology report.

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.

Only the final submission for each team will be evaluated. To allow for sanity checks, the organizers will provide the participants results on data selected from the training dataset.

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

Release of training data (1st part): 1st of May
Release of training data (2nd part): 1st of July
Start of evaluation: 1st of September
Submission deadline: 15th of September
Challenge Day: Day of EndoVis 2026

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).

As the data consists of anonymized robotic and laparoscopic videos, no ethics approval is required.

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)

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.

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

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.

The script(s) for computing metrics and rankings will be made available with the second training data set.

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

To participate in the challenge, each team has to submit a Docker image capable of producing results on the testing examples. The Docker images will not be shared by the organizers. Participants are encouraged, but not required to make their code available as open source

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.

As sponsoring is still to be determined, no information regarding conflicts of interest can be provided. Only the organizers and some members of their institutions will have access to the test case labels.

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

Assistance,Surgery,Intervention follow up,Training,Research

Task category(ies)

State the task category(ies)

Examples:

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

Segmentation,Classification

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.

Patients undergoing laparoscopic or robotassisted esophagectomy (thoracic part only) for esophageal cancer.

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

Patients undergoing laparoscopic or robotassisted esophagectomies from participating surgical centers of the TIGER study.

Imaging modality(ies)

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

Robotic and laparoscopic video stream

Context information

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

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

Full scene segmentation image data in 31 categories for the 14 lymph node stations resection area in the thoracic part of minimally invasive esophagectomy, as well as merged labels in 15 categories.
Frame-level multi-label classification of lymph node stations visibility (multiple stations can be visible in a single frame).

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

None

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.

Thorax shown in laparoscopic or robotic video data.

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.

Full scene segmentation of 31 anatomical (i.e. lymphatic tissue, esophagus) and non-anatomical (i.e. instruments, other) structures, merged labels of 16 anatomical and non-anatomical structures, visibility of 14 lymph node stations. Detailed descriptions of the annotation classes including example images are provided in the annotation protocol

(Supplement 1). 27 anatomical (Lymph node, Trachea, Left Main Bronchus, Right Main Bronchus, Esophagus, Aorta, Azygos Vein, Superior Caval Vein, Pleura, Left Inferior Pulmonary Ligament, Right Inferior Pulmonary Ligament, Pericardium, Inferior Pulmonary Vein, Right Subclavian Artery, Right Vagal Nerve, Right Recurrent Laryngeal Nerve, Left Recurrent Laryngeal Nerve, Pulmonary Artery, Left Subclavian Artery, Right Bronchial Artery, Lung, Pool of Blood, Resection Area, Gastric Conduit, Fatty Tissue, Fatty Tissue Esophagus, Omentum, Thoracic Duct) and 3 non-anatomical (Background, Instrument, Other) labels are annotated.

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)).

Three subchallenges will be posed. 1. Develop an algorithm for semantic segmentation of the merged 16 anatomical and non-anatomical structures. 2. Develop an algorithm for semantic segmentation of the full 31 anatomical and non- anatomical structures. 3. Develop an algorithm to determine which lymph node stations are visible in a frame.

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).

Recordings from varying types of laparoscopic and robotic cameras.

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

Video collected during routine esophagectomies for esophageal cancer. After the end of the resection stage, an overview of all of the relevant lymph node stations is shown.

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 dataset is derived from 7 surgical centers from 5 different countries. 50% of the videos are laparoscopic (25 videos) and 50% are robotic (25 videos). This split is applied to both the training and the test videos.
One center (UMC of the Johannes Gutenberg University Mainz) was intentionally excluded from the training set to assess the model's performance on an independent center. This

center contributed three videos: one laparoscopic and two robotic.
Due to data availability, not all centers are equally represented in the dataset (see tables below). The centers are the following:

Amsterdam UMC (Netherlands)
UMC of the Johannes Gutenberg University Mainz (Germany)
Clarunis Universitäres Bauchzentrum Basel (Switzerland)
Centre Hospitalier Universitaire de Lille (France)
IRCCS Istituto Clinico Humanitas (Italy)
NIGU Niguarda Milan Hospital (Italy)
SANRA Osepedale San Raffaelo (Italy)

Full dataset:
Laparoscopic (25 videos) Robotic (25 videos)
UMC Mainz: 1 video
Amsterdam UMC: 5 videos
IRCCS: 9 videos
CHU Lille: 6 videos
SANRA: 4 videos

UMC Mainz: 2 videos
Amsterdam UMC: 13 videos
Clarunis Basel: 4 videos
NIGU: 6 videos

Training set:
Laparoscopic (20 videos) Robotic (20 videos)
Amsterdam: 4 videos
IRCCS: 8 videos
CHU Lille: 5 videos
SANRA: 3 videos

Amsterdam UMC: 12 videos
Clarunis Basel: 3 videos
NIGU: 5 videos

Test set:
Laparoscopic (5 videos) Robotic (5 videos)
UMC Mainz: 1 video
Amsterdam UMC: 1 video
IRCCS: 1 video
CHU Lille: 1 video
SANRA: 1 video

UMC Mainz: 2 videos
Amsterdam UMC: 1 video
Clarunis Basel: 1 video
NIGU: 1 video

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).

Surgeons for human surgical cases (laparoscopic or robotic esophagectomy).

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.

One case consists of 14 frames (one per lymph node station) from one human thoracic esophagectomy. The frames are clearly attributed to a specific lymph node station. However, multiple lymph node stations may appear in the same frame due to the anatomical closeness. In those 14 frames, full scene segmentation in 31 categories is performed. Additionally, merged labels in 15 categories will be released.

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

At least 50 cases in total: 40 training and 10 test cases.

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

In total, 140 images extracted from 10 videos have already been semantically segmented by a single rater, corresponding to 20% of the planned annotated dataset (50 videos, 700 frames). For the final dataset, all currently annotated frames are expected to be used as training data (40 videos, 560 frames), as these videos originate from one center (UMC Amsterdam). To date, no frames have been annotated for the test set test, which is planned to include 10 videos, 140 frames.

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

The total number of cases was chosen due to annotation effort with a 80% / 20% split for training and test data.

e) 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.

No further characteristics.
The distribution of classes in the data is the real-world distribution.

f) 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.

For this challenge only new, unpublished data will be used.

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.

We annotated the surgical boundary structures of thoracic lymphadenectomy during esophageal resection for esophageal cancer according to the TIGER classification and TIGER-SQA protocol [8] [9], yielding 31 categories.
Additionally, we merged these labels into 16 structures for a simplified version of the dataset. Merging of the classes was chosen based on closely related anatomical structures, which can clinically be seen as one functional unit (i.e. trachea, right bronchus and left bronchus).
By providing a simplified dataset, we wish to evaluate the performance gap of the algorithms trained on 16 vs. 31 structures and lower the threshold for participation and incentivize more people to apply. As a long-term goal, we aim to investigate whether the fewer, merged classes could also be sufficient to assess lymphadenectomy completeness. However, to address this future challenge, the human ratings of lymphadenectomy completeness from the TIGER-SQA study are required.
For the current challenge, one human annotator (advanced surgical resident) will annotate the training and test cases, a validation team (consisting of another surgical resident and an experienced esophageal surgeon) will review the full-scene segmented frames. To mitigate the single rater bias, the validation team will perform iterative reviews and quality checks at several stages throughout the annotation process, rather than evaluating the complete dataset only at the end. Feedback will be discussed directly with the annotator, and necessary adjustments will be implemented both retrospectively for completed annotations

and prospectively for subsequent annotations.
The visibility of lymph node stations will be annotated on a frame-wise level as a multirater annotation by two annotators (either two surgical residents or by one surgical resident and an experienced esophageal surgeon). Multiple lymph node stations can be visible in one frame.

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.

We established 31 categories for annotation based on the surgical boundary structures with detailed descriptions of the structures given in an annotation protocol (Supplement 1). For identification of the lymph node stations, a separated annotation protocol was developed (Supplement 2).

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.

Annotation is performed by an advanced surgical resident with four years of experience. The validation team consists of a surgical resident and an experienced esophageal surgeon. The challenge protocol was established in close cooperation with an experienced esophageal surgeon as well as a board-certified general surgeon and professor for surgical AI.

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.

N/A

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 training and test data consists of videos recorded from the endoscopic video feed during surgery and compressed using MPEG-4.

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.

Annotation errors include poor lighting conditions (over-/underexposure), motion artifacts, distortion structures covered by blood, smoke, or tissue. The label “other” includes all structures not otherwise defined by the annotation protocol, including structures where clear identification is not possible. Multiple lymph node stations can be visible in one frame. We do not expect relevant differences between the training and test cases.

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

N/A

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)

Segmentation tasks: The DICE similarity coefficient and the normalized Hausdorff Distance (nHD) will be used for ranking.
Classification Task: Performance is evaluated using two complementary metrics: the F1 score and the Area Under the Receiver Operating Characteristic curve (AUROC). To ensure balanced evaluation across all categories, both metrics are computed per class and then averaged over all classes (macro-averaging). The F1 score assesses the balance between precision and recall, while the AUROC provides a threshold-independent assessment of class separability.

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

Segmentation tasks:
The DICE coefficient (and the positively correlated IoU) is a metric used commonly in the segmentation and the surgical workflow community and is generally an accepted metric to assess the overlap between reference and predicted segmentation masks. For segmentation tasks the Hausdorff distance complements the DICE coefficient as it examines contour while the DICE coefficient examines the pixel overlap [10].
Classification task:
F1 is a suitable metric for multi-class image classification because it balances precision and recall, which is especially important when images may contain multiple classes and both false positives and false negatives matter. Computing F1 per class and averaging treats each class equally in a balanced setting, providing a clear and interpretable summary of performance across all classes in a multi-label context.
In addition, the AUROC is reported as a complementary, threshold-independent metric. AUROC quantifies how well the predicted probabilities of the positive class are separated from those of the negative class and thus reflects the model's ability to discriminate between class presence and absence across all decision thresholds.

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. Ideally, provide the ranking scheme as a concrete pseudo code.

Final performance rankings are determined through a hierarchical aggregation of weighted metric results. To prevent small but clinically critical structures from being statistically

overshadowed by large anatomical volumes, we employ a Weighted Macro-Average across N classes. Each class is assigned to one of three weighting tiers based on a clinical expert consensus:

Triple Weight (w=3): Small, high-priority clinical structures including Lymph nodes, Inferior Pulmonary Ligaments (L/R), Subclavian Arteries (R/L), Vagal Nerve (R), Recurrent Laryngeal Nerves (R/L), Pulmonary Artery, and Right Bronchial Artery.

Double Weight (w=2): Standard-sized, high-priority structures including the Trachea, Main Bronchi (L/R), Esophagus, Aorta, Azygos Vein, Superior Caval Vein Pleura, Pericardium, and Inferior Pulmonary Vein.

Simple Weight (w=1): All remaining categories such as Background, Lung, Instruments, Resection Area, Gastric Conduit, and various tissue types.

Aggregation Hierarchy

For the segmentation task, Dice and Normalized Hausdorff Distance (HD) are computed per class for every frame. These are aggregated into a weighted average per image, then averaged per video sequence (case), and finally averaged across the entire dataset. This "mean-of-means" approach ensures that every surgical case contributes equally to the final score, regardless of video duration. The classification task follows a similar macro-averaging (w=1) schema for F1-score and AUROC.

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

As participants submit Docker images for centralized evaluation, missing cases are not anticipated. However, to handle algorithmic hallucinations or missing predictions: if a class is absent in the ground truth but detected by the algorithm, the Dice/F1-score for that class is set to 0 and the normalized Hausdorff Distance is set to 1 (maximum penalty).

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

We utilize a weighted macro-average to prioritize small, clinically critical structures, ensuring that high accuracy on large volumes cannot mask significant failure cases on high-stakes targets. Using the mean rather than the median ensures that algorithms are held accountable for all outliers, rewarding consistent reliability across heterogeneous surgical environments. This rank-of-ranks approach ultimately identifies the most robust algorithm by requiring excellence across both overlap and boundary dimensions.

Statistical analyses

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.

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.
Our statistical analysis focuses on two primary objectives: assessing the robustness of the challenge rankings and determining significant performance differences between the competing methods. To achieve this, we employ a non-parametric workflow. First, hierarchical bootstrapping [11] is utilized to quantify the stability of the rankings for valid submissions, ensuring that the reported winners are robust to dataset variability. Second, the Wilcoxon signed-rank test is applied to statistically compare the performance differences between methods, offering a rigorous assessment of improvement without assuming a normal distribution of errors.

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).

Precision of the performance estimates is assessed by computing 95% Confidence Intervals (CI) for the mean of each metric (DSC, F1-score, AUROC, normalized Hausdorff Distance). This quantifies the uncertainty and reliability of the reported scores.

Methodology:
We employ non-parametric percentile bootstrapping to calculate the CIs. For each algorithm, 1,000 bootstrap samples are drawn with replacement from the test set. The lower and upper bounds of the CI are defined by the 2.5th and 97.5th percentiles, respectively, of the bootstrapped mean distribution.

Justification:
This method is selected because the strict penalties (e.g., DSC=0 for failures) create a non-normal, bimodal distribution of scores. Unlike standard parametric approaches, bootstrapping does not assume normality, ensuring robust precision estimates even in the presence of outliers and skewed data.

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

The variability and stability of algorithmic performance across test cases are assessed using a combination of parametric and robust statistical measures, supplemented by visual analysis.
Methodology:
We report the Standard Deviation (SD) alongside the mean for every metric. Additionally, the Median and Interquartile Range (IQR) are calculated. Box-and-Whisker plots are generated for each algorithm to visualize the performance spread, explicitly demarcating the core

quartiles from specific outlier cases.
Justification:
While SD provides a standard measure of dispersion, the Median and IQR are essential because penalty scores (e.g., 0 for failures) create a skewed, non-normal distribution. These robust metrics, combined with Box plots, allow us to distinguish between general algorithmic instability and specific edge-case failures that would otherwise distort a simple mean/SD analysis.

Provide a description of how variability of rankings is assessed.

Rankings variability is compared by computing 95% Confidence Intervals (CIs) using hierarchical bootstrapping.
Methodology:
Hierarchical Resampling: To account for data dependencies (e.g., multiple images per patient), we resample the test data at the cluster level (e.g., patient ID) rather than the individual observation level.
Interval Calculation: We will perform 1,000 bootstrap iterations to generate a distribution of the performance metric. The 95% CI is derived using the percentile method, identified by the 2.5th and 97.5th percentiles of the bootstrapped distribution.
Justification:
This method ensures that the confidence intervals correctly reflect the variability between subjects, preventing overconfident estimates that would arise from treating clustered data as independent. Hence, bootstrapping ensures insights on the ranking stability and variability.
Generally, ranking stability will be analyzed by following the methods suggested in https://www.nature.com/articles/s41598-021-82017-6, including bootstrap based ranking variability and variability related to different ranking schemes.

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

To assess whether differences in performance between algorithms are statistically significant, we utilize the Wilcoxon signed-rank test.
Methodology:
Pairwise Comparison: We perform pairwise comparisons between algorithms (e.g., comparing the top-ranked method against all others) using the observed performance metric values on the test cases.
Threshold: A significance level of alpha = 0.05 is used to determine statistical significance.
Justification:
This non-parametric test is chosen because the data is paired (all algorithms predict on the same test cases) and performance metrics often follow a non-normal distribution. The Wilcoxon signed-rank test is robust to outliers and does not assume a normal distribution of errors, unlike the Student's t-test. Furthermore, it provides information about the ranking stability because the result is whether one method is significantly better/worse than another.

Provide a description of the missing data handling.

Due to the containerized evaluation workflow (based on participant-submitted Docker images), the inference pipeline is executed automatically on all test cases; therefore, missing results are not anticipated. Scoring penalties are applied as follows:

False Positives: If a target class is detected by the algorithm but is absent in the Ground Truth, the Dice Similarity Coefficient (DSC), F1-Score, and AUROC score is penalized to 0, normalized Hausdorff Distance is set to 1.

Runtime Exceedance: We will set a runtime limit of 1 minute per frame. If the total time is exceeded, we will not count the submission as valid.

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

The data analysis framework is implemented exclusively in Python, utilizing the standard scientific ecosystem. Pandas is employed for structured data manipulation and feature engineering, while predictive modeling and performance benchmarking (e.g., Dice Score) are conducted using Scikit-learn or custom implementations, which will be made publically available. NumPy supports high-performance numerical computations, and visualizations are generated using Matplotlib and Seaborn.

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.

N/A

TASK 5: CLiMB2026: Colonoscopy Localization and Mapping Benchmark

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.

Colonoscopy is the gold standard for colorectal cancer (CRC) screening and prevention, enabling detection and removal of precancerous lesions. Despite its clinical value, navigating the long, tubular colon and maintaining reliable orientation while achieving complete mucosal inspection is difficult. A localization and mapping system could support quality assurance by highlighting regions that may not have been sufficiently inspected, assist navigation during procedures, and provide realistic training and simulation. However, accurate camera motion estimation and 3D reconstruction from colonoscopy video remain unsolved. Traditional approaches (both feature-based and direct methods) frequently break down due to occlusions, specular reflections, low-texture areas or sudden viewpoint changes.Learning-based methods can learn key components of the 3D reconstruction pipeline directly from data, such as depth estimation, feature extraction and matching, camera pose estimation, and scene priors, potentially improving robustness in colonoscopy videos. We propose a challenge to benchmark sequential 3D reconstruction algorithms (visual odometry/SLAM) in a real clinical setting, using a curated subset of EndoMapper with particularly challenging real colonoscopy sequences. The data will be shared with participants and later made publicly available. The objective is to accurately estimate camera trajectory, with an emphasis on accuracy and robustness across patients and conditions.

Keywords

List the primary keywords that characterize the task.

Colonoscopy, SLAM, VO, Pose estimation

ORGANIZATION

Organizers

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

Javier Morlana, Javier Rodríguez-Puigvert, Richard Elvira, Jose M. M. Montiel (Universidad de Zaragoza)

b) Provide information on the primary contact person.

Javier Morlana: jmorlana@unizar.es

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

no

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 fixed conference submission deadline

Challenge venue and platform

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

EndoVis26@MICCAI26

b) Report the platform used to run the challenge.

synapse.org or grandchallenge.org

c) Do you agree that the 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 won't 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

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

Part of EndoVis Challenge (https://endovis.org). Sub-challenge web still TBD.

Participation policies

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

We intend the challenge to be fully automatic at inference time on the test set. User interactions is allowed in any stage of training/validation, including using external/public data.

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 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.

Publicly available data is allowed.

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.

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

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

There will be cash awards and certificates for the winner and first runner-up, provided at least 3 teams submit the results - depending on the availability of sponsoring.

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.

The results for all teams will be announced during the EndoVis Challenge at MICCAI2026. The results will also be made publicly available on the sub-challenge website. The submitted results will also be presented publicly in the joint publication.

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).

Team members will be asked to participate in a shared publication and will be listed as author with sufficient contributions. Participating teams may publish their own results after the publication of the shared publication.

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.

Docker container via Synapse. Submission instructions will be made available on the challenge website. Submissions should also include a short methodology report.

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.

The participating team will not be allowed to evaluate their algorithms on the test data, which will be completely unseen and hidden. We will not provide results or leaderboard to participants before the challenge day. Only the last submitted docker container, output files and report will be used for the evaluation.

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

Challenge website/registration: 1st April 2026.
Training data release: 15th April 2026.
Team registration open until: 15th August 2026.
Test data release: 15th August 2026.
Submission deadline: 7th September 2026.
Report submission: 7th September 2026
Challenge Day: Day of Endovis 2026

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).

The data for all the tasks come from the EndoMapper dataset, which is fully anonymised and ethically approved. The recordings were made under the ethical approval of the CEICA Ethics Committee (Comité de Ética de la Investigación de la Comunidad Autónoma de Aragón (CEICA), meetings 04/03/2020 acta 05/2020, 23/09/2020 acta 18/2020, 20/04/2022 acta 08/2022 and 16/11/2022 acta 20/2022).

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)

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.

EndoMapper terms and conditions (https://www.synapse.org/Synapse:syn26707219/wiki/615178). "Our conditions for accessing the data are: 1) Limited to research on how to obtain relevant medical information from images or video. 2) Redistribution of the data is not allowed. 3) Requires a Statement of Intended Use, which includes a description of how you intend to use this data. 4) You further agree to cite the DOI of the collection and the publication in any publication resulting from this content as follows: Azagra, P. et al. Endomapper dataset of complete calibrated endoscopy procedures. https://doi.org/10.7303/syn26707219 (2022) Azagra, P., Sostres, C., Ferrández, Á. et al. Endomapper dataset of complete calibrated endoscopy procedures. Sci Data 10, 671 (2023).

https://doi.org/10.1038/s41597-023-02564-7. 5) Images of the collection can be included in the scientific citing publications. 6) Video segments can be used to produce multimedia material in the citing scientific publications. 7) The Universidad de Zaragoza will create a register of users of the dataset. The University of Zaragoza will store your Full Name, Synapse User ID, and Statement of Intended Use Any questions or concerns may be directed to The Unizar DPO at the following email address: dpd@unizar.es"

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.

Evaluations script will be made available via GitHub along with the detailed instructions on docker submission with dummy docker example.

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

Participating teams are encouraged (but not required), to provide their code as open access

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.

As sponsoring is still to be determined, no information regarding conflicts of interest can be provided. Only the sub-challenge organisers will have access to the test data labels. There is no conflict of interest.

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

Assistance,Screening,Research

Task category(ies)

State the task category(ies)

Examples:

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

Camera pose estimation, localization, SLAM

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.

Target cohort is real patient data

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

Challenge cohort is real patient data, from a single hospital (Hospital Clinico Universitario Lozano Blesa, Zaragoza, Spain)

Imaging modality(ies)

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

Colonoscopy

Context information

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

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

No further information other than image data will be provided

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

none

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.

Colon imaged in video with a monocular endoscope.

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.

Camera pose estimation and 3D reconstruction using sequential colonoscopy images

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)).

Accuracy of predicted poses on real colonoscopic video. Percentage of tracked frames. Runtime.

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).

Data is acquired with a colonoscope during CRC screening.

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

We use real colonoscopy videos from EndoMapper, acquired with a standard clinical monocular RGB colonoscope. For training/validation, we provide complete EndoMapper sequences unmodified. For testing, short clips were selected from routine screening segments where it is possible to obtain COLMAP pseudo-GT trajectories and reconstructions.

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 real data is based on a public dataset, Endomapper dataset. Azagra, P., Sostres, C., Ferrandez, A. et al. Endomapper dataset of complete calibrated endoscopy procedures.Sci Data 10, 671 (2023).

https://doi.org/10.1038/s41597-023-02564-7

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).

The data were acquired by trained clinicians during routine colonoscopy procedures. We further sub-sampled the video recordings into shorter sequences by targeting clinically and visually challenging segments, selecting videos of varying difficulty based on factors such as camera motion speed, texture level, and the presence of fluids.

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.

One case is a colonoscopy sequence. For training, a set of full EndoMapper sequences is available. Also external public data is also allowed. For test, a set of short subsequences will be used. The output expected is a camera trajectory and a 3D map. The evaluation is done by comparing the output results against a pseudo-ground truth trajectory.

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

About 50 EndoMapper sequences can be used for training/validation. We selected ~30–40 test cases (short clips) extracted from ~10–12 source sequences. To avoid data leakage between training and test, those source sequences cannot be used for training/validation.

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

Training/validation data is NOT annotated, but public data with other annotation (depth, poses) can be used. The test set will be 100% annotated, but the ground truth will be kept private for evaluation

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

EndoMapper dataset provides enough variation of trajectories, texture type, dynamic elements and deformation levels. Our test cases are balanced to evaluate the robustness and accuracy of the methods in a real endoscopy setting.

e) 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.

EndoMapper includes ~24h of real colonoscopy video, providing substantial variability in appearance, motion, lighting, and anatomy to develop colonoscopy-tailored VO/SLAM methods. Training is not strictly required

(classical methods can run without learning), but learned components (e.g., depth prediction, feature matching, pose priors) have shown improved robustness in challenging endoscopic conditions and can benefit VO/SLAM pipelines. Test cases are short clips curated from routine screening segments and selected to span a range of practical difficulties (e.g., texture scarcity, specularities, blur, occlusions) while remaining predominantly locally rigid, so that COLMAP/SfM pseudo-GT can be obtained reliably and evaluation remains stable and interpretable.

f) 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.

We plan to use EndoMapper sequences (no new data acquisition) with curated segments and COLMAP-derived ground truth annotations.

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.

Human annotator: 1. Automatic annotations were obtained via COLMAP, using an endoscopy-specific tuning that we found to be the most robust (although computationally costly). Test cases were selected from EndoMapper as segments where a colonoscopy VO/SLAM system is expected to work, covering a range of difficulties (e.g., low texture, motion blur, specularities). A segment was included only if a COLMAP reconstruction could be reliably obtained, which we validated via human visual inspection of the resulting reconstruction and trajectory quality.

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.

Test: the annotator was instructed to apply a global scale factor so that the reconstructed colon has an assumed average diameter of 50 mm, enabling metric-scale error reporting.

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.

Test: Pseudo-ground-truth trajectories and reconstructions are generated with COLMAP using an endoscopy-specific tuning. Inclusion/exclusion of test clips is based on human visual inspection of the resulting trajectory and 3D reconstruction to filter out obvious failures/artifacts (e.g., duplicated walls, strong drift/scale inconsistencies). The curators are members of the organizing team (technical researchers in VO/SLAM/endoscopic reconstruction; non-clinicians). In addition, a member of the organizing team manually sets the metric scale of each accepted reconstruction by scaling it to an assumed average colon diameter of 50 mm, so that pose error metrics are reported in metric units.

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.

N/A

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.

N/A

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.

Test: The main potential error sources are (i) SfM failure modes in colonoscopy such as low texture, specular highlights, motion blur, and occasional non-rigidity, which can lead to drift and scale inconsistencies, and (ii) sensitivity to imperfect camera calibration/photometric effects. Since true sensor ground truth is infeasible in real colonoscopy, we cannot provide an absolute error range in metric units. Instead, we mitigate and bound annotation noise through strict human curation: reconstructions are visually validated and any clip showing clear artifacts (e.g., duplicated walls, strong drift/scale inconsistencies) is excluded. Inter-/intra-curator variability is minimized by using consistent acceptance criteria and, when needed, re-review of borderline cases by multiple organizers. Additionally, metric scale is set by manually scaling each accepted reconstruction to an assumed average colon diameter of 50 mm. Since this is a single global scaling step, we expect the induced scale uncertainty to be small compared to major SfM failures, which are filtered out during curation.

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

N/A

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)

The task will be evaluated using ATE and a relative rotation error between the predicted camera poses and the visually verified COLMAP reconstruction (pseudo-GT). We also report RTF, the ratio of tracked frames (0 if no frames are processed and 1 if the full clip is processed), and T, the average runtime.

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

ATE and relative rotation error are standard metrics for camera pose estimation and SLAM/VO evaluation [1,2]. It is not possible to acquire true ground truth with the current setup for real colonoscopy data, but visually verified COLMAP (structure-from-motion) reconstructions can be obtained as pseudo-GT. A similar previous challenge reports relative motion errors (e.g., RTE/ROT), which are conceptually aligned with relative pose error components [3]. Additionally, RTF and T capture the robustness and efficiency needed for real data.[1] Sturm, Jürgen, et al. “A benchmark for the evaluation of RGB-D SLAM systems.” IROS 2012.[2] Ozyoruk, K. B., et al. “EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos.” Medical Image Analysis, 2021.[3] Rau, A., et al. “SimCol3D—3D reconstruction during colonoscopy challenge.” Medical

Image Analysis, 2024.

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. Ideally, provide the ranking scheme as a concrete pseudo code.

Lowest average ATE over all test clips, weighted by tracking coverage (RTF) and runtime (T). For each clip, the primary score is Score_ATE = ATE × W_rtf × W_t, and the final score is the average of Score_ATE across all test clips (lower is better). The weight for RTF is defined as W_rtf = 1 + 0.5 × (1 − RTF). The weight for runtime T is defined as W_t = 1 + 0.5 × min(1, max(0, T/fps − 1)), where fps is the video frame rate. Methods whose final Score_ATE differs by ≤5% (relative) are tie-broken using the analogous weighted rotation score Score_Rot = RotErr × W_rtf × W_t (lower is better).

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

Missing results will be treated as failures and penalized in the final score. For any test clip with missing/invalid output, RTF is set to 0 and ATE/RotErr are set to a fixed maximum cap, yielding a high weighted error for that clip and discouraging incomplete submissions.

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

It accounts both for accurate trajectory, robust tracking under colonoscopy challenging data and real-time performance.

Statistical analyses

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.

Each method will be evaluated per test sequence by running it 5 times (to account for stochasticity/non-determinism) and taking the median performance across runs as the sequence-level score. The leaderboard score will be obtained by aggregating the sequence-level scores across all test sequences, as mentioned in the metrics section (aggregate-then-rank). We will also report a brief robustness check of the ranking to reasonable aggregation variants (e.g., mean vs median across sequences). Ties will be handled by assigning the minimum rank (co-winners if applicable).

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).

Precision will be assessed via bootstrap confidence intervals over test sequences (resampling sequences with replacement) on the final aggregated leaderboard score.

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

Variability across test cases will be reported using distribution summaries across sequences (e.g., median and IQR of sequence-level scores) and visualizations (box/violin plots).

Provide a description of how variability of rankings is assessed.

Ranking stability will be assessed using bootstrap resampling of test sequences: for each bootstrap sample we recompute the aggregated score and ranking, and then report the distribution of ranks per method (e.g., probability of being top-k / rank histograms). This reflects sensitivity of the ranking to the specific set of test sequences.
Generally, ranking stability will be analyzed by following the methods suggested in https://www.nature.com/articles/s41598-021-82017-6, including bootstrap based ranking variability and variability related to different ranking schemes.

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

Differences between methods will be assessed using paired tests across sequences on the sequence-level median-of-5 score (paired permutation test on per-sequence score differences, or Wilcoxon signed-rank as a non-parametric alternative). If multiple pairwise comparisons are reported, we will apply a multiple-comparisons correction (e.g., Holm).

Provide a description of the missing data handling.

If an algorithm fails to return a valid output for a test case (e.g., no trajectory / invalid format), we will treat it as a failure and assign a predefined worst-score/penalty for that case, as discussed in the metrics section. We can also separately report failure rate as an additional metric.

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

We plan to run the analysis in Python, with scripts shared for transparency/reproducibility.

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.

N/A

TASK 6: iMED2026: Multi-Endoscope Dataset Novel View Synthesis Challenge

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.

Vision-based 3D reconstruction is an exciting application in robot-assisted minimally invasive surgery (MIS). For instance, Simultaneous Localization and Mapping can enable real-time guidance systems, improve both humans' and robots' surgical scene modeling, and actualize digital twins for intraoperative interventional systems. However, MIS presents extreme challenges: constrained viewpoints through narrow trocar ports, tissue deformation from respiratory motion and tool interactions, and non-Lambertian surfaces. Traditional methods like Structure-from-Motion require joint optimization of all frames, limiting real-time use, and struggle with featureless anatomical regions. Neural network training is hindered by small datasets with noisy pose ground truth from calibration and kinematics, while foundation models trained on natural images fail on surgical data. The iMED Challenge addresses these limitations through two tasks, pose estimation (PE) and deformable novel view synthesis (NVS), using synchronized dual-endoscope data from the full iMED dataset (340 sequences). The EndoVis challenge uses task-specific subsets: Pose Estimation (PE) uses 105 cases and Deformable Novel View Synthesis (NVS) uses 113 cases. Subsets and splits are defined at the anatomical-scene level, and no sequences appear in both tasks or across splits. iMED PE provides the first benchmark for camera pose estimation in deformable surgical environments, using ArUco markers fixed to anatomy for accurate relative poses. iMED NVS introduces the first deformable novel view synthesis benchmark with a train-on-one-endoscope, test-on-another protocol evaluating geometric understanding over photometric interpolation. Advancing robust 3D surgical vision will enable intraoperative guidance, autonomous camera control, and realistic simulation.

Keywords

List the primary keywords that characterize the task.

Pose Estimation, Novel view synthesis, Deformable reconstruction, Surgical endoscopy, Multi-view stereo

ORGANIZATION

Organizers

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

Sierra Bonilla (University College London), John Han (Vanderbilt University), Tianyi Song (University College London, Adam Schmidt (Intuitive Surgical) , Omid Mohareri (Intuitive Surgical), Francisco Vasconcelos (University College London), Sophia Bano (University College London)

b) Provide information on the primary contact person.

Sierra Bonilla: sierra.bonilla.21@ucl.ac.uk

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

We are exploring this option currently

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 with annual fixed conference submission deadline

Challenge venue and platform

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

EndoVis26@MICCAI26

b) Report the platform used to run the challenge.

Synapse for the challenge, UCL Public Research Database for dataset hosting

c) Do you agree that the 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 won't 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

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

Part of EndoVis Challenge (https://endovis.org). Sub-challenge web still TBD.

Participation policies

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

No user interaction is allowed during the testing phase, whereas user interaction is permitted during the training phase. Submissions must run fully automatically within the provided Docker environment. But the data preprocessing and curating are allowed during the training phase. But any data curating or preprocessing method must be reproducible and described in detail in reports.

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 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 entirety of the iMED training dataset will be publicly available by the time participants will work on the challenge. Any publication, preprint, or presentation that uses the challenge data (including methods developed or evaluated with it) must properly cite the original dataset paper and the iMED Challenge (Pose Estimation and Deformable NVS subtasks). Combination with external data: • Participants are allowed to pretrain or co-train their models on publicly available external datasets, provided that those datasets are independently accessible under their own licenses and are clearly documented in the method description. • Participants are not allowed to use private clinical data or non-public institutional datasets.

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.

Teams affiliated with the organizing institutes are allowed to participate in iMED Challenge under the following conditions: • Organizer-affiliated teams may submit methods and appear in the official leaderboard, but are not eligible for monetary prizes or sponsored hardware/software awards. • Organizer-affiliated teams are eligible to be listed in the top-performing methods (e.g., top 3) for scientific comparison and transparency, and their results may be shown on the public leaderboard, clearly marked as “organizer-affiliated”. • Members of organizer-affiliated teams who have been directly involved in data curation, annotation, test set construction, or evaluation pipeline implementation are not allowed to access unreleased test labels or intermediate evaluation results before the challenge is closed to all participants. • Any organizer-affiliated participation must strictly follow the same data access rules, submission limits, and deadlines as all other participants. • In any subsequent challenge paper or related publications, organizer-affiliated methods will be explicitly flagged as such, to avoid confusion with fully external submissions.

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

for each task: 3 monetary prizes for 1st, 2nd , and 3rd place – depending on the availability of sponsoring.

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.

Public leaderboard: • A public leaderboard will be maintained on the official challenge website. By default, teams will appear on the leaderboard under their chosen team name. • The leaderboard will display the official ranking metrics and aggregated scores for best submission.Top-performing methods: • The top 3 teams for each primary task will be announced publicly on the challenge website and during the EndoVis/MICCAI workshop. • Their method names, affiliations, and summarized results will be presented in the workshop and in any subsequent challenge overview paper (unless explicitly requested otherwise)

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).

Team members will be asked to participate in a shared publication and will be listed as author with sufficient contributions. Participating teams may publish their own results after the publication of the shared publication.

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.

Docker container via Synapse. Submission instructions will be made available on the challenge website. Submissions should also include a short methodology report.

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.

For both subtasks, the organizers will provide official evaluation scripts, dataset loaders, and baseline implementations so that participants can perform self-evaluation on public development data and submit intermediate results to a public validation leaderboard. Subtask-specific details (e.g., available sequences, metrics, and baselines) are described in the Assessment Methods section and on the challenge website.

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

Challenge website/registration/dataset release: April 15th
Participants may optionally submit docker containers beginning: May 30th
May 30th: Participants may submit results qualitative (video from second view) and quantitative (PSNR, SSIM) on public test sequences for inclusion on public validation leaderboard.
Submission deadline. August 20th
Report submission: September 1st
Challenge Day: Day of Endovis 2026

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).

N/A

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)

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.

CC BY (Attribution)

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.

Accessibility: PublicLink: GitHub repository (URL TBA)Supported platforms: Linux-based systems with Docker support, NVIDIA GPU required (tested on RTX 6000 Ada), CUDA compatible versions quoted on the website and githubContents: • Official evaluation script with PSNR, SSIM, Train/Render FPS metrics • Dataset loader for iMED • Docker template and integration guidelines • Baseline method implementations and example benchmark reproduction code • Train/test split definitions for all 340 sequences

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

Accessibility: Teams must submit a Docker container with their complete methodRequirements: • Docker image containing all dependencies and runtime environment • For Task 1, the Docker must implement get_pose() • For Task 2, the Docker must implement new_view(). • Must integrate with provided evaluation framework • Code must be runnable by organizers on unseen test scenesPost-challenge: Teams are encouraged but not required to open-source their code after the challenge paper publication

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.

Only the organizing team (within Intuitive and University College London) will have access to test cases and labels, so there will be no conflict of interest with any other institutions. All awards will also be sponsored by Intuitive, while any team from within Intuitive or Surgical Vision Group supervised by Sophia Bano or Francisco Vasconcelos at University College London wanting to participate will not be eligible for any award.

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

Assistance,Intervention planning,Research,Training

Task category(ies)

State the task category(ies)

Examples:

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

Reconstruction,Tracking,Modeling

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.

The target cohort consists of patients undergoing minimally invasive abdominal surgery (e.g., laparoscopic or robot-assisted procedures) in which accurate deformable scene understanding and novel view synthesis of soft-tissue anatomy and surgical tools can support intra-operative guidance, safe instrument manipulation, and advanced surgical training and simulation. The primary objects of interest are deformable abdominal organs (e.g., liver, kidney, bowel, surrounding soft tissue) and surgical tools interacting with them under realistic operating room conditions.

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

The iMED dataset comprises: Ex vivo: organ specimens imaged ex vivo, forming multiple anatomical scenes with camera motion, tissue manipulation, and tool–tissue interactions. In vivo post-mortem: anatomical scenes recorded on human cadavers, capturing human abdominal anatomy, tissue deformation, and tool usage. In vivo live: endoscopic sequences acquired from one live pig, providing physiologic tissue motion such as breathing or heartbeat-induced deformation, and tool-tissue interactions.Across all sessions, the objects include deformable abdominal organs, soft tissues, and surgical instruments. No subject-identifying information is present; only image data, pseudo-labeled anatomy and tool masks, pose estimations, and metadata required for iMED challenge tasks are provided.Task 1 – Pose Estimation (iMED PE): Uses the subset of sequences with moving cameras and reliable ArUco-based relative pose estimation between the two endoscopes. Task 2 – Deformable NVS (iMED NVS): Focuses on sequences with static cameras and moving anatomy for deformable novel view synthesis and reconstruction.

Imaging modality(ies)

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

The challenge data consist of multi-view, multi-session endoscopic imaging acquired with two stereo endoscopes:Dual stereo endoscopy:Each scene is recorded simultaneously by two stereo endoscopes, resulting in four synchronized video streams per sequence (Endoscope 1: left/right, Endoscope 2: left/right). The two endoscopes have overlapping fields of view, and their image streams are time-synchronized with an accuracy of approximately 3 ms.

Context information

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

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

For each sequence, a rich set of metadata is provided that is directly linked to the image data, including: • Acquisition category (ex vivo, in vivo post-mortem, in vivo live). • Session ID, scene ID, and sequence ID, reflecting the hierarchical structure (session / scene / sequence). • Detailed anatomical region description for each scene. • Activity type labels describing the dominant source of motion in the sequence: Camera motion (zoom-in,

translation, circular exploration), Tissue motion (e.g., off-camera manipulation, breathing/heartbeat-induced deformation), Tool–tissue interaction (e.g., cutting, mock suturing, coagulation). • Tool presence metadata, including the set of surgical tools present in each sequence, linked to a global tool taxonomy (~20 tool types) via tool IDs. • Indicators for the use of energy devices (e.g., electrocautery). • Links to relative camera pose information between the two stereo endoscopes, including calibration files and, where applicable (when camera moving), frame-wise pose trajectories. • Pseudo-labeled anatomy and tool segmentations

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

The dataset is anonymized and does not include individual-level clinical attributes such as age, gender, or medical history.

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.

For the intended clinical application, the image data originate from the abdominal or thoracic (pleural) cavity during minimally invasive laparoscopic or robot-assisted surgery.

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.

Task 1 – Pose Estimation (iMED PE): Algorithms focus on estimating accurate relative camera poses between the two endoscopes to enable multi-view 3D reconstruction and surgical navigation (e.g., SLAM). Methods must be robust to deformable tissue, specular reflections, limited texture, and constrained viewpoints.Task 2 – Deformable NVS (iMED NVS): Algorithms focus on deformable novel view synthesis and 3D reconstruction of soft-tissue anatomy and surgical tools. They must reconstruct geometry, appearance, and motion, maintain temporal consistency, and correctly handle occlusions, specularities, and complex tissue deformation.

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)).

Task 1 – Pose Estimation (iMED PE): Geometric accuracy of relative pose estimation between the two endoscope views, measured on per-frame trajectories under deformable surgical conditions.Task 2 – Deformable NVS (iMED NVS): Geometric accuracy and photometric realism of deformable novel view synthesis and 3D reconstruction, including temporal consistency and robust handling of occlusions, specularities, and tissue deformation.For both subtasks: Computational efficiency, including training and inference (rendering) speed, to assess the feasibility of deploying these methods in time-constrained surgical scenarios.

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).

Using Da Vinci 5 surgical robot and Da Vinci 5 stereo endoscope for data collection. Endoscopic image data were acquired using two synchronized stereo endoscopes per session, resulting in four video streams per sequence (Endoscope 1: left/right, Endoscope 2: left/right).Both endoscopes are clinical-grade rigid stereo laparoscopic/endoscopic cameras with visible-light RGB sensors. For each session, a specific pair of stereo endoscopes and corresponding calibration parameters (intrinsics and extrinsics for all four cameras) were used and documented.Endoscope 1 is mounted in a stationary configuration (fixed viewpoint), while Endoscope 2 is robot-mounted and actively moved during acquisition, also providing standard MIS illumination.

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

A da Vinci 5 endoscope is used to collect calibrated synchronized stereo data from two endoscopes with overlapping FOVs along with ground truth relative pose labels between endoscopes.

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.

At Intuitive Surgical, Inc. in Sunnyvale in a USDA-licensed and AAALAC International–accredited clinical lab using one porcine model and three human cadaver models sourced from a willed body program

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 dataset was collected by clinical engineers with hundreds of hours of experience using the da Vinci robots.

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.

For subtask (1) Pose Estimation: A case is defined as one synchronized stereo video pair between 10-15 seconds in length where one endoscope remains stationary while the other undergoes robotic movement (e.g., circular or scanning motion). Ground truth relative poses between endoscopes are computed from ArUco markers in the

scene and are provided for all frames. The ArUcos are then inpainted for training and testing. Training and test cases are structurally identical but captured on different anatomical scenes. Test cases are hidden and never visible to participants.For subtask (2) Deformable Novel View Synthesis: A case is defined as one stereo video clip between 10-15 seconds in length from endoscope 2 (left stacked on top of right) with ground truth camera pose of endoscope 1 relative to endoscope 2 and second stereo video clip from endoscope 1 as test view. Neither camera is moving, but the anatomical scene is moving. Training and test cases are the same in structure apart from being captured on different anatomical scenes. Test cases are hidden and never visible to participants.

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

The subset of the dataset that is used for each subtask is different. For subtask (1) Pose Estimation: 78 training, 21 validation, 6 hidden test casesFor subtask (2) Deformable Novel View Synthesis: there are 84 training, 23 validation, 6 hidden test cases

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

For both subtasks, 100% of our data is labeled. More specifically, all images in the pose estimation subtask are annotated with its corresponding camera pose and all images in the NVS subtask have a corresponding second camera view.

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

Subtask (1) Pose Estimation: The 74% / 20% / 6% split provides sufficient diversity for both classical feature-based methods (which require no training) and data-driven approaches (which benefit from larger training sets). The training set contains post-mortem, and live porcine sequences from 3 cadavers and 1 live porcine subject to ensure exposure to varied tissue properties and motion characteristics. The validation set (21 sequences) enables iterative method refinement and hyperparameter tuning. Each test case contains approximately 1200 frame pairs, yielding 7200 pose estimation instances across the 6 hidden test cases, providing statistically robust per-frame accuracy evaluation while maintaining held-out scenes to prevent overfitting.Subtask (2) Deformable Novel View Synthesis: The 75% / 20% / 5% split reflects the per-scene training paradigm of most neural rendering methods, where models are optimized individually for each sequence rather than across sequences. The training set includes ex-vivo, post-mortem and live porcine sequences from 10 ex-vivo organs, 3 cadavers and 1 live porcine subject respectively , exposing methods to diverse deformation patterns (tool-tissue interaction, respiratory motion, tissue manipulation). The validation set (23 sequences) supports per-scene optimization hyperparameter selection. Each test case contains approximately 1200 test frames, yielding 7200 novel view synthesis evaluations across 6 hidden test scenes for robust photometric quality assessment.Data leak prevention: Scenes are randomly sampled such that all sequences from a given anatomical scene appear exclusively in one split (train, validation, OR test—never mixed).
The data is split at the level of independent anatomical scenes rather than entire specimens. Each specimen contains multiple anatomically distinct regions (scenes) that are spatially and visually independent (e.g., different organs or separated anatomical areas). No anatomical scene used for testing appears in the training set.

e) 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.

Both subtasks use randomly sampled scenes with no data leakage—all sequences from a given anatomical scene appear exclusively in one split (train, validation, or test). All splits contain the full range of camera movement types (circular, lateral, zooming) and tissue motion patterns (tool-tissue interaction, respiratory motion, manipulation). Stereo camera pairs are positioned at 2.6cm ± 0.9cm baseline and 18.0° ± 6.9° angular separation

across all splits. Random surgical tools from a taxonomy of ~20 tool types appear across all splits. Test sequences contain only tools seen in the training set, ensuring methods are evaluated on in-distribution tool appearances. Test scenes are within one standard deviation of training/validation camera configurations and contain the same distribution of motion types, anatomical regions, and tool types, ensuring fair evaluation of generalization to unseen anatomical scenes rather than unseen imaging conditions.

f) 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.

N/A

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.

Subtask (1) Pose Estimation: Ground truth camera poses are derived from COLMAP structure-from-motion reconstruction. ArUco fiducial markers are placed in the surgical field to ensure sufficient feature correspondences for reconstruction in textureless surgical scenes. We validate the quality via reprojection error and ArUco marker dimension consistency. Any poor reconstructions (failed to converge) will use fallback opencv detect ArUco method.Subtask (2) Deformable Novel View Synthesis: Although there are many annotations for the dataset, this task uses the relative pose between the cameras. For every case for this subtask, a pre-calibration sequence was run using a custom external calibration system using custom calibration cubes with ArUco fiducial markers. ArUco markers are square fiducial tags with an internal bit pattern commonly used in robotics for pose estimation and unique identification. Each marker consists of an outer perimeter of black cells surrounding an inner data payload region. For these cases, ArUco cubes are moved around to densely sample the working volume, then removed before tissue manipulation sequences proceed with fixed cameras.No human annotators used. Same protocol applied to all splits.

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.

N/A

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.

The data was filtered by Sierra Bonilla who has worked on surgical robotic vision systems for 2.5 years.

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.

N/A

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.

Frame synchronization and organization: All four video streams (two stereo endoscopes: Endoscope 1 left/right, Endoscope 2 left/right) are temporally synchronized (≈3 ms accuracy) and organized into a consistent folder and naming structure (session / scene / sequence name).Endoscope matching:To simulate realistic illumination protocols within MIS, only endoscope 2 (the robot endoscope) had active illumination. Post-processing of gain and contrast was adapted for endoscope 1 to match endoscope 2. Camera calibration: For each session, intrinsic and extrinsic calibration parameters are estimated for all four cameras using ArUco-based calibration protocols. These calibration files (camera intrinsics, relative poses) are stored and referenced in the metadata.For static-camera protocols, ArUco markers are used during calibration sequences and are removed from subsequent “actual” sequences (no markers in the field of view).COLMAP pose estimation:Ground truth camera trajectories are computed using COLMAP with the following pipeline: SIFT features are detected in all frames from both endoscopes. ArUco markers in the scene provide dense, reliable features with known 3D geometry. Exhaustive matching is performed between the temporal frame pairs (consecutive frames from same camera), Stereo pairs (left-right frames from same endoscope at same time), Cross-endoscope pairs (frames from different endoscopes at same time). Initial camera poses are estimated from stereo pairs with highest feature correspondences. Additional frames are incrementally registered via PnP. Local bundle adjustment optimizes poses and 3D points in sliding windows. All camera poses and 3D points are jointly optimized to minimize reprojection error across all observations. Sequences with poor reconstruction quality (high reprojection error, insufficient feature tracks, or failed convergence) use openCV ArUco detection, RANSAC, and temporal filtering. For each synchronized frame pair between Endoscope 1 and Endoscope 2, the relative pose (rotation matrix R and translation vector t) is computed from the absolute poses estimated by COLMAP.

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.

For Subtask (1) Pose Estimation, COLMAP-based reference poses may be affected by scene deformation and specularities. We therefore apply an explicit quality control procedure before including a sequence in the hidden test set. For each sequence, we compute and report reprojection error statistics and stereo left–right pose consistency. Sequences exceeding predefined thresholds are excluded from the hidden test set. If COLMAP fails or does not pass quality control, we use a fallback pipeline based on ArUco detection with PnP, RANSAC, and temporal smoothing to obtain a stable relative pose trajectory. ArUco markers are digitally removed via inpainting for training and testing, which may introduce minor visual artifacts.
For Subtask (2) Deformable Novel View Synthesis, tool segmentations are pseudo-labels generated by a pretrained model and may contain boundary inaccuracies or missed detections. Camera calibration is performed prior to tissue manipulation, and small pose drift may occur during live sequences. We quantify pre- and post-manipulation drift and report summary statistics to contextualize residual calibration uncertainty. No additional major sources of error are expected.

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

No other error is expected.

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)

Subtask (1) Pose Estimation (iMED PE): We report the same pose metrics as CLiMB for consistency: ATE (Absolute Trajectory Error, translation) (Primary ranking), Relative rotational error (rotation-only component of relative pose error; ROT) (Primary ranking), % successfully registered frames/images – fraction of frames for which a valid pose is produced (Reported only) Subtask (2) Deformable Novel View Synthesis (iMED NVS): PSNR (Primary ranking), SSIM (Primary ranking), Training FPS (Reported only), Rendering FPS (Reported only)

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

ATE and relative rotational error are standard camera pose/SLAM metrics [1,2]. True ground truth is not available in our setup, so we evaluate against visually verified SfM/COLMAP reference trajectories supported by ArUco correspondences, consistent with prior
endoscopy SLAM benchmarks/challenges (e.g., translation/rotation components reported in SimCol3D) [3]. We additionally report % successfully registered frames/images to quantify robustness to tracking failures, which is critical for downstream surgical
navigation and reconstruction.
[1] Sturm, J., et al. “A benchmark for the evaluation of RGB-D SLAM systems.” IROS, 2012.
[2] Ozyoruk, K. B., et al. “EndoSLAM dataset…” Medical Image Analysis, 2021.
[3] Rau, A., et al. “SimCol3D—3D reconstruction during colonoscopy challenge.” Medical Image Analysis, 2024.

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. Ideally, provide the ranking scheme as a concrete pseudo code.

Subtask (1) : For each sequences: Compute ATE over all valid frames. Compute relative rotational error over all valid frames. Compute RTF (Registered Tracking Fraction) = number of valid pose frames / total frames.
Per-sequence metrics are first computed, then averaged across all test sequences. Primary ranking metric: Lowest mean ATE across all test sequences. Secondary reported metrics: Mean relative rotational error, Mean RTF (% successfully registered frames)
Runtime (T) is reported separately for transparency.
Subtask (2)): Compute PSNR per frame and average per sequence. Compute SSIM per frame and average per sequence.Per-sequence values are averaged across all test sequences. Primary ranking metric: Highest mean PSNR across sequences. Secondary reported metric: Mean SSIM across sequences. Training FPS and Rendering FPS are reported but do not affect ranking.For neural networks that overfit to a specific sequence (e.g. NeRFs), there will be a time limit for practicality.

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

All methods are required to provide some result. Otherwise they will be considered invalid. For subtask (1) Missing results will count as a failure case, adding a high error to the average. RTF will be set to 0 and ATE/ROT set

to a fixed max cap. For subtask (2) Any missing data will contribute 0 psnr and averaged into per frame PSNR values.

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

These are conventional metrics for pose estimation and novel view synthesis respectively. Although rendering speeds are important, we do not treat this as a main scoring metric to prioritize higher quality 3D reconstructions rather than computational efficiency.

Statistical analyses

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.

We report 95% confidence intervals via hierarchical bootstrap (1000 iterations). Each iteration: (i) resample test sequences with replacement; (ii) within each sampled sequence, resample frames with replacement; (iii) compute per-sequence metrics (ATE, ROT, RTF for PE; PSNR, SSIM for NVS); (iv) average per-sequence metrics across the resampled sequence set to obtain the iteration score and ranking.
Hierarchical bootstrap respects the nested structure (frames within sequences) and provides non-parametric uncertainty estimates on aggregate metrics and ranks.

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).

We assess the precision of performance estimates by computing 95% confidence intervals via percentile bootstrap resampling (1,000 iterations over the 6 test sequences). This non-parametric approach provides robust uncertainty estimates for mAA, PSNR, and SSIM without assuming a normal distribution.

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

To evaluate the consistency of algorithm performance across different surgical scenarios, we assess variability using standard deviation (SD) and interquartile range (IQR) for all per-sequence metrics. These quantitative measures are complemented by boxplots to visualize the distribution of results and identify potential outliers or specific failure modes as suggested in https://www.nature.com/articles/s41598-021-82017-6

Provide a description of how variability of rankings is assessed.

Our challenge contains two subchallenges, namely pose estimation and novel view synthesis. There will be two prize pools for each subchallenge. For the first subchallenge, there is only one metric (mAA) which will be the sole determinant for ranking.
For the second subchallenge, we use multiple image quality metrics, meaning all metrics should be aggregated to determine a team's score. We use a min-max normalization followed by arithmetic mean to calculate the score. For every metric, we normalize it between [0,1] using its minimum and maximum over all teams. Then we will calculate the mean of all normalized metrics for a given team to calculate their score.
Generally, ranking stability will be analyzed by following the methods suggested in https://www.nature.com/articles/s41598-021-82017-6, including bootstrap based ranking variability and variability related to different ranking schemes.

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

Although metrics are imperfect methods to comprehensively evaluate an algorithm's performance, we will not assess statistical significance between scores following other works in the literature.

Provide a description of the missing data handling.

We handle missing or invalid algorithm outputs as follows:
If a submission fails to return a valid output for a test case (e.g., missing trajectory, invalid format, or runtime failure), that test case is treated as a failure and assigned a predefined worst-case score (normalized score of 0) for the corresponding metric.
Submissions that produce all missing outputs or invalid files are rejected by the evaluation system and automatically marked as failed submissions.

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

We will programmatically implement all data analysis pipelines with Python and its common libraries, e.g. NumPy, PyTorch, Scikit-learn, etc.

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.

Using the hierarchical bootstrap samples, we will report rank variability (e.g., distribution of ranks per method and probability of appearing in the top-k). We will provide stratified results by acquisition type (ex vivo / cadaver / live porcine) on the public splits; hidden test is reported only as an overall aggregate.

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.

TIGER-SQ-AI2026:
REFERENCES
[1] Ovrebo et al. (2012). Long-term survival from adenocarcinoma of the esophagus after transthoracic and transhiatal esophagectomy. World J Surg Oncol. https://doi.org/10.1186/1477-7819-10-130
[2] Chen et al. (2017). Survival benefit of surgery to patients with esophageal squamous cell carcinoma. Sci Rep. https://doi.org/10.1038/srep46139
[3] Okholm et al. (2014). Status and prognosis of lymph node metastasis in patients with cardia cancer - a systematic review. Surg Oncol. https://doi.org/10.1016/j.suronc.2014.06.001

[4] Mariette et al. (2008). The number of metastatic lymph nodes and the ratio between metastatic and examined lymph nodes are independent prognostic factors in esophageal cancer regardless of neoadjuvant chemoradiation or lymphadenectomy extent. Ann Surg. https://doi.org/10.1097/SLA.0b013e31815aaadf
[5] Greenstein et al. (2008). Effect of the number of lymph nodes sampled on postoperative survival of lymph node-negative esophageal cancer. Cancer. https://doi.org/10.1002/cncr.23309
[6] Low et al. (2015) International consensus on standardization of data collection for complications associated with esophagectomy: esophagectomy complications consensus group (ECCG). Ann Surg 262:286–294. https://doi.org/10.1097/SLA.0000000000001098
[7] Hagens et al. (2019). Distribution of lymph node metastases in esophageal carcinoma [TIGER study]: study protocol of a multinational observational study. BMC Cancer. https://doi.org/10.1186/s12885-019-5761-7
[8] Lam et al. (2022). Machine learning for technical skill assessment in surgery: a systematic review. NPJ Digit Med. https://doi.org/10.1038/s41746-022-00566-0
[9] Henckens et al. (2025). Assessment of the extent of lymphadenectomy in esophageal cancer surgery in the observational TIGER study: [TIGER-SQA] study protocol. Art Int Surg. http://dx.doi.org/10.20517/ais.2024.47
[10] Maier-Hein, Reinke, Godau et al. (2024). Metrics reloaded: recommendations for image analysis validation. Nat Methods. https://doi.org/10.1038/s41592-023-02151-z
[11] Reinke et al. (2025). Current validation practice undermines surgical AI development (arXiv:2511.03769). arXiv. https://doi.org/10.48550/arXiv.2511.03769
SUPPLEMENTS

Supplement 1:
https://docs.google.com/document/d/1-WxFyD1sHrlrqyQ1Yq4-02ngZwgxyhKrbHDvY_yxM3c/edit?usp=sharing

Supplement 2:
https://docs.google.com/document/d/1GfBsblzTvPRpX4AOHwSeO7a2gFnCpFsw/edit#heading=h.gjdgxs

Further comments

Further comments from the organizers.

N/A