The SAGES Critical View of Safety Challenge: A Global Benchmark for AI-Assisted Surgical Quality Assessment

Deepak Alapatt [a,c,1,*], Jennifer Eckhoff[d,e,1], Zhiliang Lyu[d], Yutong Ban[d,f], Jean-Paul Mazellier[b], Sarah Choksi[g,h], Kunyi Yang[f], 2024 CVS Challenge Consortium[3], Quanzheng Li[d], Filippo Filicori[g], Xiang Li[d], Pietro Mascagni[b,i], Daniel A. Hashimoto[d,j], Guy Rosman[k], Ozanan Meireles[d,l,2], Nicolas Padoy[a,b,2]

[a] *University of Strasbourg, CNRS, INSERM, ICube, UMR7357, Strasbourg, France*
[b] *IHU Strasbourg, France*
[c] *Scialytics SAS, France*
[d] *Massachusetts General Hospital, Harvard Medical School, USA*
[e] *University Hospital Cologne, Germany*
[f] *Global College, Shanghai Jiao Tong University, China*
[g] *Lenox Hill Hospital, Northwell Health, USA*
[h] *Albany Medical Center, USA*
[i] *Fondazione Policlinico Universitario A. Gemelli IRCCS, Italy*
[j] *University of Pennsylvania, USA*
[k] *Massachusetts Institute of Technology, USA*
[l] *Duke University, USA*

A B S T R A C T

Advances in artificial intelligence (AI) for surgical quality assessment promise to democratize access to expertise, with applications in training, guidance, and accreditation. This study presents the SAGES Critical View of Safety (CVS) Challenge, the first AI competition organized by a surgical society, using the CVS in laparoscopic cholecystectomy, a universally recommended yet inconsistently performed safety step, as an exemplar of surgical quality assessment. A global collaboration across 54 institutions in 24 countries engaged hundreds of clinicians and engineers to curate 1,000 videos annotated by 20 surgical experts according to a consensus-validated protocol. The challenge addressed key barriers to real-world deployment in surgery, including achieving high performance, capturing uncertainty in subjective assessment, and ensuring robustness to clinical variability. To enable this scale of effort, we developed EndoGlacier, a framework for managing large, heterogeneous surgical video and multi-annotator workflows. Thirteen international teams participated, achieving up to a 17% relative gain in assessment performance, over 80% reduction in calibration error, and a 17% relative improvement in robustness over the state-of-the-art. Analysis of results highlighted methodological trends linked to model performance, providing guidance for future research toward robust, clinically deployable AI for surgical quality assessment.

K E Y W O R D S :   Surgical activity recognition, Critical View of Safety, Cholecystectomy, Safety, SAGES, Challenge.

---

*Corresponding author:

e-mail: `deepak@scialytics.io` (Deepak Alapatt)

[1]Deepak Alapatt and Jennifer Eckhoff contributed equally and share co-first authorship.
[2]Ozanan Meireles and Nicolas Padoy contributed equally and share co-last authorship.
[3]This manuscript represents version 1 of the SAGES CVS Challenge report. The final author list is currently being finalized and will be updated in subsequent versions.

1. Introduction

Surgery is among the most critical components of modern healthcare, accounting for about a third of all healthcare expenditure and a comparable share of the global burden of disease (Meara et al., 2015). Despite this awesome financial and human impact, surgical practice remains variable, both in processes and outcomes. Historically, the operating room (OR) has functioned as a data-poor environment, limiting scalable efforts to understand and reduce this variability. The digital transformation of the OR, particularly driven by the widespread adoption of minimally invasive surgery, has begun to shift this trend. Surgical video, now a routine byproduct of surgical care, is enabling a new paradigm in which data-driven approaches can uncover the link between the quality of surgical care and patient outcomes.

Surgical quality, spanning decision making, technical skills, and adherence to best practices, is now a central focus in understanding how intraoperative factors influence patient outcomes. Since the seminal work of Birkmeyer et al. (2013) demonstrating a quantifiable link between technical skills and patient outcomes, video based assessment has emerged as a powerful tool for evaluating surgical quality. Surgical societies, most notably the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES)[4], have led efforts in this space through dedicated initiatives aimed at advancing the science of video-based surgical assessment. However, despite growing interest, most existing efforts are constrained by limited scale. The experts best equipped to assess surgical quality, experienced surgeons, are often too burdened to evaluate large volumes of video. Compounding this issue is the inherent subjectivity in surgical assessment, with variability even among expert reviewers. Artificial intelligence offers a natural solution to these challenges, with the potential to replicate and disseminate expert consensus. These AI assessments could democratize access to expertise when it's needed the most, in the OR, and enable a mature understanding of the variability that we have come to accept as an intrinsic part of surgical care today.

A notable example of progress toward AI-assisted surgical quality assessment is the automated evaluation of the Critical View of Safety (CVS) (Strasberg et al., 1995) in laparoscopic cholecystectomy. CVS is a universally recommended safety step that ensures adequate dissection for the safe identification of anatomical structures, yet it remains inconsistently implemented in clinical practice. Prior research has demonstrated the feasibility of using AI to assess CVS achievement (Mascagni et al., 2022), document this step (Mascagni et al., 2021b), and even provide real-time intraoperative feedback (Mascagni et al., 2024). However, most existing approaches rely on black-box deep learning models, raising important concerns about their robustness. Variations in surgical instrumentation, technique, workflow, lighting conditions, and image quality across centers introduce significant challenges to generalization—challenges that must be rigorously understood before these models can safely influence clinical decisions.

To thoroughly evaluate and improve AI models for surgical quality assessment, structured biomedical challenges offer a powerful framework. These challenges allow researchers to compete and benchmark diverse methodological approaches on curated datasets. The SAGES CVS Challenge represents a landmark step in this direction. As the first biomedical challenge organized by a surgical society, it leverages the unique infrastructure of the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES), whose 7,000+ members include global leaders in minimally invasive surgery. In this challenge, CVS serves as an exemplar of surgical quality: a well-defined, clinically meaningful target with clear guidelines for assessment. Through a global effort with 54 institutions across 24 countries, we assembled a diverse dataset of 1,000 laparoscopic cholecystectomy videos, each curated with rigorous quality-control protocols and annotated by a panel of 20 expert surgeons.

This work describes the resulting benchmark and the EndoGlacier infrastructure that enabled it: a reproducible, automation-driven framework for coordinating global video sourcing, expert training, multi-annotator workflows, and quality control. Participants were evaluated across three complementary dimensions essential for clinical deployment—absolute performance requirements, alignment with expert uncertainty, and robustness under real-world domain shifts. Submissions from 13 international teams delivered measurable gains over state-of-the-art in all three areas, offering both methodological advances and insight into the design choices that influence performance. By uniting scale, clinical focus, and reproducible evaluation, this benchmark aims to catalyze the development of robust AI systems for surgical quality assessment.

2. Related work

The SAGES CVS challenge relates to several research topics, for which we present the relevant literature in the following paragraphs.

2.1. Video-based Surgical Quality Assessment

The first work linking technical skills to outcomes dates back to 2013 (Birkmeyer et al., 2013), and since then, the growing body of literature on the topic (Grüter et al., 2023) has prompted the need for solutions to measure surgical performance objectively. Minimally invasive surgery and surgical video are quickly becoming some of the most promising sources of information to make these assessments. Being able to track how performance changes over time, essentially the learning curve, means we can begin to shift toward competency-based credentialing models (Pryor et al., 2023) and offer personalized feedback (Naik et al., 2018). Beyond the individual surgeon, this could help identify systematic issues and offer potential solutions at the department, hospital, or even healthcare system level. The applications are innumerable, fundamentally boiling down to being able to identify gaps in current practice and offer actionable insights and support. Most quality assessment tools, scoring rubrics used to measure quality, generally fall into one of four categories: assessing errors (i.e., suboptimal techniques), assessing events (i.e., consequential errors), technical skills (e.g., precision, smoothness), and procedure-specific assessments (e.g., quality of performing a certain step, choice of technique) (Grüter et al., 2023).

Relative positioning of our work: Previous work on manual video-based surgical quality assessment (Grüter et al., 2023) is

---

[4]Society of American Gastrointestinal and Endoscopic Surgeons (SAGES). https://www.sages.org
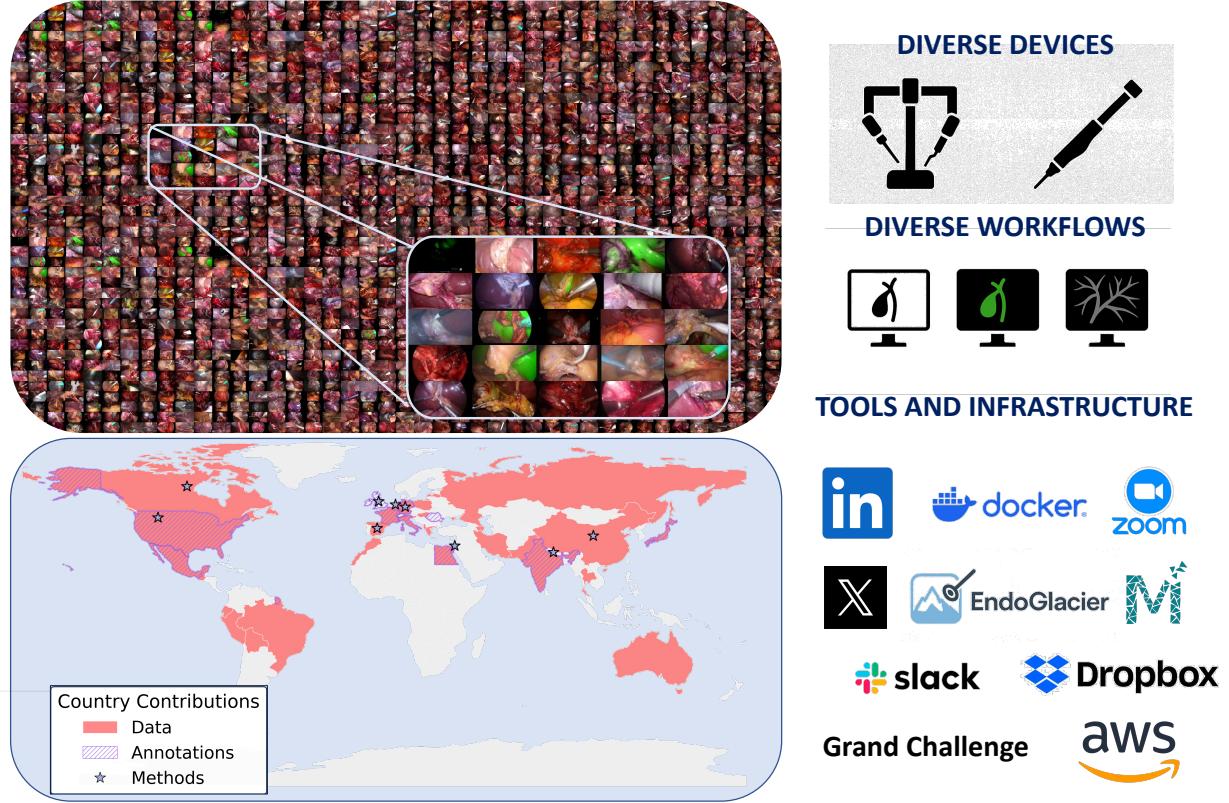
Fig. 1. A global benchmark for surgical safety AI, the challenge brought together over 1000 annotated cholecystectomy videos, representing diverse acquisition devices (hardware, instrumentation, etc.) and workflows (case difficulty, techniques, preferences, instrumentation, etc.). The dataset was constructed through a global collaborative effort spanning three years, with the bottom-left panel illustrating countries contributing data, annotations, and methods to this international initiative.

increasingly validating the potential, if not the need, for systems to perform objective assessments. Most of this work is limited in the number of videos evaluated (Birkmeyer et al., 2013; Scally et al., 2016; Varban et al., 2020, 2021; Stulberg et al., 2020; Kurashima et al., 2022), restricting both the number of surgeons assessed and the number of cases reviewed per surgeon. This natural limitation of highly specialized, manually performed reviews is often compromised by assumptions, like the idea that a surgeon broadly delivers consistent quality across cases, discounting important factors like case-specific complexity and learning curves. Our work positions itself as one of the largest surgical video-based surgical quality assessment initiatives to date, surpassing previous work by an order of magnitude, with a pool of 20 experts evaluating 1,000 cases representing a diverse range of factors often overlooked, such as acquisition device characteristics, hospital-specific practices, and technical variation.

2.2. AI-based Surgical Quality Assessment

Alongside the manual video-based assessment literature, there is a growing body of work demonstrating the feasibility and value of automating surgical quality assessments. To date, much of the work in this space has focused on technical skill evaluation (Lam et al., 2022), driven in part by the lower data privacy concerns when working with simulator data and the availability of rigorously validated scoring frameworks. More recently, efforts have begun to extend to other types of quality metrics, such as the detection of adverse events (Bose et al., 2025; Eppler et al., 2023) like bleeding, thermal injuries,

or mechanical damage to anatomical structures. Procedure-specific models are also becoming more common, with several focused on evaluating the adequacy of safety step implementation (Mascagni et al., 2022) or the approach taken during key procedural moments (Laplante et al., 2023). Downstream applications of these models are increasingly being demonstrated, from automated documentation of safety steps (Mascagni et al., 2021b) to evaluation of training interventions (Mascagni et al., 2022), and even delivery of real-time intraoperative feedback (Mascagni et al., 2024) and predictive analytics (Yin et al., 2024).

Relative positioning of our work: Our work falls within the category of procedure-specific assessments of surgical quality, and extends it by foregrounding several understudied but critical considerations for real-world deployment. We explore and benchmark methods that explicitly account for the inherent ambiguity in surgical quality assessments. We also systematically evaluate the robustness of these methods under distribution shifts related to device characteristics, case complexity, regional differences in technique, and other real-world factors. In doing so, we contribute a valuable benchmark and highlight key failure modes that warrant deeper investigation if these models are to be used safely and reliably in clinical settings.

2.3. AI for CVS assessment

Among the various initiatives in surgical quality, Critical View of Safety (CVS) assessment stands out for its well-established clinical relevance and the growing volume of work exploring its automation. Initially proposed by Strasberg et al.

(1995), CVS is a safety step designed to ensure adequate dissection for the secure identification of anatomical landmarks, thereby preventing bile duct injuries—a devastating complication that costs over a billion dollars annually in the United States alone (Carroll et al., 1998). Now supported by more than 30 years of literature, CVS is universally recommended by multi-society guidelines (Brunt et al., 2020) and is known to prevent over 97% of major bile duct injuries when properly implemented (Way et al., 2003). Despite strong evidence suggesting that CVS can be achieved in 90–95% of cases (Avgerinos et al., 2009; Sanjay et al., 2010; Tsalis et al., 2015), its real-world implementation rate remains alarmingly low, often between 10–15% (Korndorffer Jr et al., 2020). This discrepancy has been attributed to the subjectivity of the assessment, as well as overconfidence among surgeons.

In response to this critical gap, researchers have developed structured annotation protocols (Mascagni et al., 2021a), AI models for CVS assessment and documentation (Korndorffer Jr et al., 2020; Mascagni et al., 2022, 2021b; Murali et al., 2023c,b; Ban et al., 2023; Yin et al., 2024), and systems to provide intraoperative feedback (Mascagni et al., 2024). Over time, a growing body of literature has emerged around this task, including multiple public benchmarks (Ríos et al., 2023; Murali et al., 2023a; Mascagni et al., 2025) and purpose-built methodological frameworks.

Relative positioning of our work: We build on this strong foundation by validating, refining, and expanding previously proposed annotation protocols through structured expert consensus, and by assembling the largest and most diverse benchmark dataset in this domain. Not only does our dataset exceed prior efforts by an order of magnitude in scale, with 1,000 laparoscopic cholecystectomy videos, but it also includes annotations from a more diverse panel of expert reviewers (20 annotators). The data spans a wide range of acquisition settings, including different institutions, devices, and regional practices, with representation from low- and middle-income countries. This diversity enables new lines of investigation into sources of bias related to annotators, devices, geography, and workflow, pushing the field toward more robust and generalizable AI models for surgical safety.

2.4. Biomedical challenges

Biomedical challenges are a well-established format prompting teams from academia and industry alike to compete in advancing research or benchmarking on emerging or underserved topics. Within surgical computer vision, challenges have regularly served as a springboard for insight into promising research directions, especially in establishing new benchmarks in this relatively niche domain (Maier-Hein et al., 2022). To date, most endoscopic vision challenges have focused on foundational tasks aimed at understanding surgical context from video data. These efforts span a range of spatial and temporal granularities—from coarse procedural phases (Maier-Hein et al., 2021) to fine-grained actions (Nwoye et al., 2023a,b), from frame-level assessments (Al Hajj et al., 2019) to pixel-precise segmentations—across various types of surgical procedures (Allan et al., 2019, 2020; Roß et al., 2021; Maier-Hein et al., 2021; Luengo et al., 2021; Bano et al., 2021).

Relative positioning of our work: Leveraging the platform of SAGES, and as the first biomedical challenge led by a surgical society, our work seeks to bridge two persistent gaps in the challenge literature: clinical relevance and scale. Building on prior challenges that have successfully advanced foundational, context-aware systems for surgical video understanding, we pivot upstream toward a task with a clearly defined and direct clinical purpose: the automatic assessment of the Critical View of Safety. In doing so, we surface critical issues that emerge when moving from procedural understanding to clinical quality assessment, such as managing subjectivity in labeling, capturing ambiguity in human assessment, and building models robust to real-world data heterogeneity.

In terms of scale, our challenge departs from the common approach of sourcing data and annotations from one or a small number of institutions. Prior challenges have demonstrated immense value, but their limited scale can introduce subtle biases tied to local practices, individual annotators, or specific devices. By expanding the dataset across 54 centers in 24 countries and involving a diverse panel of globally dispersed annotators, we open up new possibilities for studying generalization, robustness, and fairness in clinical AI models. This scale also brings previously hidden challenges to the surface, ranging from how to design evaluation protocols that account for inter-rater variability, to the technical and logistical hurdles of coordinating global annotation efforts.

Naturally, scaling up biomedical challenges in this way introduces new demands. From data governance and privacy compliance to the infrastructure needed to engage, track, and manage a large number of stakeholders, we encountered and addressed a range of practical challenges. We hope that the design choices and tooling developed for this challenge serve as a framework for future initiatives operating at similar or larger scales, especially those seeking to tackle clinically meaningful tasks with real-world constraints.

3. Coordinated Data Flows for Large-Scale Annotation

Managing diverse dataset generation at a global scale requires more than assembling data, it demands infrastructure capable of sustaining continuous, high-quality flows of videos, expert annotations, and coordination across dozens of institutions and contributors. Here, we break down our dataset generation process into three coordinated streams visualized in Figure 2: (1) annotator management, (2) video management, and (3) orchestration and automation.

3.1. Annotator Management

Our fundamental goal was to assemble the largest and most diverse pool of annotators possible. Size was critical to capture the inherent subjectivity in surgical quality assessments robustly, and diversity was essential to minimize bias from individual raters.

To achieve this, we conducted dedicated onboarding sessions that introduced potential annotators to the challenge, its goals, the team, and the commitments expected. Quality control mechanisms were put in place to identify individuals with the expertise, resources, and motivation to meaningfully contribute.

Annotators were provided with expert-consensus protocols, flashcards, and training materials, along with written instructions and video tutorials to familiarize themselves with MO-SaiC (Mazellier et al., 2023), a web-based annotation platform developed at IHU-Strasbourg. Beyond providing the annotation
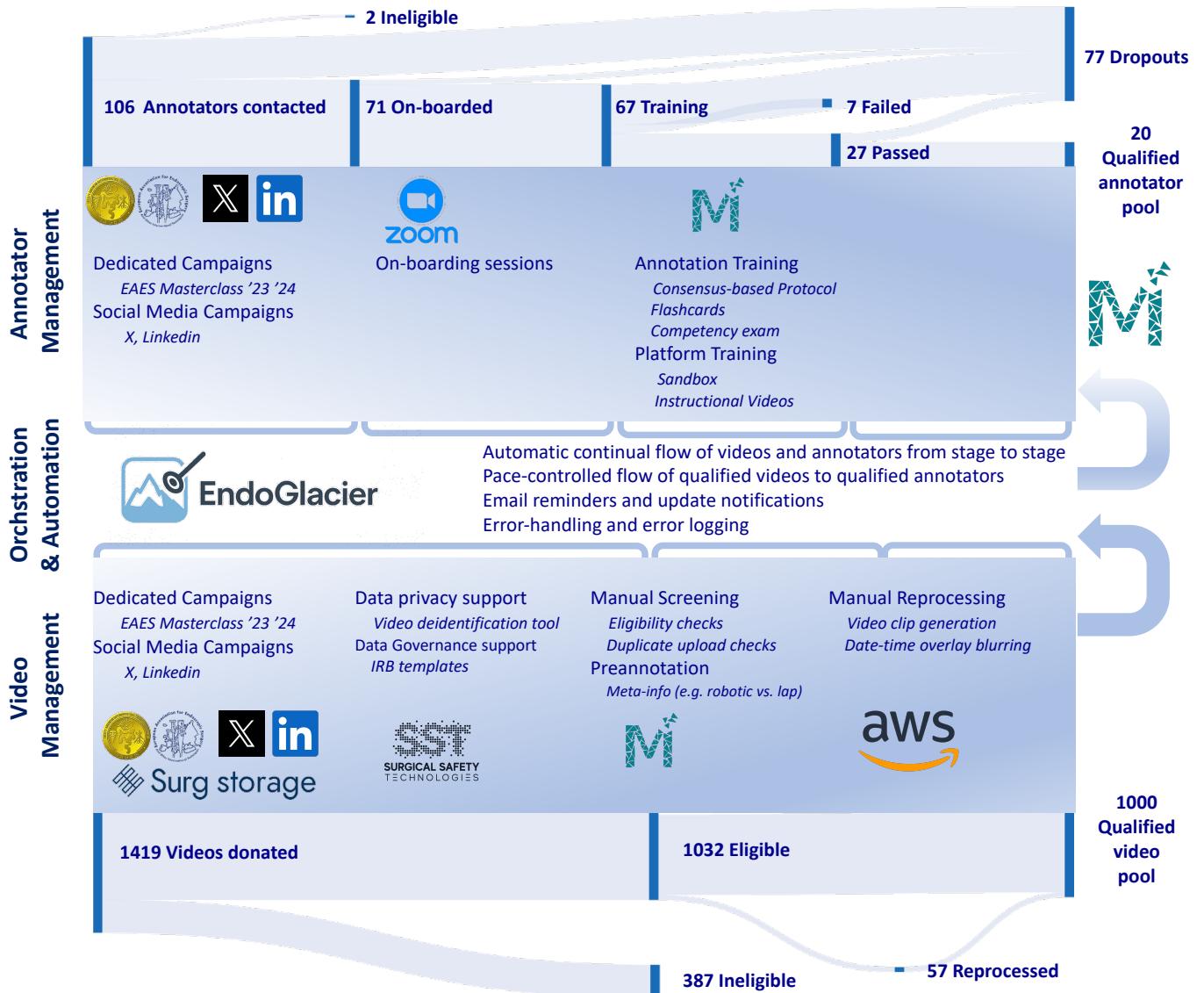
Fig. 2. Overview of the annotation infrastructure. The pipeline integrated three coordinated flows: (i) Annotator management, encompassing recruitment, onboarding, training, and competency validation; (ii) Orchestration and automation, ensuring blinded allocation of videos, controlled pacing of assignments, reminders, and error logging; and (iii) Video management, including privacy safeguards, eligibility screening, and reprocessing to yield a curated pool of qualified surgical videos.

interface, MOSaiC enabled blinded scoring of every video, centralized all instructional material for quick reference, and APIs that allowed automated assignment of videos and annotators, ensuring consistency and efficiency at scale. As a final gate, we implemented a competency-based exam requiring at least 75% correspondence with expert ratings to ensure conceptual alignment and minimum annotation quality.

Through this structured pipeline, we contacted or were contacted by 106 potential annotators, recruited through channels including social media and in-person onboarding campaigns. Of these, 71 met basic eligibility (with 2 excluded due to non-clinical backgrounds). Sixty-seven proceeded to the competency exam, and 27 passed, ultimately yielding a qualified pool of 20 expert annotators. Seventy-seven of the original 106 dropped out at various stages, most commonly due to time constraints and competing professional commitments.

3.2. Video Management

The primary goal of the video pipeline was to assemble the largest and most diverse dataset of laparoscopic cholecystectomy procedures possible, while ensuring each sample was clinically meaningful and task-relevant for the challenge. The core prediction task in the CVS Challenge is to determine whether the Critical View of Safety (CVS) has been achieved during the 90-second window immediately prior to clipping of the cystic duct or artery, the critical decision point at which misidentification can lead to a bile duct injury.

To support this objective, we launched a broad video sourcing campaign over a three-year period, including social media outreach, in-person recruitment events, and direct institutional engagement. This effort yielded a total of 1,419 donated surgical videos. We supported contributors with practical tools and resources, including institutional review board (IRB) templates and a custom-built de-identification platform for redacting out-of-body segments that may contain patient or staff identifiers.

All incoming videos were subjected to a manual screening and preannotation phase conducted by members of the organizing team. This process had two main purposes:

1. **Eligibility verification**, based on a predefined set of criteria:

   - The video must be of a minimally invasive cholecystectomy (either laparoscopic or robotic).

   - A continuous 90-second segment prior to clipping of the cystic duct or artery must be available, during which the operative field is clearly visible.

   - Videos were excluded if they involved a change in surgical strategy due to unsafe conditions, referred to as a *bailout procedure* , such as conversion to open surgery or subtotal cholecystectomy.

   - Videos were also excluded if they were incomplete or did not capture the clipping of the cystic structures.

2. **Preannotation of objective metadata**, with minimal room for interpretation:

   - The timestamp of the cystic duct or artery clipping, used to extract the 90-second target segment.

   - The use of adjunctive visualization techniques, such as intraoperative cholangiography (IOC) or indocyanine green (ICG) fluorescence, which may indicate case difficulty or reinforce CVS implementation.

   - The surgical approach used (laparoscopic vs. robotic), which significantly affects procedural workflow.

To ensure quality and consistency, each video was independently reviewed and preannotated by two raters. In cases of disagreement, a third rater was introduced to adjudicate. This process was repeated iteratively until two consecutive raters produced concordant annotations across all eligibility and metadata criteria.

For privacy compliance, 57 videos required additional reprocessing to blur overlaid date-time stamps. All qualified videos were then clipped to produce 90-second segments corresponding to the critical decision window. This resulted in a final pool of 1,000 high-quality, clinically relevant qualified video clips.

Each qualified clip was subsequently assigned to three independent, qualified annotators, who were asked to assess whether the CVS had been achieved and to rate their confidence in the assessment. This multi-annotator protocol was designed to capture the inherent subjectivity of surgical quality assessment and provide a basis for evaluating inter-rater agreement.

3.3. EndoGlacier: Orchestration and Automation

Handling this volume of data, annotators, and institutional contributors over a multi-year period would have been intractable without automation. To address this, we developed EndoGlacier, a Python-based framework for managing large-scale surgical data flows.

EndoGlacier served two critical purposes:

- Accelerating and parallelizing the flow of videos and annotators through their respective pipelines. Automated stage-to-stage progression, participant status updates, email reminders, error logging, and error handling mechanisms (e.g. automatic retries) enabled the organizing team to monitor, adjust, and intervene when needed.

- Controlling the flow of videos from the qualified video pool to the qualified annotator pool. While quality controls ensured conceptual alignment with the annotation protocol, we deliberately preserved the diversity of annotator opinions. EndoGlacier prevented over-enthusiastic annotators from monopolizing the dataset while ensuring balanced contributions across the pool. Specifically, it assigned each annotator a bucket of 20 videos, drawn from the qualified pool, on a bi-weekly basis.

4. SAGES CVS challenge

4.1. Benchmark Task: Assessing the Critical View of Safety

The Critical View of Safety (CVS) is a universally recommended surgical safety step in laparoscopic cholecystectomy. Its purpose is to ensure that the key aspects of dissection have been performed with sufficient quality to allow conclusive identification of the cystic duct and cystic artery, before proceeding to clip and cut these structures. This step is central to preventing bile duct injuries, a devastating and largely preventable complication.

CVS is composed of three binary criteria, all of which must be met to achieve the critical view:

- **Criterion 1 (C1):** Two and only two tubular structures are visible entering the gallbladder.

- **Criterion 2 (C2):** The hepatocystic triangle is cleared of fat and fibrous tissue.

- **Criterion 3 (C3):** The lower third of the gallbladder is detached from the liver bed.

The decision to proceed is most relevant in the final moments prior to clipping and cutting the cystic duct and artery. To reflect this clinical reality, the CVS Challenge focuses assessment on the 90-second window immediately preceding this critical juncture in the procedure.

To balance clinical relevance with annotation effort and modeling tractability, the benchmark is structured such that CVS achievement can be assessed at a granularity of one frame every five seconds within this 90-second window. This design provides both temporal resolution and practical annotation density, aligning the benchmark with the real-world clinical decision context.

4.2. Dataset Characteristics

The SAGES CVS Challenge dataset comprises a total of 1,000 laparoscopic cholecystectomy cases, each represented by a 90-second clip taken from the critical decision window preceding clipping of the cystic duct or artery. The dataset is split into a training set of 700 videos and a test set of 300 videos (Table 1). To enable practical annotation density while preserving temporal resolution, one frame every five seconds (18 frames per video) is annotated throughout the clip.

Annotator confidence, self-reported on a per-clip basis (0–1 scale), shows a mean of 0.64 ± 0.28 on the training set and 0.58 ± 0.27 on the test set. As expected, confidence varies across cases, reflecting case complexity and clinical ambiguity.

The dataset captures substantial clinical and technical diversity. 23 countries are represented in the training split and 18 countries in the test split, with an average of 30.43 ± 46.54 and

16.57 ± 23.18 videos per country, respectively. Data acquisition was performed across the same 8 device vendors in both splits. A portion of videos lack device metadata (156 in train, 114 in test), but the remaining videos are evenly distributed, with a mean of 68.0 ± 65.47 videos per device in the training set and 23.25 ± 24.44 in the test set.
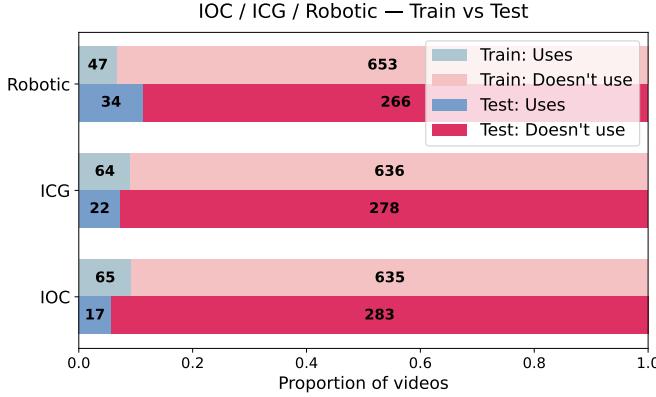

Fig. 3. Distribution of adjunct imaging techniques and surgical platform across training and test sets. Proportion of videos using intraoperative cholangiography (IOC), indocyanine green (ICG) fluorescence imaging, and robotic-assisted surgery. The distribution demonstrates clinical and workflow diversity across the dataset, with variation between training and test splits done randomly to reflect unselected real-world patterns.

The use of robotic platforms, Indocyanine Green (ICG) fluorescence, and Intraoperative Cholangiography (IOC) varies across the dataset and is summarized in Figure 3. Most procedures are performed laparoscopically (653/700 in train, 266/300 in test), with ICG and IOC used in a minority of cases, providing an opportunity to examine model robustness across clinically heterogeneous workflows.

Annotation of Critical View of Safety (CVS) achievement is performed for each frame along three binary criteria (C1, C2, C3). The dataset captures not only frame-level classification but also annotator agreement, which is known to vary in subjective quality assessment tasks. We distinguish between full agreement (all 3 annotators concur) and partial agreement (2 out of 3 annotators agree). Such variation is expected and informative: all annotators were screened through structured training and competency validation to minimize sources of bias unrelated to inherent subjectivity, including lack of alignment with the protocol, inattention, or annotation fatigue.

Figure 4 presents the distribution of frame-level achievement rates and agreement levels for each criterion. As expected, C2 (clearing the hepatocystic triangle) shows the highest frame-level achievement rate, while C1 (two and only two tubular structures visible) and C3 (lower third of gallbladder detached) are achieved less frequently. Notably, partial agreement is observed across all criteria and is particularly common in borderline cases, reinforcing the challenge of consistent quality assessment in surgical video.

At the video level, annotators assessed whether each CVS criterion was achieved during the 90-second window. Across the dataset:

- **C1**: 413 videos achieved, 587 not achieved
- **C2**: 600 videos achieved, 400 not achieved
- **C3**: 395 videos achieved, 605 not achieved

Overall, the dataset reflects the real-world variability, subjectivity, and technical heterogeneity that models deployed in surgical environments will encounter. By capturing this complexity, the CVS Challenge provides a benchmark that goes beyond technical optimization and instead tests the robustness and clinical relevance of AI-based surgical quality assessment.

Table 1. Summary statistics of the SAGES CVS Challenge dataset. Key characteristics of the video dataset and annotations used for training and evaluation. Annotator confidence is self-reported by annotators on a per-clip basis. Country and device metadata are reported separately for training and test splits to highlight dataset diversity.

Number of Videos (90-second clips)			**Total**	**1000**
			Train	700
			Test	300
Number of frames annotated per video			**18**	
Annotation frequency			1 frame every 5 seconds	
Annotator Confidence (per clip, 0–1 scale)				
Split	**Min**	**Max**	**Mean**	**SD**
Train	0	1	0.64	0.28
Test	0	1	0.58	0.27
Country metadata		**Train**		**Test**
Total number of countries		23		18
Videos from unknown country		0		0
Mean # videos per country (± SD)		30.43 ± 46.54		16.57 ± 23.18
Device metadata		**Train**		**Test**
Total number of device vendors		8		8
Videos from unknown device		156		114
Mean # videos per device (± SD)		68.00 ± 65.47		23.25 ± 24.44

4.3. Challenge Tasks and Evaluation Metrics

Participants were evaluated on a held-out test set of 300 laparoscopic cholecystectomy videos submitted through the Grand Challenge platform using Docker-based containers. Test videos and annotations were withheld, and all models were required to operate causally using only past and current frames at inference time. Each submission was evaluated across three subchallenges designed to probe different facets of the modeling problem.

Subchallenge A: CVS Achievement. This task evaluated classification performance in predicting the most common (majority) expert assessment for each of the three CVS criteria. The evaluation metric was mean average precision (mAP), computed per frame and averaged across criteria. The ground truth for each frame was defined as:

$$y = \mathrm{mode}(\{l_i\}), \quad i \in \{1, 2, 3\}$$

where l_i is the binary label provided by annotator i.

Subchallenge B: Uncertainty Quantification. This task tested whether models could express uncertainty in line with annotator disagreement and confidence. Participants submitted a probability between 0 and 1 for each frame and criterion. The evaluation metric was the Brier Score (BS), calculated using a soft label incorporating annotator confidence:

$$y = \frac{1}{3} \sum_{i=1}^{3} (0.5 + (l_i - 0.5) \cdot c_i)$$

where l_i is the binary label and $c_i \in [0, 1]$ is the confidence score provided by annotator i. The Brier Score was computed per frame, then averaged across frames and criteria.
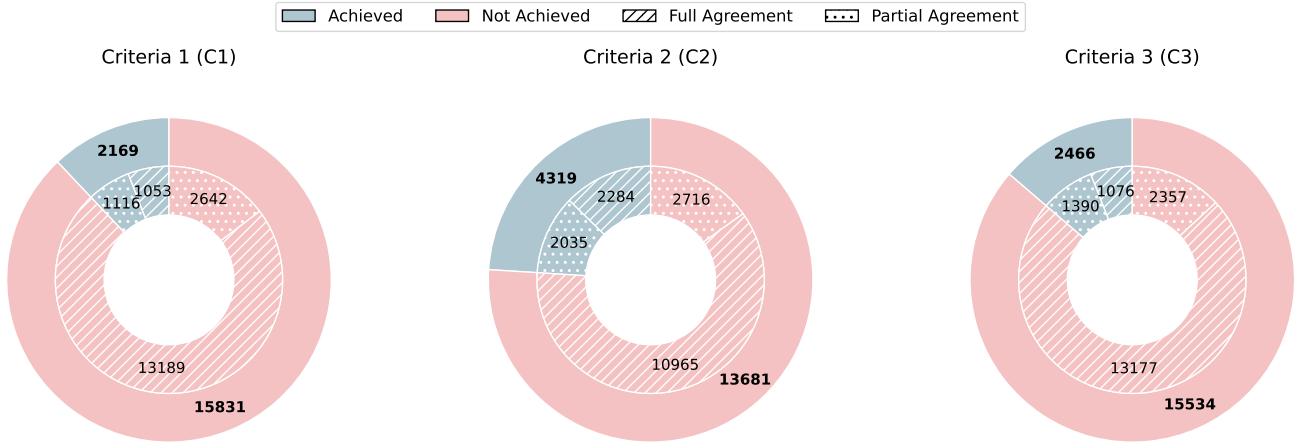
Fig. 4. Distribution of CVS assessment outcomes and annotator agreement. For each CVS criterion (C1, C2, C3), we report the number of clips rated as "Achieved" vs. "Not Achieved," and the corresponding level of annotator agreement (full agreement across all three annotators vs. partial agreement). This highlights both the subjective nature of CVS assessment and the distribution of clinical outcomes in the dataset.

Subchallenge C: Domain Shift Robustness. This task evaluated model generalization under real-world distribution shifts (Ben-David et al., 2010; Shao et al., 2024). Several variant test sets were constructed from the full test split to introduce differences in imaging modality (IOC, ICG), surgical platform (robotic vs. laparoscopic), device type, country of origin, and annotator confidence. mAP was calculated per variant set, and the final score was defined as:

$$\text{Domain Robustness Score} = \min_{\text{variant } v \in \mathcal{V}_{10\text{--}100}} \text{mAP}_v$$

where $\mathcal{V}_{10\text{--}100}$ excludes the bottom 10th percentile of variant scores to reduce sensitivity to extreme outliers. Ground truth labels were again computed using majority vote:

$$y = \text{mode}(\{l_i\}), \quad i \in \{1, 2, 3\}$$

4.4. Challenge Design and Submission Framework

The SAGES CVS Challenge was designed to balance clinical realism with reproducibility and ease of participation (Fig. 5). Submissions were made through the Grand Challenge platform, using Docker containers to ensure standardized evaluation. Participants were required to implement causal models, making predictions at each timepoint using only the current and preceding frames, in line with intraoperative constraints. No restrictions were placed on external data sources, as long as they were publicly accessible, ensuring a fair playing field across teams.

To support participation, an open-source submission template was provided, including example code, a test video, and scripts for local building, testing, and packaging. Each model was expected to ingest a single 90-second, 1 fps video and output predictions for all three CVS criteria at each of the 90 frames.

Evaluation was automated via Grand Challenge's infrastructure. Submissions were uploaded as Docker containers, registered and validated on the platform, and then tested against the held-out dataset. Participants were required to test locally using the provided inference interface, and could optionally validate on-platform using an example input video. Participants submitted a single Docker container, which was evaluated across all three subchallenges. This design reflected real-world deployment constraints: in surgical AI systems, raw predictive performance, reliability under distribution shifts, and appropriate uncertainty estimation are all critical, yet not inherently aligned. Optimizing for one may compromise another. Requiring a single model to be evaluated across all criteria forced participants to navigate these trade-offs, as would be required for clinically deployable systems.

4.5. Awards

Monetary prizes were awarded to the top 3 ranking teams for each of the 3 subchallenges. The overall winner across the 3 subchallenges was additionally awarded an NVIDIA IGX edge AI platform.

5. Methods

5.1. Baseline: LG-DG

As a baseline for our challenge, we adapted LG-DG (Satyanaik et al., 2024), a recently proposed object-centric classification model optimized for domain generalization in surgical video. LG-DG builds on a previous latent graph framework (LG-CVS, (Murali et al., 2023c)) by explicitly disentangling semantic, visual, and image-level features through a combination of architectural and loss-based design choices.

The model operates in two stages: first, a Mask-RCNN (He et al., 2017) detector identifies and localizes surgical tools and anatomical structures, encoding their spatial, semantic, and visual properties as nodes and edges in a latent graph. This graph is passed to a graph neural network-based classification head for downstream prediction of CVS criteria. To improve robustness, the model is trained with an auxiliary reconstruction objective and a disentanglement loss that regularizes the predictions from masked graph variants, emphasizing different feature categories.

For this challenge, we retrained LG-DG using spatial supervision from the Endoscapes2023 dataset and CVS annotations from the SAGES CVS Challenge training set.
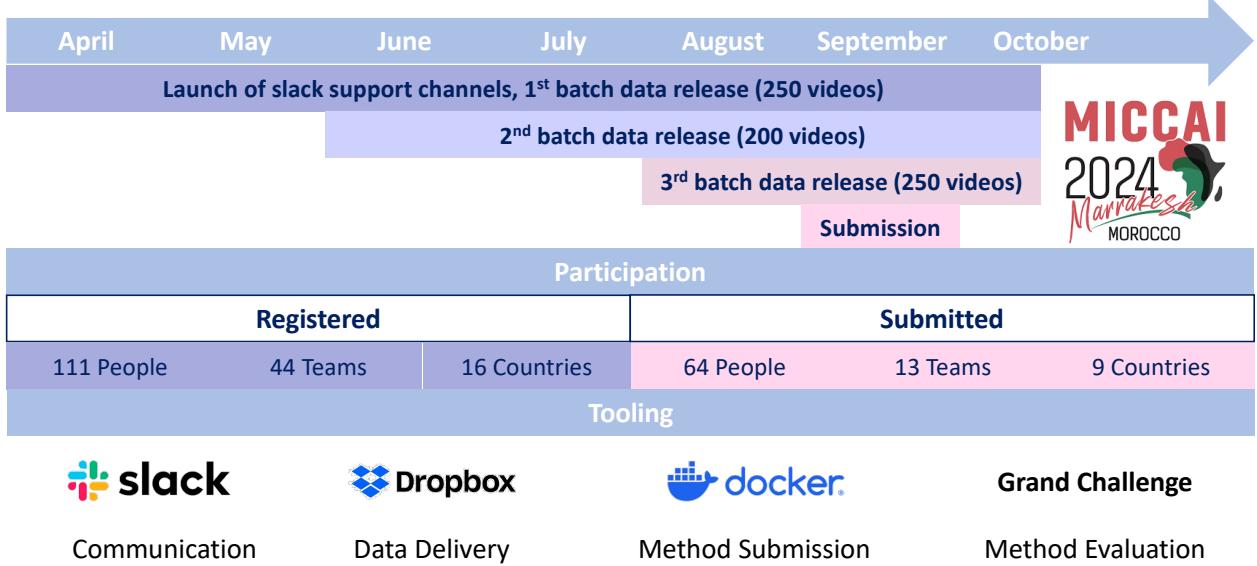
Fig. 5. Overview of the SAGES CVS Challenge. The top panel shows dataset release and timeline, the middle panel summarizes team participation, and the bottom panel highlights the core tooling stack used to manage communication, data delivery, and evaluation.

5.2. Competing methods

In addition to the baseline model, LG-DG, 13 teams submitted competing methods to the SAGES CVS Challenge. The goal was to predict per-frame achievement of the three CVS criteria using only the labeled 700-video training set. All teams developed their own training pipelines and architectures and submitted Dockerized inference code, which was automatically evaluated across all subchallenges on a hidden 300-video test set.

The diversity of approaches submitted reflects the multifaceted nature of the task and the real-world complexity of surgical video understanding. Teams variously framed the problem as framewise classification, sequence modeling, or multitask learning; built on top of pretrained image encoders or segmentation models; and handled label uncertainty through loss design, label smoothing, or ensemble calibration. Several groups also integrated auxiliary tasks or domain adaptation strategies to improve generalization. Many of the top-performing methods combined multiple strategies through architectural design and ensembling.

Below, we describe the methodological design of each submission, with attention to key architectural components, auxiliary objectives, supervision strategies, and other distinctive modeling choices. For an overview of each method's high-level design choices, see Table 2.

5.2.1. Team TUE-VCA: *Scalable CVS Classification via Self-Supervised Pretraining and Semi-Supervised Distillation*

The TUE-VCA team addressed the challenge of class imbalance and label uncertainty through a two-stage training approach combining large-scale self-supervised pretraining and semi-supervised knowledge distillation. A Pyramid Vision Transformer v2 (PVT-v2) (Wang et al., 2022) backbone was first pretrained using DINO on a SurgeNet, a collection of over 4 million unlabeled frames from public surgical datasets (Jaspers et al., 2024), and the GenSurgery dataset, which compiles 680 hours of YouTube videos depicting general surgery

procedures (Schmidgall et al., 2024). This self-supervised model was then fine-tuned on the labeled training set from the SAGES CVS Challenge.
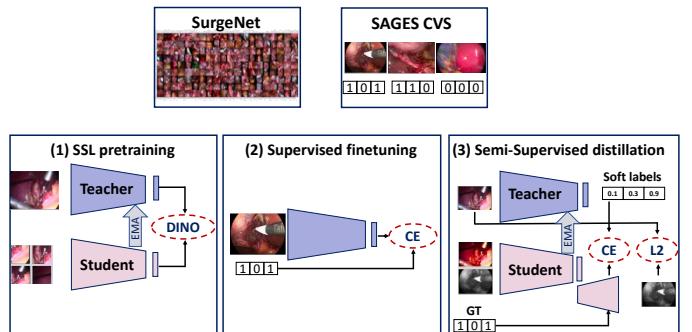

Fig. 6. Overview of various stages of the TUE-VCA method.

To further improve sensitivity to rare positive cases, the team employed a student-teacher distillation framework. A teacher model trained solely on labeled data was used to identify confident positive frames within the unlabeled data pool. These pseudo-labeled frames were then incorporated into training a student model on both labeled and pseudo-labeled data. Strong augmentations and a diminishing auxiliary reconstruction loss were applied to improve generalization and representation robustness.

The final model operated at the frame level with no temporal modeling. An ensemble of five PVT-v2 models was used for inference, and temperature scaling was applied to calibrate the sigmoid output probabilities following Kumar et al. (2022). To better account for skewed label distributions, the model was trained using label sampling.

5.2.2. Team Farm: *Efficient Adaptation of Vision Foundation Models with TCNs*

This method leverages recent advances in vision foundation models and parameter-efficient fine-tuning to address the CVS classification task. Multiple pretrained image back-

bones—including DINOv2 (giant and large) (Oquab et al., 2023), SigLIP (Zhai et al., 2023), InternImage (Wang et al., 2023), and ConvNeXt2 (Woo et al., 2023)—were adapted using Low-Rank Adaptation (LoRA) (Hu et al., 2022), allowing only a small subset of weights to be fine-tuned on the CVS Challenge training set. Some of these backbones were paired with a Temporal Convolutional Network (TCN) (Bai et al., 2018) to aggregate frame-level features into temporally-aware representations. Models were trained using both full-sequence (90-frame) and shorter-sequence (18-frame) settings, the latter approximating 1 FPS inference using overlapping sliding windows.
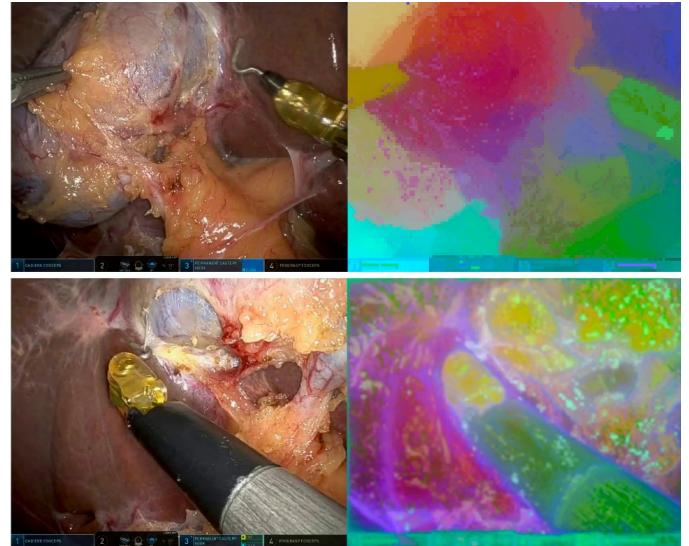
Fig. 7. Overview of various stages of the FARM method. Foundation models, despite being trained primarily on natural images, extract informative features from surgical video frames. By virtue of their large size and extensive pretraining datasets, they produce robust representations of distinct objects and structures - instruments, gallbladder, liver, connective tissue, and other recurring patterns. These representations can be visualized using UMAP (McInnes et al., 2018) (top) and PCA (bottom).

The final ensemble includes seven models: (1) DINOv2-giant with 90-frame TCN; (2) frame-level fine-tuned DINOv2-giant; (3) SigLIP with 18-frame TCN; (4) DINOv2-large with 18-frame TCN; (5) DINOv2-giant with 18-frame TCN; (6) ConvNeXt2-L with 18-frame TCN; and (7) frame-level fine-tuned InternImage. All models output sigmoid probabilities for each CVS criterion at each annotated frame, trained with binary cross-entropy loss.

5.2.3. Team CVS HUST: *Cyclic CNN-LSTM Training with Bidirectional Parameter Sharing*

This method proposes a two-stream cyclic training strategy to jointly leverage frame-level and video-level supervision. The approach alternates between two training modes: (1) a frame-level CNN stream trained on annotated frames, and (2) a video-level CNN-LSTM stream trained on full 90-frame clips. After each batch, CNN parameters are transferred between the two streams in both directions, promoting mutual refinement across granularities.

ConvNeXt (Liu et al., 2022), pretrained on natural images, serves as the backbone. Its first three stages are frozen during training, while later stages are fine-tuned. The temporal stream adds a single-layer LSTM to the CNN backbone. Binary cross-entropy loss is used for classification, and L1 loss is used to

Table 2. Summary of teams, architectures and training details.

#	Team	Affiliation(s)	Architecture	Temporal Component	Auxiliary Objective	Pretraining
1	TUE-VCA	Eindhoven University of Technology (Netherlands)	PVT-v2 (x5 ensemble)	None	L2 reconstruction loss	Self-supervised (DINO on SurgeNet and GenSurgery datasets)
2	Farm	Stanford University (USA)	DINOv2 (giant, large); SigLIP ConvNeXt2-L, InternImage	Stacked dilated TCN	None	Self-supervised (DINO, SigLIP)
3	CVS HUST	Huazhong University of Science and Technology (China)	ConvNeXt + MLP	LSTM	ℓ_1 loss for confidence prediction	Supervised (natural image weights)
4	SDS-HD	German Cancer Research Center (DKFZ) (Germany)	EVA02-Large Transformer	None	Segmentation loss on Endoscapes-pretrained	Self-supervised (MoCov2 on Cholec80, HeiChole, Endoscapes)
5	SRV-WEISS	University College London (UK)	EndoViT + Dense Prediction Transformer, EfficientNet, ConvNeXt	None	Segmentation + classification fusion; YOLOv8 pseudolabels	Self-Supervised (EndoViT) + Supervised (EndoViT on Endoscapes segmentation; EfficientNet on Endoscapes classes)
6	Ostrich	Anonymous	DenseNet-121 (5x ensemble)	None	None	Supervised (ImageNet weights)
7	FightTumor	University Health Network (Canada)	ConvNeXt-B	None	None	Supervised (ImageNet weights)
8	HUFT-MedIA	Hefei University of Technology (China)	Vision Transformer (ViT) + MLP	None	None	Supervised (ImageNet weights)
9	RCV-URV	Universitat Rovira i Virgili (Spain)	EfficientNet-B5 + FPN + spatial attention (5x ensemble)	None	None	Supervised (ImageNet weights)
10	mnnl	NAAMII (Nepal)	1. ResNet-50 + Mask R-CNN + Transformer encoder (temporal 10-frame); 2. ResNet-50 + Mask R-CNN + Transformer encoder (single frame tokens); 3. LG-CVS	Transformer	Reconstruction Loss, MSE to average rater score	Supervised (Endoscapes-BBox201)
11	Pandas	ChengDu Withai Innovations Technology Company (China)	ConvNeXt V2	None	SimCLR-style contrastive loss (NT-Xent)	Supervised (ImageNet weights)
12	Theator	Theator (Israel)	EVA02-Large Transformer (6x model soup ensemble)	Transformer	Log-cosh for confidence estimation, BCE for video-level CVS labels, Focal loss for pre-temporal component classification	Supervised (ImageNet-1k weights)
13	Transformers	University Hospital Cologne (Germany)	U-Net (EfficientNet-B0) + Vision Transformer (ViT)	None	Segmentation	Supervised (Endoscapes for U-Net)

supervise confidence scores. The model predicts frame-level outputs for all three CVS criteria.
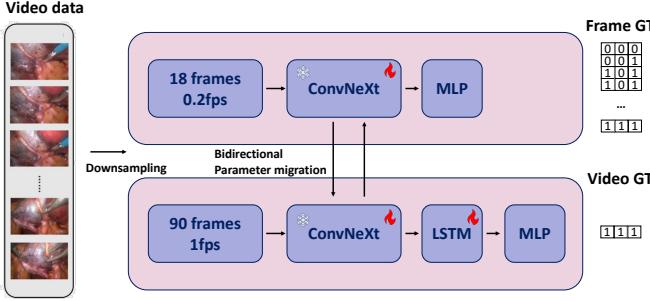

Fig. 8. Overview of various stages of CVS HUST method.

5.2.4. Team SDS-HD: *Multitask Transformer with Pseudo-Labels and Uncertainty-Aware Training*

This method formulates CVS prediction as a multitask learning problem using the EVA02-Large Vision Transformer (Fang et al., 2024) as backbone. The model simultaneously learns to classify CVS criteria and perform segmentation, with the two objectives optimized jointly using a weighted combination of binary cross-entropy, Dice, and Lovász losses. Classification is supervised using a smoothed version of the mean rater score per frame. Specifically, rater labels of 0, 0.33, 0.66, 1 are mapped to 0.15, 0.3975, 0.645, 0.90 using a linear label smoothing function to mitigate overconfidence and reflect annotator disagreement.

Segmentation is guided by pseudo-labels from a YOLOv8 model pretrained on Endoscapes (Jocher et al., 2023; Murali et al., 2023a), which are used during training but not inference. In parallel, a second EVA02-Large model pretrained with MoCoV2 (Chen et al., 2020) on Cholec80 (Twinanda et al., 2016), HeiChole, and Endoscapes datasets was trained with the same smoothed labels, forming a complementary representation stream. Finally, both models were used to generate 1 FPS pseudo-labels for the training set.

The final submission ensembles five models trained via 5-fold cross-validation on the 700 training videos (140 per fold), grouped by video and stratified by CVS prevalence.

5.2.5. Team SRV-WEISS: *Hybrid Segmentation–Classification Pipeline for CVS Prediction*

This method combines segmentation- and classification-based representations to predict CVS criteria. The architecture uses a Dense Prediction Transformer (DPT)(Ranftl et al., 2021) with an SSL-pretrained EndoViT encoder (Batić et al., 2024) for anatomical and tool segmentation, and an EfficientNet classifier for high-level recognition of surgical elements. Both models were additionally pretrained on subsets of the Endoscapes dataset and fine-tuned on the challenge task. For each input frame, segmentation masks from DPT are concatenated with the RGB image and passed through a ConvNeXt-tiny encoder, while EfficientNet outputs six class probabilities corresponding to anatomical and tool categories. These outputs are fused and processed by a two-layer multilayer perceptron (MLP) to generate final sigmoid predictions for each CVS criterion.

To adapt to the expanded input dimensionality caused by concatenated segmentations, the ConvNeXt input layer was modified using a weight averaging strategy (Wang et al., 2016). All predictions were made per-frame, and training was fully supervised using a weighted binary cross-entropy loss. The weighting was conditioned on the prevalence of each class in the training data. The DPT weights were frozen during training to reduce overfitting and stabilize learning.

5.2.6. Team Ostrich: *DenseNet Ensemble for Multi-Criteria CVS Prediction*

This method frames CVS assessment as a multi-label classification task and employs an ensemble of five independently trained DenseNet-121 (Huang et al., 2017) models to improve robustness and generalization. Each model is fine-tuned from ImageNet weights and adapted for multi-label prediction by modifying the final classification head to output probabilities for the three CVS criteria.

To address class imbalance, the team used weighted binary cross-entropy loss. During inference, predictions from the five models were averaged to form the final output. All predictions were made independently for each frame, with no additional temporal modeling, attention mechanisms, or auxiliary tasks. Data augmentation included affine transformations, horizontal flipping, and color jittering. This setup prioritized architectural simplicity and training stability while leveraging ensemble diversity for better performance across criteria.

5.2.7. Team FightTumor: *ConvNeXt-Baseline with AutoAugment for CVS Prediction*

This method adopts a straightforward multi-label classification architecture consisting of a ConvNeXt-B backbone and a simple multilayer perceptron (MLP) head. The model outputs sigmoid probabilities for each of the three CVS criteria per frame. Training is fully supervised using a weighted binary cross-entropy loss to address class imbalance.

To enhance generalization, the team applied AutoAugment (Cubuk et al., 2019) as the primary data augmentation strategy. The model was initialized with ImageNet-pretrained weights, and no auxiliary losses, attention mechanisms, or temporal modeling were used. All predictions are generated independently for each frame using a single-pass, end-to-end model.

5.2.8. Team HFUT-Media: *Vision Transformer with Annotator-Averaged Multi-Label Classification*

This method uses a Vision Transformer (ViT) (Dosovitskiy et al., 2020) backbone followed by a multilayer perceptron (MLP) head to perform multi-label classification of the three CVS criteria. The ViT encoder, pretrained on ImageNet, extracts spatial features which are passed through a two-layer MLP including ReLU, dropout, and batch normalization. The output is processed via sigmoid activation to produce per-criterion frame-level probabilities.

Training is fully supervised using the binary cross-entropy loss, with no auxiliary tasks or temporal modeling.

5.2.9. Team IRCV-URV: *EfficientNet-FPN with Attention Mechanisms for CVS Criteria Classification*

This method uses EfficientNet-B5 (Tan and Le, 2019) as a backbone, integrated with a Feature Pyramid Network (FPN) with spatial attention to classify the three CVS criteria at the frame level. The architecture is designed to extract multi-scale contextual features relevant to CVS, with the FPN enabling hierarchical feature aggregation and the spatial attention mechanism refining focus on anatomically significant regions implicitly.

Each frame is processed independently using a classification head composed of convolutional and fully connected layers with dropout regularization. The model is trained using a binary cross-entropy loss with full supervision, using 5-fold cross-validation (80-20 splits), with the final model averaging.

5.2.10. **Team mmll:** *Temporal Object-Token Transformer with Graph-Based Ensemble*

This method ensembles three complementary models for frame-level CVS prediction: a temporal object-token transformer, a non-temporal variant, and a latent graph network. Each model leverages structured object-level representations derived from a Mask R-CNN (He et al., 2017) detector trained on Endoscapes-Seg50 (Murali et al., 2023c). The detector produces class-labeled bounding boxes, which are used to extract visual features from a separate ResNet-50 backbone. For each object, these features are concatenated with class embeddings and detection scores to form object tokens. To capture spatial relationships, relative layout embeddings are computed between every token pair based on bounding box geometry, and injected as attention biases during transformer encoding.

The temporal model processes object tokens from the current frame and nine preceding frames, supplemented by a global token representing each frame's overall features. Absolute positional embeddings encode frame order, while spatial biases remain relative. All tokens pass through a 4-layer transformer encoder; the final global token is used for frame-level classification. For the non-temporal model, only the current frame's objects are used. Each model predicts CVS criteria using MSE loss against the mean annotator label. Final predictions are produced by a weighted ensemble, giving the temporal model twice the weight of each static model.

5.2.11. **Team Pandas:** *ConvNeXt V2 with Contrastive Learning and Color-Based Augmentation*

This method builds on the ConvNeXt V2 architecture for multi-label classification of the three CVS criteria. To enhance feature discrimination, the team incorporates a contrastive learning objective during training, encouraging the model to learn subtle distinctions between frames with different CVS states.

In addition to architectural tuning, the team applies extensive color-based augmentation to improve generalization under diverse imaging conditions. Transformations include color jittering across brightness, contrast, saturation, and hue.

5.2.12. **Team theater:** *Leveraging visual and temporal transformers for CVS Criteria Analysis*

This method integrates vision-language pretrained transformer backbones with causal temporal modeling to predict CVS criteria and associated uncertainty. The team uses EVA-02-ViT-L, a large-scale Vision Transformer pretrained on ImageNet-1k, as the visual backbone. They train five such models via k-fold cross-validation and one on the full dataset, then aggregate them using the uniform recipe for model soups—an efficient weight-averaging ensembling method that improves generalization under distribution shifts.

The resulting model extracts one feature vector per second for each video, which serves as input to a causal temporal transformer composed of an encoder-decoder architecture. This temporal model predicts frame-level CVS achievement and simultaneously estimates per-frame uncertainty and video-level CVS

scores through dedicated heads. The frame model is optimized with focal loss, while the temporal model uses binary cross-entropy for classification and log-cosh loss for uncertainty estimation. Multi-view inference and RandAugment are employed during training to enhance robustness.

5.2.13. **Team Transformers:** *Vision Transformer with U-Net-Based Tissue Localization*

This method introduces a two-stage architecture designed to incorporate anatomical localization into CVS classification. The first stage uses a U-Net (Ronneberger et al., 2015) with an EfficientNet-B0 backbone to segment key anatomical structures involved in the Critical View of Safety. The segmentation model is trained on the Endoscapes-Seg50 dataset to leverage limited annotated surgical data.

The second stage combines the original video frames with the predicted segmentation masks and inputs them into a Vision Transformer (ViT) to classify CVS criteria. This integration allows the classification model to operate with localized anatomical context, potentially improving recognition of subtle cues.

5.3. *Comparative analysis of methods*

The purpose of this subsection is to summarize and categorize the methodological strategies employed by competing teams in the SAGES CVS Challenge. This categorical grouping provides a structured basis for subsequent performance analysis, allowing trends to be examined both within and across methodological types. Full details per team are given in Table 2.

Data strategy categories.

- **Standard general pretraining only:** Rely on generic computer vision dataset (e.g., ImageNet) pretraining, without additional surgical datasets. Teams: theator, IRCV-URV, HFUT-MedIA, Ostrich, CVS HUST, Pandas.

- **Surgical dataset–driven training:** Use models pretrained on surgical datasets before fine-tuning on CVS training data.

  - Detection pretraining on Endoscapes-BBox201: mmll (Mask R-CNN detector).

  - Segmentation pretraining on Endoscapes-Seg50: SRV-WEISS (EndoViT encoder) and SDS-HD (YOLOv8 pseudo-labels)

  - CVS labels extended from Endoscapes-CVS201: Farm

- **Self-supervised surgical pretraining:** SDS-HD (MoCoV2 on Cholec80, HeiChole, Endoscapes). TUE-VCA (DINO on multi-procedure SurgeNet + GenSurgery datasets). SRV-WEISS (EndoViT initialized with pretrained weights across several distinct surgical procedures (Batić et al., 2024))

  Note that SDS-HD and SRV-WEISS employ both self-supervised pretraining and fully supervised training on additional surgical datasets.

Architecture categories.

- **CNN:** FightTumor (ConvNeXt-B), Pandas (ConvNeXt V2), IRCV-URV (EfficientNet-B5 + Feature Pyramid Network with spatial attention), Ostrich (DenseNet-121).

- **Transformer:** Theator & SDS-HD (EVA-02-ViT-L), HFUT-MedIA (Vision Transformer), TUE-VCA (PVT-v2).

- **Hybrid CNN + Transformer:** Transformers (U-Net with EfficientNet-B0 backbone for segmentation + ViT for classification), SRV-WEISS (Dense Prediction Transformer with EndoViT encoder + EfficientNet + ConvNeXt + MLP), mmll (ResNet-50 backbone + Mask R-CNN object detector + Transformer encoders + LG-CVS), Farm (DINOv2, SigLIP, ConvNeXt2-L, InternImage backbones paired with TCN).

Temporal modeling.

- **Used temporal modeling:** Theator (temporal transformer), mmll (temporal object-token transformer), Farm (stacked dilated TCNs in some ensemble members), CVS HUST (LSTM).

- **No temporal modeling:** All other teams.

Ensembling.

- **Used ensembling:** mmll (3 heterogeneous graph- and transformer-based models), Ostrich (5 DenseNet-121), Theator (6 EVA-02-ViT-L via model soups), Farm (7 varied backbones), TUE-VCA (5× PVT-v2).

- **No ensembling:** All other teams.

Learning objective categories.

- **Classification only:** Ostrich, FightTumor, HFUT-Media, IRCV-URV, Farm

- **Classification + auxiliary objectives:** SDS-HD (segmentation loss from YOLOv8 pseudo-labels, label smoothing), SRV-WEISS (segmentation + classification fusion), Transformers (segmentation feeding ViT classification), mmll (graph-based reconstruction loss (Murali et al., 2023c), MSE loss to average rater score), Pandas (contrastive loss), TUE-VCA (L2 reconstruction loss during distillation), Theator (log-cosh loss for uncertainty, BCE for video-level CVS labels, focal loss for backbone), CVS-HUST (ℓ_1 loss for confidence)

Note that SDS-HD, mmll, Theator, CVS-HUST all perform some kind of optimization for uncertainty or confidence estimation.

6. Results and discussion

6.1. Overview

Headline gains. Table 3 shows large, measurable improvements over the LG-DG baseline across all three subchallenges. In subchallenge A (mAP), the best score is 69.09 (Farm), a +10.04 absolute gain over LG-DG at 59.05, or a 17.0% relative improvement; 6 of 13 teams surpass the baseline. In subchallenge B (Brier; lower is better), the best score is 0.022 (theator), cutting error by 0.102 points versus 0.124 for LG-DG, an 82% reduction; 12 of 13 teams beat the baseline. In subchallenge C (DRS using the consistent-subset robust-min mAP), the best score is 59.06 (SDS-HD), a +8.44 absolute gain over 50.62 for LG-DG, or a 16 .7% relative improvement; 6 of 13 teams exceed the baseline. Taken together, these headline gains indicate the challenge catalyzed tangible advances over the state of the art.

Common traits of top-performing entries (across subchallenges). To provide a compact view of what worked well, we summarize patterns among leaders in each subchallenge using the methodological taxonomy in Section 5.3:

- **Architectures:** Transformer or hybrid transformer–cnn models dominate the very top. Among teams appearing in any top-3 across A/B/C (theator, sds-hd, farm, pandas), 3 of 4 use transformer-based or hybrid designs. In the broader top-5 sets across the three tasks, 7 of 8 unique teams rely on transformers or hybrids.

- **Pretraining beyond imagenet:** Leveraging surgical or self-supervised surgical pretraining is common among accuracy/robustness leaders. In both subchallenges A and C, 4 of the top 5 teams use either surgical datasets or self-supervised pretraining on surgical video. Subchallenge B rankings are more mixed on this dimension.

- **Ensembling:** Ensembling is frequently present among Subchallenge A and C leaders but rare among calibration leaders. In both subchallenge A and C, 4 of the top 5 use ensembles; however, in subchallenge B, only 1 of the top 5 ensembles.

- **Temporal modeling:** temporal components appear in 3 of the top 5 teams in subchallenge A and C, despite only 4 of 13 submissions explicitly leveraging temporal context.

- **Learning objectives:** Auxiliary objectives are ubiquitous among top performers. All of the top 5 in subchallenge B employ auxiliary losses or multi-task formulations (e.g., uncertainty, segmentation, contrastive, or distillation objectives). In subchallenges A and C, 4 of the top 5 also incorporate auxiliary objectives.

6.2. Cross–subchallenge relationships

Using the ranks in Table 4, A and C track each other closely (Spearman = 0.96), while A vs B and B vs C are only weakly aligned (both 0.38). This is consistent with Table 3: teams that classify well (A) also tend to be robust (C), whereas B behaves more independently.

Method traits by top five counting presence within the top five of each subchallenge:

- **Pretraining beyond ImageNet:** A 4/5, C 4/5, B 2/5

- **Ensembling:** A 4/5, C 4/5, B 1/5

- **Temporal modeling:** A 3/5, C 3/5, B 1/5

in short, the ingredients that lead to robust performance are not prerequisites for capturing clinical ambiguity.

Table 3. We show detailed metrics for each subchallenge per CVS criteria (C1–C3) and the average across them. For Subchallenge C (Domain Robustness Score; DRS), we use a *robust-min* evaluation: from the 10 predefined variation splits we drop the worst 10% (one split) and take the minimum mAP over the remaining splits. We apply this independently to the overall score (computed by first averaging across C1–C3 per sample) and to each criterion (C1–C3). As a result, the average of the per-criterion DRS (C1–C3) may not equal the "Avg" DRS, which is obtained by applying the Domain Robustness Score on the average mAP across criteria. The baseline *LG-DG* is included and italicized.

Subchallenge A (mAP, ↑ is better)					Subchallenge B (Brier, ↓ is better)					Subchallenge C (DRS, ↑ is better)				
Team	C1	C2	C3	Avg	Team	C1	C2	C3	Avg	Team	C1	C2	C3	Avg
Farm	56.65	85.30	65.32	**69.09**	theator	0.024	0.023	0.020	**0.022**	SDS-HD	41.51	73.51	41.13	**59.06**
theator	59.06	83.61	63.94	**68.87**	Pandas	0.025	0.025	0.020	**0.023**	theator	39.85	76.39	46.88	**58.11**
SDS-HD	56.53	83.89	65.99	**68.80**	SDS-HD	0.026	0.026	0.021	**0.024**	Farm	37.08	77.13	35.35	**57.71**
mmll	54.92	78.91	58.96	**64.26**	SRV-WEISS	0.028	0.042	0.028	**0.033**	mmll	46.37	67.64	36.37	**56.33**
TUE-VCA	48.12	80.60	59.94	**62.89**	Transformers	0.032	0.050	0.031	**0.038**	TUE-VCA	32.71	71.61	30.57	**53.13**
Pandas	48.24	79.82	56.42	**61.50**	CVS HUST	0.038	0.055	0.036	**0.043**	Pandas	36.86	70.39	47.74	**51.66**
LG-DG (baseline)	48.85	73.42	54.89	**59.05**	mmll	0.047	0.046	0.040	**0.044**	*LG-DG (baseline)*	40.87	60.61	37.47	**50.62**
FightTumor	40.85	76.34	47.15	**54.78**	TUE-VCA	0.059	0.050	0.046	**0.052**	Ostrich	19.72	63.79	16.82	**49.24**
Ostrich	40.35	75.32	44.13	**53.27**	Farm	0.063	0.057	0.055	**0.058**	FightTumor	27.81	63.22	30.91	**42.83**
SRV-WEISS	38.16	66.27	42.29	**48.91**	Ostrich	0.092	0.091	0.077	**0.086**	IRCV-URV	26.00	59.83	14.32	**41.38**
IRCV-URV	34.32	69.42	37.19	**46.98**	IRCV-URV	0.102	0.108	0.094	**0.101**	SRV-WEISS	31.21	54.21	25.94	**40.89**
Transformers	12.98	28.94	18.17	**20.03**	FightTumor	0.103	0.109	0.093	**0.102**	Transformers	11.39	23.50	6.44	**18.01**
CVS HUST	13.76	25.14	9.15	**16.01**	*LG-DG (baseline)*	0.130	0.119	0.121	**0.124**	HUFT-MedIA	8.01	21.94	7.80	**13.65**
HUFT-MedIA	9.91	23.44	9.51	**14.29**	HUFT-MedIA	0.119	0.172	0.110	**0.134**	CVS HUST	8.11	19.58	5.69	**11.65**

Table 4. Ranks for Subchallenge A (CVS Classification), Subchallenge B (Uncertainty Quantification), and Subchallenge C (Robustness) are shown. A lower rank number indicates a better performance.

Team	Overall Rank	Subchallenge A	Subchallenge B	Subchallenge C
theator	1	2	1	2
SDS-HD	2	3	3	1
Farm	3	1	9	3
Pandas	4	6	2	6
mmll	5	4	7	4
TUE-VCA	6	5	8	5
SRV-WEISS	7	9	4	10
Ostrich	8	8	10	7
FightTumor	9	7	12	8
Transformers	10	11	5	11
IRCV-URV	11	10	11	9
CVS_HUST	12	12	6	13
HFUT-MedlA	13	13	13	12

Table 5. Overall and per-criterion macro-F1 and accuracy for the top five Subchallenge A teams (ordered by Subchallenge A mAP in Table 3), compared to expert references. Accuracy is reported as overall subset accuracy (all three criteria correct) and per-criterion accuracy; macro-F1 is reported overall and per criterion. Expert (upper bound) predicts the per-frame, per-criterion rater-mode label; Expert (lower bound) predicts the consensus on unanimous frames and the minority label on disagreement frames. All values are percentages.

	macro-F1				accuracy			
Team	Overall	C1	C2	C3	Overall	C1	C2	C3
Expert (upper bound)	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Farm	62.94	53.64	76.09	59.10	77.65	89.85	89.35	92.26
theator	64.30	55.34	75.13	62.42	75.59	89.93	87.59	92.24
SDS-HD	61.12	51.31	74.41	57.64	77.33	90.37	88.33	92.41
mmll	59.35	52.20	71.54	54.30	73.11	88.74	86.31	90.06
TUE-VCA	53.81	39.05	69.76	52.63	76.28	89.54	87.20	91.83
Expert (lower bound)	36.89	31.12	47.44	32.11	56.96	77.30	72.83	82.15

6.3. Focused Analysis: Overall Performance

Table 5 shows clear criterion-wise differences within Subchallenge A. Macro-F1 is consistently highest on C2 and lowest on C1 (with C3 in between), while per-criterion accuracies remain uniformly high (about 86–92 percent). This gap between accuracy and macro-F1 is consistent with the fact that most frames capture negative examples of the CVS criteria: teams (and the mode-of-raters labels) agree on negatives, but miss more positives—macro-F1 exposes those misses in a way accuracy does not.

To anchor these results, we compare against two expert references, evaluated against the same mode-of-raters labels. Expert (upper bound) predicts the per-frame, per-criterion mode label and thus attains 100 percent by construction. Expert (lower bound) predicts the consensus on unanimous frames and the minority label on frames where there was partial agreement between the 3 annotators. These robust expert references are based on 19 annotators who underwent the training and quality–control pipeline described in Section 3 and, within the 300–video test set, each annotated 47.4±25.0 videos on average. The top five Subchallenge A teams in Table 5 exceed this floor by an average of about 23.4 macro-F1 points (60.30 vs 36.89) and 19.0 accuracy points (75.99 vs 56.96), indicating performance that these systems are approaching expert-level capability at CVS assessment.

6.4. Focused Analysis: Calibration (Subchallenge B)

Subchallenge B evaluates how well a model's predicted probabilities align with a confidence-aware target (Section 4): the target approaches 0.5 when raters are uncertain and moves toward 0 or 1 when they are confident. Figure 9 visualizes, per team and per criterion (C1–C3), the signed error $p - y_{\text{conf}}$ (predicted probability minus confidence-aware label). Distributions centered tightly around 0 indicate good calibration; large absolute deviations and wide spreads indicate miscalibration.

Who is calibrated? Rankings mirror the violins. Top Subchallenge B teams theator, Pandas, and SDS-HD show small biases and relatively narrow spreads across criteria (means within about ±0.06).

Auxiliary objectives, uncertainty objectives, and localization. All 7 top-ranking teams in subchallenge B, without exception, make use of auxiliary objectives. Among them, the top performers (theator, Pandas, SDS-HD) achieve tight, near-zero distributions, suggesting auxiliary supervision can help calibration when paired with strong training recipes. Explicit uncertainty objectives (SDS-HD, theator, mmll, CVS HUST) are not a guarantee on their own: theator and SDS-HD are well-centered and near the top; CVS HUST is mixed; mmll remains negatively biased. Similarly, using localization signals (SDS-HD, SRV-WEISS, Transformers, mmll) yields heterogeneous
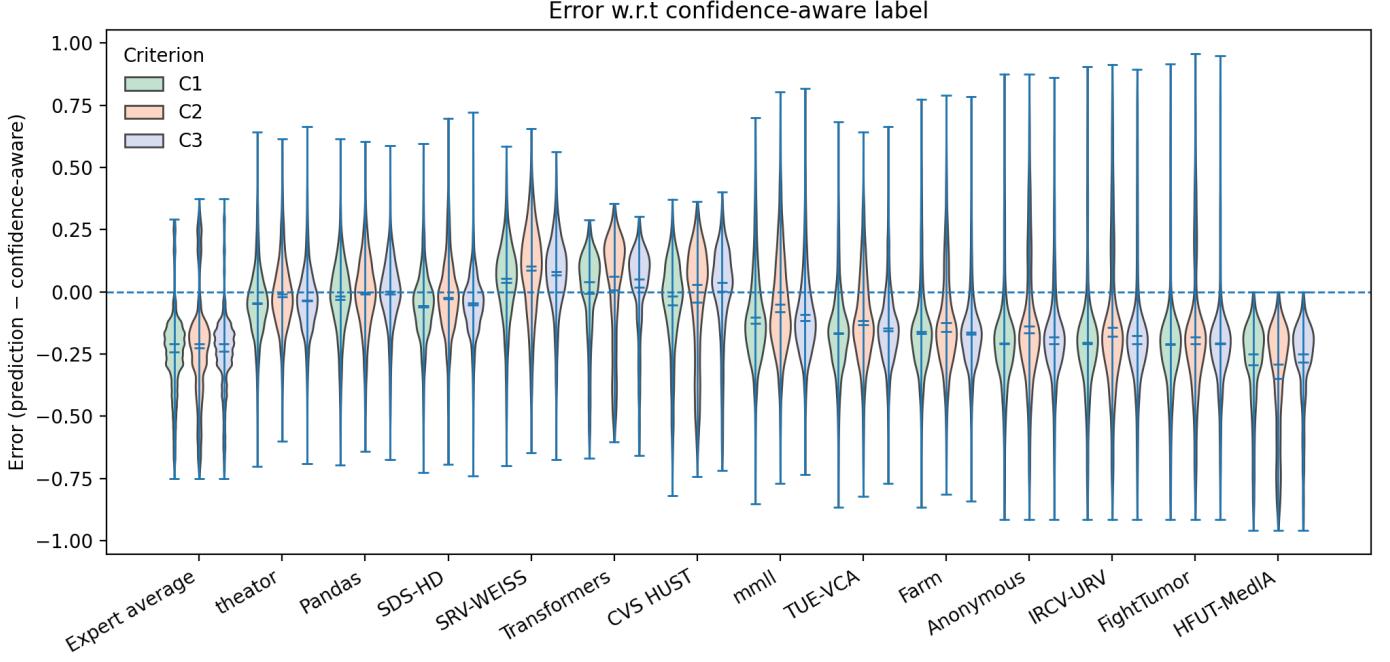
Fig. 9. Violin plots of signed error relative to the confidence-aware label described in Section 4. Each group shows one team, with the three CVS criteria (C1, C2, C3) in different colors. Error is defined as predicted probability minus the confidence-aware target for a criterion; zero (dashed line) means perfect alignment with the target, while larger absolute values indicate greater deviation. Positive values indicate a tilt toward the positive label relative to the target; negative values indicate a tilt toward the negative label. The Expert average violin on the left shows the difference between the mean rater label and the confidence-aware label, providing a human reference. Teams are displayed left to right in order of Subchallenge B performance.

outcomes: excellent for SDS-HD, positively tilted for SRV-WEISS, broader for Transformers, and under-confident (negative) for mmll. In short, how these ingredients are integrated matters as much as their presence.

Direction of error and what models learn. Relative to the confidence-aware target, most mid- and lower-ranked systems skew negative (probabilities below y_{conf}), indicating conservative estimates when raters express partial confidence. Top teams nevertheless sit much closer to y_{conf} than to the hard mean label, indicating they learn to place probabilities in the soft region that reflects rater uncertainty, whether they learn to do so implicitly or explicitly.

A note on Farm. Farm is typically top-3 in Subchallenges A/C, yet shows sizable negative bias and wider spread in Subchallenge B, aligning with its drop in Brier ranking in Table 3. This supports the takeaway that optimizing for discriminative performance (for example, mAP) does not inherently yield confidence alignment; calibration requires targeted choices.

6.5. Focused Analysis: Robustness across Variant Splits

Figure 10 summarizes each team's behavior by plotting mean mAP across the variant test splits of Subchallenge C against its standard deviation. We did not observe statistically significant methodological patterns, within the submitted methods, linking specific architectural or training choices to robustness. Instead, teams populate all four quadrants of the plot: some achieve high means with low dispersion, others reach similar means with notably higher dispersion, and several exhibit uniformly lower means with either narrow or wide spread. This heterogeneity reinforces a practical takeaway for deployment: optimizing average performance alone is not sufficient. Methods that explicitly prioritize, or at minimum validate, equitable
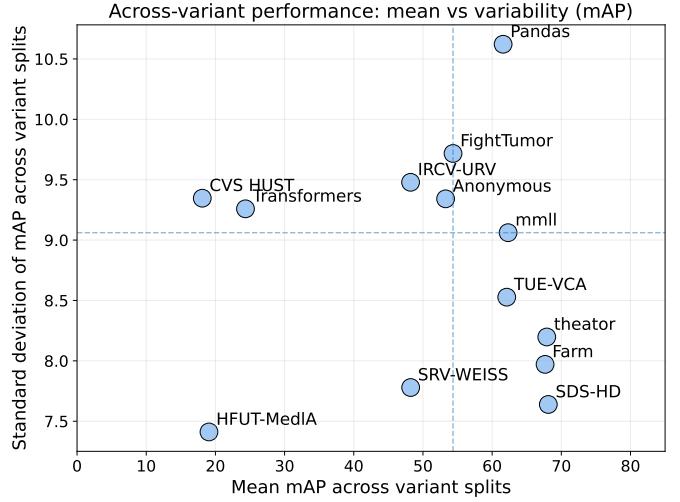

Fig. 10. Mean mAP (x-axis) vs. standard deviation across variant splits (y-axis) for each team. Toward the right indicates better average performance; toward the bottom indicates more consistent performance across splits.

performance across different data distributions are needed to ensure responsible and fair real-world deployment.

7. Limitations

Our work comes with several limitations that we would like to be transparent about.

A limitation of challenge design in general, including ours, is that it prioritises broad exploration of design choices over definitive conclusions about any single choice. In return, we surfaced clear, community-relevant signals across diverse ar-

chitectures, auxiliary objectives, and training datasets. The immediate value is a sharper starting point for focused, controlled ablations and shared training recipes that can isolate what truly drives gains.

Our confidence-aware ground truth was designed to be practical for experts and clinically meaningful. Annotator confidence was captured at the video level, and labels were multiplexed so that predictions are encouraged to move toward 0.5 when raters are unsure and toward 0 or 1 when raters are confident. This makes calibration an explicit objective rather than an afterthought. Complementary next steps include testing finer-grained confidence, learning separate confidence heads, and comparing alternative probabilistic targets, which should further align scores with the uncertainty clinicians naturally express.

This manuscript concentrates on methodological trends across submissions rather than a full data-centric analysis of performance across all dataset attributes. The dataset and manual quality assessments enable several high-value directions. First, the quality assessments themselves constitute one of the largest video-based initiatives in surgery, even without the AI component. Mining these assessments for predictors of difficult frames, and using those signals to trigger targeted model development, could directly benefit clinical reliability. Second, the data support stratified analyses that each warrant independent investigation, for example performance in procedures from LMIC settings, under-represented countries, or other cohort characteristics. In this work we prioritised a single coherent synthesis of methodological trends to provide clear takeaways the community can build on.

8. Conclusion

The SAGES Critical View of Safety (CVS) Challenge brought together clinicians and engineers, industry and academia, in a coordinated effort to address a single, clinically relevant problem: the automated assessment of a key surgical safety step. CVS serves as an exemplar of surgical quality assessment, clear in its definition, supported by strong clinical evidence, yet inconsistently performed worldwide. The resulting benchmark spans continents, involves data from 54 hospitals, and engaged hundreds of clinicians in its creation, offering a rare real-world testbed for AI-based surgical quality assessment.

Delivering this benchmark required infrastructure capable of sustaining global collaboration at scale. The EndoGlacier framework enabled the orchestration of diverse video and annotation flows, rigorous multi-annotator quality control, and the coordination of contributors across institutions and countries. We hope its design will serve as a reproducible model for future large-scale benchmarks in surgery.

The challenge attracted submissions from 13 international teams and produced substantial advances over the baseline across all evaluation axes. In Subchallenge A, the best method improved mean average precision by 17% relative; in Subchallenge B, the top performer reduced Brier error by over 80%; and in Subchallenge C, the leading method achieved a 17% relative gain in domain robustness. These results underscore both the potential of current methods and the importance of evaluating accuracy, calibration, and robustness together when developing clinically deployable systems.

Methodological trends observed across top-performing submissions, including the effectiveness of transformer-based or hybrid architectures, surgical or self-supervised surgical pre-training, temporal modeling, and auxiliary learning objectives, provide a springboard for future research. As such, this benchmark not only measures progress but also guides it, helping the field move toward robust, trustworthy AI systems capable of improving surgical safety and quality worldwide.

CRediT authorship contribution statement

DA : Conceptualization, Data Curation, Data Analysis and Interpretation, Methodology, Software, Investigation, Validation, Evaluation, Formal Analysis, Visualization, Writing - Original Draft, Writing - Review & Editing, Challenge Organization, Resources. **DN, OB, AY, JH, XW, RL, LL, YW, SK, BB, TJ, ZM, AW, JM, YX, ZW, AMOA, HAFR, BS, KY, YZ, HW** : Methodology, Software, Writing - Review & Editing. **PI** : Conceptualization, Writing - Review & Editing, Supervision, Challenge Organization, Resources, Funding Acquisition, Project Administration.

References

Al Hajj, H., Lamard, M., Conze, P.H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O., Dedmari, M.A., Zhao, F., Prellberg, J., et al., 2019. Cataracts: Challenge on automatic tool annotation for cataract surgery. Medical image analysis 52, 24–41.

Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al., 2020. 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 .

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al., 2019. 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 .

Avgerinos, C., Kelgiorgi, D., Touloumis, Z., Baltatzi, L., Dervenis, C., 2009. One thousand laparoscopic cholecystectomies in a single surgical unit using the "critical view of safety" technique. Journal of Gastrointestinal Surgery 13, 498–503.

Bai, S., Kolter, J.Z., Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271 .

Ban, Y., Eckhoff, J.A., Ward, T.M., Hashimoto, D.A., Meireles, O.R., Rus, D., Rosman, G., 2023. Concept graph neural networks for surgical video understanding. IEEE Transactions on Medical Imaging 43, 264–274.

Bano, S., Casella, A., Vasconcelos, F., Moccia, S., Attilakos, G., Wimalasundera, R., David, A.L., Paladini, D., Deprest, J., De Momi, E., et al., 2021. Fetreg: Placental vessel segmentation and registration in fetoscopy challenge dataset. arXiv preprint arXiv:2106.05923 .

Batić, D., Holm, F., Özsoy, E., Czempiel, T., Navab, N., 2024. Endovit: pretraining vision transformers on a large collection of endoscopic images. International Journal of Computer Assisted Radiology and Surgery 19, 1085–1091.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W., 2010. A theory of learning from different domains. Machine learning 79, 151–175.

Birkmeyer, J.D., Finks, J.F., O'Reilly, A., Oerline, M., Carlin, A.M., Nunn, A.R., Dimick, J.B., 2013. Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine 369, 1434–1442. doi:10.1056/NEJMsa1300625.

Bose, R., Nwoye, C.I., Lazo, J., Lavanchy, J.L., Padoy, N., 2025. Feature mixing approach for detecting intraoperative adverse events in laparoscopic roux-en-y gastric bypass surgery. arXiv preprint arXiv:2504.16749 .

Brunt, L.M., Deziel, D.J., Telem, D.A., Strasberg, S.M., Aggarwal, R., Asbun, H., Bonjer, J., McDonald, M., Alseidi, A., Ujiki, M., et al., 2020. Safe cholecystectomy multi-society practice guideline and state-of-the-art consensus conference on prevention of bile duct injury during cholecystectomy. Surgical endoscopy 34, 2827–2855.

Carroll, B., Birth, M., Phillips, E., 1998. Common bile duct injuries during laparoscopic cholecystectomy that result in litigation. Surgical endoscopy 12, 310–314.

Chen, X., Fan, H., Girshick, R., He, K., 2020. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 .

Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V., 2019. Autoaugment: Learning augmentation strategies from data, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 113–123.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Eppler, M.B., Sayegh, A.S., Maas, M., Venkat, A., Hemal, S., Desai, M.M., Hung, A.J., Grantcharov, T., Cacciamani, G.E., Goldenberg, M.G., 2023. Automated capture of intraoperative adverse events using artificial intelligence: a systematic review and meta-analysis. Journal of Clinical Medicine 12, 1687.

Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y., 2024. Eva-02: A visual representation for neon genesis. Image and Vision Computing 149, 105171.

Grüter, A.A., Van Lieshout, A.S., van Oostendorp, S.E., Henckens, S.P., Ket, J.C., Gisbertz, S.S., Toorenvliet, B.R., Tanis, P.J., Bonjer, H.J., Tuynman, J.B., 2023. Video-based tools for surgical quality assessment of technical skills in laparoscopic procedures: a systematic review. Surgical endoscopy 37, 4279–4297.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al., 2022. Lora: Low-rank adaptation of large language models. ICLR 1, 3.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Jaspers, T.J., de Jong, R.L., Al Khalil, Y., Zeelenberg, T., Kusters, C.H., Li, Y., van Jaarsveld, R.C., Bakker, F.H., Ruurda, J.P., Brinkman, W.M., et al., 2024. Exploring the effect of dataset diversity in self-supervised learning for surgical computer vision, in: MICCAI Workshop on Data Engineering in Medical Imaging, Springer. pp. 43–53.

Jocher, G., Chaurasia, A., Qiu, J., 2023. Ultralytics yolov8. URL: https://github.com/ultralytics/ultralytics.

Korndorffer Jr, J.R., Hawn, M.T., Spain, D.A., Knowlton, L.M., Azagury, D.E., Nassar, A.K., Lau, J.N., Arnow, K.D., Trickey, A.W., Pugh, C.M., 2020. Situating artificial intelligence in surgery: a focus on disease severity. Annals of surgery 272, 523–528.

Kumar, A., Ma, T., Liang, P., Raghunathan, A., 2022. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift, in: Uncertainty in Artificial Intelligence, PMLR. pp. 1041–1051.

Kurashima, Y., Kitagami, H., Teramura, K., Poudel, S., Ebihara, Y., Inaki, N., Nakamura, F., Misawa, K., Shibao, K., Nagai, E., et al., 2022. Validation study of a skill assessment tool for education and outcome prediction of laparoscopic distal gastrectomy. Surgical Endoscopy 36, 8807–8816.

Lam, K., Chen, J., Wang, Z., Iqbal, F.M., Darzi, A., Lo, B., Purkayastha, S., Kinross, J.M., 2022. Machine learning for technical skill assessment in surgery: a systematic review. NPJ digital medicine 5, 24.

Laplante, S., Namazi, B., Kiani, P., Hashimoto, D.A., Alseidi, A., Pasten, M., Brunt, L.M., Gill, S., Davis, B., Bloom, M., et al., 2023. Validation of an artificial intelligence platform for the guidance of safe laparoscopic cholecystectomy. Surgical endoscopy 37, 2260–2268.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11976–11986.

Luengo, I., Grammatikopoulou, M., Mohammadi, R., Walsh, C., Nwoye, C.I., Alapatt, D., Padoy, N., Ni, Z.L., Fan, C.C., Bian, G.B., et al., 2021. 2020 cataracts semantic segmentation challenge. arXiv preprint arXiv:2110.10965 .

Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., et al., 2022. Surgical data science–from concepts toward clinical translation. Medical image analysis 76, 102306.

Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P.M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., et al., 2021. Heidelberg colorectal data set for surgical data science in the sensor operating room. Scientific data 8, 101.

Mascagni, P., Alapatt, D., Garcia, A., Okamoto, N., Vardazaryan, A., Costamagna, G., Dallemagne, B., Padoy, N., 2021a. Surgical data science for safe cholecystectomy: a protocol for segmentation of hepatocystic anatomy and assessment of the critical view of safety. arXiv preprint arXiv:2106.10916 .

Mascagni, P., Alapatt, D., Lapergola, A., Vardazaryan, A., Mazellier, J.P., Dallemagne, B., Mutter, D., Padoy, N., 2024. Early-stage clinical evaluation of real-time artificial intelligence assistance for laparoscopic cholecystectomy. British Journal of Surgery 111, znad353.

Mascagni, P., Alapatt, D., Murali, A., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., et al., 2025. Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. Scientific Data 12, 331.

Mascagni, P., Alapatt, D., Urade, T., Vardazaryan, A., Mutter, D., Marescaux, J., Costamagna, G., Dallemagne, B., Padoy, N., 2021b. A computer vision platform to automatically locate critical events in surgical videos: documenting safety in laparoscopic cholecystectomy. Annals of surgery 274, e93–e95.

Mascagni, P., Vardazaryan, A., Alapatt, D., Urade, T., Emre, T., Fiorillo, C., Pessaux, P., Mutter, D., Marescaux, J., Costamagna, G., et al., 2022. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. Annals of surgery 275, 955–961.

Mazellier, J.P., Boujon, A., Bour-Lang, M., Erharhd, M., Waechter, J., Wernert, E., Mascagni, P., Padoy, N., 2023. Mosaic: a web-based platform for collaborative medical video assessment and annotation. arXiv preprint arXiv:2312.08593 .

McInnes, L., Healy, J., Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 .

Meara, J.G., Leather, A.J.M., Hagander, L., Alkire, B.C., Alonso, N., Ameh, E.A., Bickler, S.W., Conteh, L., Dare, A.J., Davies, J., et al., 2015. Global surgery 2030: evidence and solutions for achieving health, welfare, and economic development. The Lancet 386, 569–624. doi:10.1016/S0140-6736(15)60160-X.

Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., et al., 2023a. The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. arXiv preprint arXiv:2312.12429 .

Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Mutter, D., Padoy, N., 2023b. Encoding surgical videos as latent spatiotemporal graphs for object and anatomy-driven reasoning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 647–657.

Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Mutter, D., Padoy, N., 2023c. Latent graph representations for critical view of safety assessment. IEEE Transactions on Medical Imaging 43, 1247–1258.

Naik, N.D., Abbott, E.F., Gas, B.L., Murphy, B.L., Farley, D.R., Cook, D.A., 2018. Personalized video feedback improves suturing skills of incoming general surgery trainees. Surgery 163, 921–926.

Nwoye, C.I., Alapatt, D., Yu, T., Vardazaryan, A., Xia, F., Zhao, Z., Xia, T.,

Jia, F., Yang, Y., Wang, H., et al., 2023a. Cholectriplet2021: A benchmark challenge for surgical action triplet recognition. Medical Image Analysis 86, 102803.

Nwoye, C.I., Yu, T., Sharma, S., Murali, A., Alapatt, D., Vardazaryan, A., Yuan, K., Hajek, J., Reiter, W., Yamlahi, A., et al., 2023b. Cholectriplet2022: Show me a tool and tell me the triplet—an endoscopic vision challenge for surgical action triplet detection. Medical Image Analysis 89, 102888.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 .

Pryor, A.D., Lendvay, T., Jones, A., Ibáñez, B., Pugh, C., 2023. An american board of surgery pilot of video assessment of surgeon technical performance in surgery. Annals of Surgery 277, 591–595.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 12179–12188.

Ríos, M.S., Molina-Rodriguez, M.A., Londoño, D., Guillén, C.A., Sierra, S., Zapata, F., Giraldo, L.F., 2023. Cholec80-cvs: An open dataset with an evaluation of strasberg's critical view of safety for ai. Scientific Data 10, 194.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Roß, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., et al., 2021. Comparative validation of multi-instance instrument segmentation in endoscopy: results of the robust-mis 2019 challenge. Medical image analysis 70, 101920.

Sanjay, P., Fulke, J.L., Exon, D.J., 2010. 'critical view of safety'as an alternative to routine intraoperative cholangiography during laparoscopic cholecystectomy for acute biliary pathology. Journal of Gastrointestinal Surgery 14, 1280–1284.

Satyanaik, S., Murali, A., Alapatt, D., Wang, X., Mascagni, P., Padoy, N., 2024. Optimizing latent graph representations of surgical scenes for zero-shot domain transfer. arXiv preprint arXiv:2403.06953 .

Scally, C.P., Varban, O.A., Carlin, A.M., Birkmeyer, J.D., Dimick, J.B., Collaborative, M.B.S., et al., 2016. Video ratings of surgical skill and late outcomes of bariatric surgery. JAMA surgery 151, e160428–e160428.

Schmidgall, S., Kim, J.W., Jopling, J., Krieger, A., 2024. General surgery vision transformer: A video pre-trained foundation model for general surgery. arXiv preprint arXiv:2403.05949 .

Shao, M., Li, D., Zhao, C., Wu, X., Lin, Y., Tian, Q., 2024. Supervised algorithmic fairness in distribution shifts: a survey, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, pp. 8225–8233.

Strasberg, S.M., Hertl, M., Soper, N.J., 1995. An analysis of the problem of biliary injury during laparoscopic cholecystectomy. Journal of the American College of Surgeons 180, 101–125.

Stulberg, J.J., Huang, R., Kreutzer, L., Ban, K., Champagne, B.J., Steele, S.R., Johnson, J.K., Holl, J.L., Greenberg, C.C., Bilimoria, K.Y., 2020. Association between surgeon technical skills and patient outcomes. JAMA surgery 155, 960–968.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR. pp. 6105–6114.

Tsalis, K., Antoniou, N., Koukouritaki, Z., Patridas, D., Christoforidis, E., Lazaridis, C., 2015. Open-access technique and "critical view of safety" as the safest way to perform laparoscopic cholecystectomy. Surgical Laparoscopy Endoscopy & Percutaneous Techniques 25, 119–124.

Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N., 2016. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging 36, 86–97.

Varban, O.A., Thumma, J.R., Carlin, A.M., Finks, J.F., Ghaferi, A.A., Dimick, J.B., 2020. Peer assessment of operative videos with sleeve gastrectomy to determine optimal operative technique. Journal of the American College of Surgeons 231, 470–477.

Varban, O.A., Thumma, J.R., Finks, J.F., Carlin, A.M., Ghaferi, A.A., Dimick, J.B., 2021. Evaluating the effect of surgical skill on outcomes for laparoscopic sleeve gastrectomy: a video-based study. Annals of surgery 273, 766–771.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L., 2016. Temporal segment networks: Towards good practices for deep action recognition, in: European conference on computer vision, Springer. pp. 20–36.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al., 2023. Internimage: Exploring large-scale vision foundation models with deformable convolutions, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14408–14419.

Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2022. Pvt v2: Improved baselines with pyramid vision transformer. Computational visual media 8, 415–424.

Way, L.W., Stewart, L., Gantert, W., Liu, K., Lee, C.M., Whang, K., Hunter, J.G., 2003. Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective. Annals of surgery 237, 460–469.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S., 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16133–16142.

Yin, L., Ban, Y., Eckhoff, J., Meireles, O., Rus, D., Rosman, G., 2024. Hypergraph-transformer (hgt) for interactive event prediction in laparoscopic and robotic surgery. arXiv preprint arXiv:2402.01974 .

Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L., 2023. Sigmoid loss for language image pre-training, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 11975–11986.