



Advancing automated phase recognition in cataract surgery through the SICS-155 challenge

Simon Mueller^{a,b}, Bhuvan Sachdeva^{c,d}, Singri Niharika Prasad^d,
 Raphael Lechtenboehmer^b, Frank G Holz^b, Robert P Finger^e, Christopher Nielsen^{f,g},
 Patrick Gooi^h, Adrian Gooi^{i,j,k}, Nils D. Forkert^{l,m,n}, Yasser El Jarida^o, Youssef Iraqi^o,
 Loubna Mekouar^o, Yiran Kong^p, Zhihai Huang^p, Jiawei Du^p, Huiqi Li^p,
 Lasse Renz-Kiefel^{q,r}, Eric L. Wisotzky^{q,s}, Dong Zheng^t, Jiacheng Lin^t, Miao Hu^t,
 Yanwu Xu^u, Mohit Jain^c, Kaushik Murali^d, Maximilian W M Wintergerst^{b,v,*},
 Thomas Schultze^{a,w}

^a B-IT and Department of Computer Science, University of Bonn, Bonn, Germany

^b Department of Ophthalmology, University Hospital Bonn, Bonn, Germany

^c Microsoft Research, Bengaluru, India

^d Sankara Eye Hospital, Bengaluru, Karnataka, India

^e Department of Ophthalmology, University Medical Center Mannheim, Heidelberg University, Germany

^f Cumming School of Medicine, University of Calgary, Calgary, Canada

^g Biomedical Engineering Program, University of Calgary, Calgary, Canada

^h Department of Ophthalmology, University of Calgary, Calgary, Canada

ⁱ Max Rady College of Medicine, University of Manitoba, Winnipeg, Canada

^j Department of Otolaryngology, University of British Columbia, Vancouver, Canada

^k School of Medicine, Simon Fraser University, Burnaby, Canada

^l Department of Radiology, University of Calgary, Calgary, Canada

^m Hotchkiss Brain Institute, University of Calgary, Calgary, Canada

ⁿ Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Canada

^o College of Computing, University Mohammed VI Polytechnic, Morocco

^p School of Information and Electronics, Beijing Institute of Technology, Beijing, China

^q Computer Vision and Graphics, Vision and Imaging Technologies, Fraunhofer HHI, Berlin, Germany

^r Klinik und Poliklinik für Hals-Nasen-Ohrenheilkunde, Kopf- und Halschirurgie Otto Körner, Universitätsmedizin Rostock, Germany

^s Visual Computing, Humboldt University of Berlin, Berlin, Germany

^t Hang Zhou Dianzi University, Hangzhou, Zhejiang, China

^u South China University of Technology, Guangdong, China

^v Augenzentrum Grischun, Chur, Switzerland

^w Lamarr Institute for Machine Learning and Artificial Intelligence, Bonn, Germany

ARTICLE INFO

Keywords:

Manual small-incision cataract surgery
 Phacoemulsification
 Artificial intelligence
 Temporal action segmentation
 Phase recognition, Video dataset
 Vision transformers

ABSTRACT

Manual Small-Incision Cataract Surgery (SICS) is a prevalent technique in low- and middle-income countries (LMICs). Automated analysis of SICS videos based on artificial intelligence (AI) would benefit self-evaluation, training, monitoring, and ultimately facilitate computer aided surgical assistance. However, this procedure remains understudied in terms of automated surgical analysis due to a lack of publicly available data. To advance this field, we organized the SICS-155 phase recognition challenge at MICCAI 2025. It was based on a dataset with 155 videos recorded at Sankara Eye Hospital in Bengaluru, India with an average duration of 13:05 minutes, annotated with 19 distinct surgical phases.

In this work, we first present the results and findings of that challenge in accordance with the BIAS guidelines. Second, we integrated key technical innovations from the participating teams to develop a new, state-of-the-art approach for phase recognition in a post-operative, offline setting. This approach uses a boundary-aware FACT architecture, which we evaluated using videos of both SICS and phacoemulsification. Our approach achieved an

* Corresponding author at: Department of Ophthalmology, University Hospital Bonn, Bonn, Germany.

E-mail address: maximilian.wintergerst@gmail.com (M.W.M. Wintergerst).

<https://doi.org/10.1016/j.media.2026.104313>

Received 4 April 2026; Received in revised form 25 August 2026; Accepted 7 September 2026

Available online 8 September 2026

1361-8415/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Open access funding enabled and organized by Projekt DEAL.

accuracy of 87.98% [95%-CI: ± 2.06], an edit score of 90.80% [95%-CI: ± 1.47], and a segmental F1-score of 89.56% [95%-CI: ± 1.92] on the SICS-155 challenge dataset, as well as an accuracy of 97.61%/ 93.36%, a precision of 98.42%/ 95.27%, and a recall of 97.97%/ 93.32% on the Cataract-101 phacoemulsification public dataset (depending on the data split). Building on these promising results, our future work will investigate potential real-time quality assessment and complication recognition in SICS and facilitate the development of clinical training and surgical quality monitoring software for ophthalmologists in LMICs.

1. Introduction

Cataract is the leading worldwide cause of avoidable blindness. Compared to high-income countries of the Global North, low- and middle-income countries (LMICs) face limitations in resources including infrastructure, funding and availability of medical professionals. This leads to a 10-fold higher rate of avoidable cataract-related blindness in LMICs (Flaxman et al., 2017; Murthy et al., 2008). In addition to increasing access to healthcare, there is a substantial unmet need to improve cataract surgery outcomes, given that surgical outcomes in LMICs are often poorer due to challenges in recognition and management of surgical complications (Bhagat et al., 2007; Flaxman et al., 2017; Goel et al., 2022; Lindfield et al., 2008; Oladigbolu et al., 2014). Nowadays, cataract surgeries are usually video recorded, which allows for self-evaluation, video-based coaching and monitoring of results. Studies have shown that implementing these approaches is directly associated with improved surgical outcomes (Hickman et al., 2022; Thia et al., 2019). However, current solutions require time- and personnel-intensive processes to review videos, so they are usually unavailable in LMICs (Baltussen et al., 2004; Lindfield et al., 2008; Tabin et al., 2008).

One way to address these problems is with artificial intelligence (AI)-based solutions, such as deep learning-based automated annotation of cataract surgical videos. These solutions can efficiently navigate and summarize surgical videos and automatically assess surgical quality. Standardized evaluation systems, such as the ICO OSCAR, could be assessed automatically (Dean et al., 2019; Golnik et al., 2011).

Important steps in the evaluation process include identifying critical phases, tracking the instruments used for each phase, and detecting complications (Garcia Nespole et al., 2022; Mueller et al., 2024; Rampat et al., 2021; Sarker, 2021; Valikodath et al., 2021). Analyzing surgical phases is important because SICS follows a standardized progression from one phase to the next during a typical surgery. This allows for quantitative comparisons between surgeons, feedback on critical steps, detection of discrepancies from established protocols, identification of high-risk phases, and guidance for other tasks, such as instrument segmentation (Sachdeva et al., 2025).

Although algorithms have been developed for the automated analysis of video data to evaluate surgical performance parameters in phacoemulsification procedures, little research has been conducted on manual small incision cataract surgery (SICS), the most widely used, appropriate, and cost-effective cataract surgical technique in countries in the Global South. A collaboration between a Sankara Eye Hospital in Bengaluru, India and the University of Bonn published the first corresponding surgical phase dataset, the first surgical instrument dataset, and a CNN-based algorithm to establish a baseline for phase detection in SICS (Mueller et al., 2025; Sachdeva et al., 2025). Now, we organized the SICS-155 challenge at the 28th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2025) to find the best approach for surgical phase recognition given current advances in DL technology (Mueller, 2025).

In this work, we report the results of the challenge according to the BIAS guidelines proposed by Maier-Hein et al. (2020). We then identify the most promising approaches for feature extraction, model architecture, and post-processing among the submissions, and combine them to obtain state-of-the-art results for phase recognition in cataract surgery, which we evaluate both on our challenge's test set, and on the

Cataract-101 dataset, the largest manually annotated dataset for phacoemulsification.

2. Related work

In recent years, there has been a surge in research regarding automated analysis of cataract surgery, but with a primary focus on phacoemulsification, which has become the preferred technique.

An important milestone was the release of the CATARACTS challenge in 2017 (Liu et al., 2017), with a dataset consisting of 50 videos. The aim of the challenge was the detection of surgical tools. Phase annotations for these videos were subsequently released in 2019, but without an associated challenge. In 2016 and 2018, the Cataract-21 and Cataract-101 datasets were released by the ITEC in Klagenfurt, Austria (Primus et al., 2018; Schoeffmann et al., 2018). The latter collection contains 101 surgeries performed by four surgeons at the clinic in Klagenfurt over a period of 9 months in 2017, with annotations for 10 distinct surgical phases.

In 2021, Ghamsarian et al. published the CatRel dataset, which contains 22 phacoemulsification videos with annotations for relevant phases and anatomical landmarks. This dataset was released in collaboration with the Klagenfurt Group. In 2024, the collaboration unveiled the Cataract-1K Dataset, a substantial, partially annotated collection of phacoemulsification surgical videos comprising 1,000 videos (Ghamsarian et al., 2024). Phase annotations were released for 56 of these videos, along with anatomical and instrument pixel-level annotations for 2,256 frames. Additionally, a small subset of the surgeries has been published that exhibit two major irregularities ("pupil reaction" and "IOL rotation").

The extensive OphNet Dataset was released in 2025. The dataset features 2,278 surgical recordings totaling 284 hours and contains 66 types of cataract, glaucoma, and corneal surgeries from various hospitals in different countries. There are 102 phases and 150 operations for recognition, detection, and prediction tasks, and more than 77% of the videos have high-definition (HD) resolutions of at least 1280×720 pixels. Of these, 743 contain detailed time-boundary annotations, resulting in 7,320 phase-level and 9,795 operation-level segments (Hu et al., 2025).

In addition, in 2025, our collaborative efforts resulted in the release of the previously mentioned SICS-105 phase dataset, which contained a subset of the dataset utilized for the MICCAI challenge (Mueller et al., 2025).

Several private datasets have been reported in the literature, for example, the BigCat dataset consisting of 208 videos (Mahmoud et al., 2024). However, their use remains limited, as they have not been made publicly available and results cannot be reproduced.

Adjacent to our domain of cataract extraction, the following surgical phase recognition challenges have been conducted:

In 2016, the M2CAI challenge presented 16 videos of cholecystectomy procedures that were annotated by phase and surgical tool. These videos were taken from the larger Cholec80 dataset, which was released in the same year. In 2018, the EndoVis challenge introduced multiple datasets focused on laparoscopic surgeries. The SensorOR subchallenge featured 30 videos containing 12 different phases. Over the years, this challenge has evolved to include surgical step and action recognition. The latest iteration of the challenge, held in 2024, presented the Video Dataset for Surgical Phase, Keypoint, and Instrument Recognition in

Laparoscopic Surgery (PhaKIR), focusing on cross-institutional generalizability. The dataset consisted of 13 videos of cholecystectomies with annotations for eight different phases (Rückert et al., 2025).

In summary, despite the existence of some publicly available datasets for phase recognition in cataract surgery, we identified the need for a dedicated challenge to define the state of the art in phase recognition with respect to the difficulties associated with the SICS surgical approach. This approach is the method of choice in the Global South.

3. Material and methods

In this section, we introduce our dataset, describe the methodology of the SICS-155 challenge, and explain how we distilled additional insights from the challenge.

3.1. Medical background of manual small-incision cataract surgery

A cataract occurs when the clear lens of the eye becomes opaque, most commonly due to age-related changes. The lens lies posterior to the iris and anterior to the vitreous body and retina. Cataract leads to decreased visual acuity and in its end-stage to blindness. It is treated through minimally invasive surgery, such as phacoemulsification and SICS (Asbell et al., 2005; Joshua Ehrlich, 2024; Nizami et al., 2022; Rossi et al., 2021; Singh et al., 2017).

Manual small-incision cataract surgery (SICS) is frequently used in low- and middle-income countries due to comparable outcomes to phacoemulsification, its high cost-effectiveness, and low complication rate (Baltussen et al., 2004; Jaggernath et al., 2014; Muralikrishnan et al., 2004). In this surgery, the ophthalmologist constructs a tunnel in the sclerocorneal region, injects an ophthalmic viscoelastic device (OVD) to maintain a stable anterior chamber, performs a capsulorhexis to access the lens, and hydroprocedures to separate lens layers from the

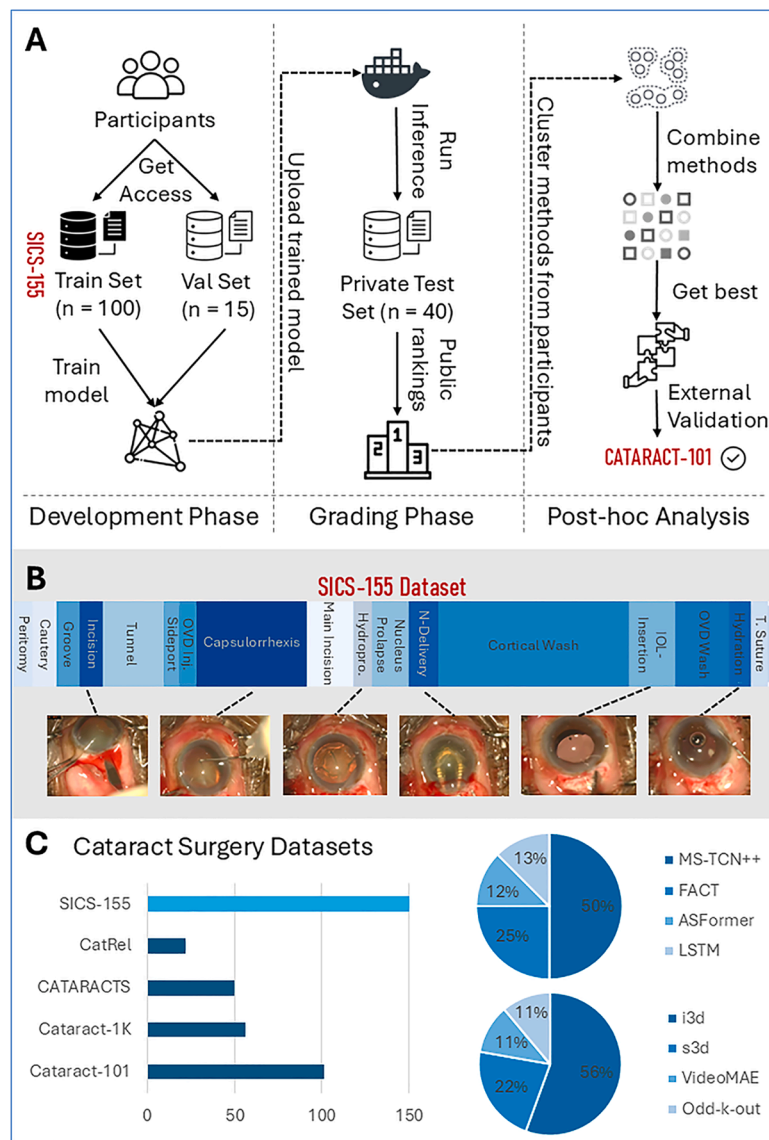


Fig. 1. Overview of our challenge design and our dataset. A shows the three phases of our project. First, the training and validation subsets were released for participants to train their DL models. Post-submission, these models were graded on the private test set by our team. Finally, the methods proposed by our challenge participants were clustered by strategy and systematically evaluated and reassembled into a novel state-of-the-art model. This final architecture was validated on an external dataset (Cataract-101). B shows examples of the phases encountered in a typical SICS procedure with some sample frames. C highlights the fact that our dataset is the biggest, public and fully annotated cataract dataset (left) and presents the architectures (top right) and feature-types (bottom right) chosen by our 12 participants.

capsular bag. The cataractous nucleus is then prolapsed and extracted (Gurnani B and K, 2023; Joshua Ehrlich, 2024; Tabin et al., 2008). The remaining lens cortex is removed by aspiration, and an intraocular lens (IOL) is placed into the capsular bag, and the access tunnels in the conjunctiva are closed. Important steps are shown in Fig. 1B.

SICS differs from phacoemulsification surgery (primarily used in the Global North) in the following ways: SICS involves a larger incision, approximately 6-8 mm (compared to 2-3 mm), which allows for easier maneuvering of tools and extraction of the cataractous lens in its entirety. The surgical phases of SICS differ from those of phacoemulsification. In SICS, steps such as peritomy, nucleus prolapse, and nucleus delivery are performed, while in phacoemulsification, trenching, nucleus emulsification, and irrigation/aspiration are performed using the handpiece instead of a Simcoe cannula. SICS utilizes specialized tools, such as the Vectis, conjunctival scissors, Simcoe cannula, cautery, and crescent blade. These tools are not used in phacoemulsification surgery, which uses tools like the phaco probe, irrigation/aspiration probe, and lens injector (Grammatikopoulou et al., 2021; Spencer, 2006).

3.2. SICS-155 dataset details

Our SICS-155 dataset consists of 155 recordings performed at the Sankara Eye Hospital hospital in Bengaluru, India and annotated by three ophthalmologists over 6 months in 2023 to 2024 with 19 distinct phases. The 155 videos have a resolution of 960×540 pixels (after downsampling from 1920×1080 pixels) and a framerate of 15 fps (downsampled from 30 fps). Surgery durations range from 5:01 to 21:58 minutes, and the average video length is 13:05 minutes. The dataset is a superset of an earlier published database and includes 50 additional videos. Most public datasets used in related phacoemulsification research have a size between 30 and 100 surgery recordings (Cataract-101, IEEE CATARACTS, ...). Some private datasets, like BigCat, have up to 208 videos. We aimed to build a dataset with a similar size for SICS.

Fig. 2 highlights that the class or phase distribution in the SICS-155 dataset is highly uneven, an imbalance that must be addressed by participants of the challenge.

The dataset can be publicly accessed under the following URL: <https://doi.org/10.5281/zenodo.15044589>. The overall study protocol, the data collection and dataset publication were approved by the Institutional Ethics Committee of the Sankara Eye Hospital in Bengaluru, India on the 11th of January 2024. The approval number is SEH/BLR/EC/2024/114. All experiments were performed on anonymized data and in accordance with relevant guidelines and regulations. As detailed below, informed consent was obtained from all participants before surgery and data collection.

Patient selection and recruitment: The inclusion criteria for prospective recruitment at Sankara Eye Hospital in Bengaluru, India were patients with impaired best-corrected visual acuity due to cataracts (as diagnosed by an ophthalmologist), who provided written informed consent, and who were candidates for SICS. The surgery was performed using a digital microscope with a field-of-view recording system. After the included patients underwent surgery, their data was anonymized and randomized into training, validation, and test sets ($n_{\text{train}}=100$, $n_{\text{val}}=15$, $n_{\text{test}}=40$) stratified by video length clustered into short, medium-length and long videos. This was done to guarantee a representative test set for the entire dataset. Patients without a consent form or with insufficient video quality were excluded from analysis.

Annotation procedure: Annotation of the surgery recordings was performed by three senior ophthalmologists from the Sankara Eye Hospital in Bengaluru, India. To ensure annotation quality, 10 random videos from the dataset were annotated by two different annotators and an inter-grader agreement analysis was performed (Ward et al., 2021). After a high agreement was observed between the annotators (Cohen's kappa of 0.96, accuracy of 98.41%), the annotators split the remaining dataset between themselves and completed the annotations. One recording corresponds to one patient and is treated as one case. Most

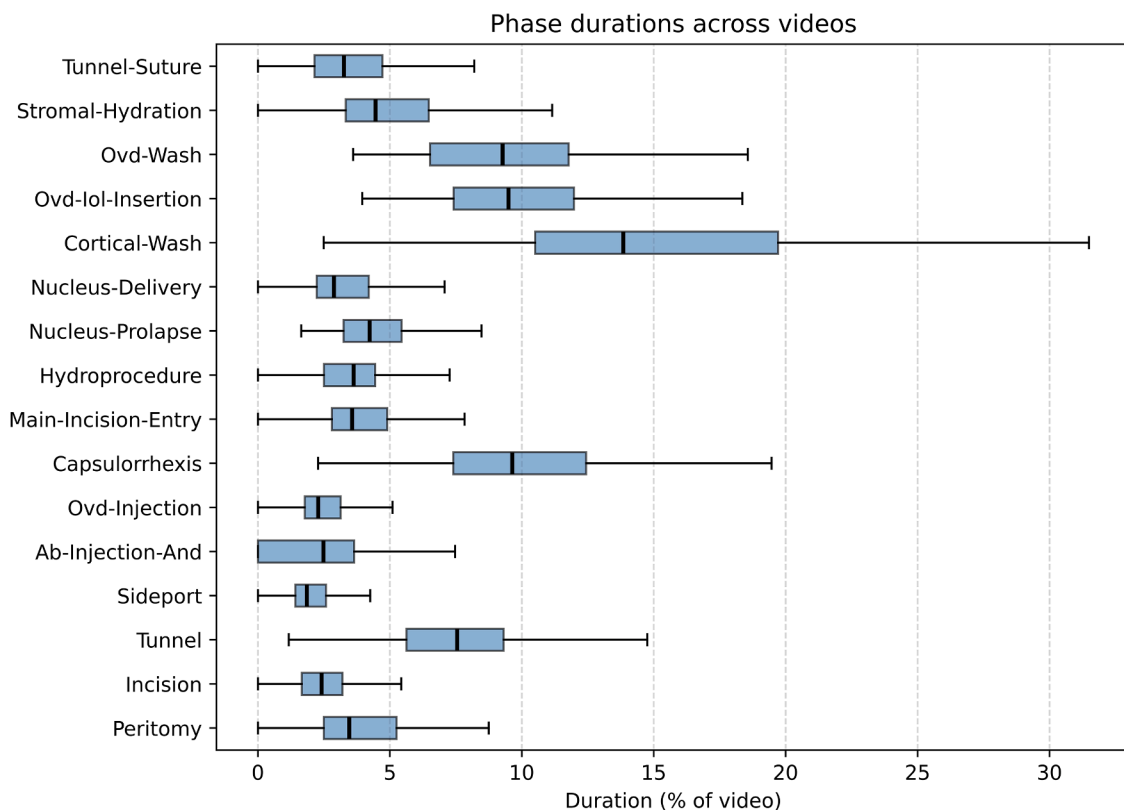


Fig. 2. Boxplot of the durations of different surgical phases in manual small-incision cataract surgery, without outliers. The plot shows the inherent class imbalance in surgical datasets, where some phases e.g. side port only represent a small number of frames and others like cortical wash are overrepresented.

errors were caused by the inherently smooth phase transitions when annotators disagreed about the specific second where a phase starts or ends.

Outlier removal and processing: Videos were manually reviewed; recordings that showed complications requiring extra phases not encoded in the 19 phases of SICS (e.g. vitrectomy) and videos longer than the mean plus two times the standard deviation of the dataset were removed (duration > 21:59 minutes). Videos were scaled down to 960×540 pixels using FFmpeg's Lanczos-resampling for better comparability with the Cataract-101 dataset and to counteract blurriness caused by compression by the recording equipment. Consequently, the present dataset primarily represents routine SICS workflows and does not cover the full spectrum of complicated or irregular cases requiring additional surgical steps.

3.3. Mission of the challenge

The Phase Recognition in Small Incision Cataract Surgery Videos (SICS-155) Challenge aims to advance the evaluation of automated surgery in ophthalmology by providing the first high-quality public dataset on Small-Incision Cataract Surgery (SICS). Although cataract extraction is the most common surgical procedure worldwide, there is a significant research gap: existing datasets primarily originate from HICs and focus on phacoemulsification. The SICS-155 Challenge addresses this disparity by providing data from LMICs, where the cataract burden is highest and SICS is the most vital and cost-effective surgical technique. Due to the high volume of SICS procedures in LMICs, the need for automation is particularly acute.

The task of this challenge is the accurate recognition of phases (temporal segmentation) of a single small-incision cataract surgery. This requires overcoming the inherent complexities of SICS, which typically features longer operative durations and a wider variety of specialized instruments than standard phacoemulsification. Even though SICS is most frequently performed in resource-limited settings, the initial step towards robust video analysis is post-operative, offline processing, which can be done off premises. Therefore, computational constraints were not a primary concern in this challenge.

Another key objective is to leverage the temporal information inherent in the provided video sequences to improve the accuracy and robustness of phase prediction. By providing the first public dataset for SICS, we aim to accelerate the development of AI tools that could support surgeon training, facilitate postoperative quality assessment, and may ultimately improve surgical outcomes in LMICs. Additionally, the availability of our dataset allows testing for generalizability of phase recognition algorithms due to the differences in surgical steps, used instruments and occurrence of complications as compared to phacoemulsification.

3.4. Organization of the SICS-155 challenge

Following acceptance of our challenge as a one-time satellite event of MICCAI 2025, held jointly with the 12th Ophthalmic Medical Image Analysis Workshop on September 27, 2025, we implemented a structured evaluation process and timeline. To facilitate benchmarking, we utilized the open-source CodaBench platform, ensuring all provided source- and evaluation-code remained publicly available on GitHub. No members of the organizers' institutions participated in this challenge. During the entire challenge period, the primary organizer provided support for technical bugs and hardware compatibility in the submission pipeline but did not assist in method development.

The SICS-155 challenge began with a pre-registration phase for interested parties that concluded on April 1, 2025. Participants then entered a development phase lasting until August 10, 2025, during which they were granted access to training ($n_{\text{train}}=100$) and validation ($n_{\text{val}}=15$) subsets, optionally accompanied by pre-computed inflated-3D (i3d) features (Carreira and Zisserman, 2017). Throughout this

phase, training was permitted on the challenge data plus any other publicly available data. In this part of the challenge, participants were able to check their results by uploading their predictions and getting the respective performance metrics on a validation set displayed on a temporary leaderboard. Submission instructions were provided on the challenge webpage and on CodaBench: <https://www.codabench.org/competitions/9961/>.

Between August 10 and August 26, the challenge transitioned to the grading phase. Participants were given this two-week window to upload their DL architectures in a predefined format, aided by a mock prediction model that we provided. These Docker containers were executed on a private virtual machine hosted via a cloud service, where they ingested the 40 videos of the still-private test set for inference. Participants were limited to 5 runs per 24 hours, with inference time restricted to a 6-h duration to enable fair access to everyone. Missing test cases were assigned a chance score. By the deadline on August 26, 12 of the 38 registered teams had submitted their source code. The resulting predictions were used to calculate accuracy, edit-score, and F1-score, with the results subsequently published to the public ranking on CodaBench.

The final ranking was officially announced via email and the challenge website on August 31. Eight of the submissions produced viable results on the test set, and 6 teams submitted short papers delineating their methodology. The teams occupying the top three ranks, amounting to five groups due to shared ranks (method described in section 3.5), were given the opportunity to name a maximum of three co-authors for this joint publication, provided they submitted a description of their method. Additionally, the overall winner received an award of €500. Detailed information regarding the entire organization of the challenge can be found on the challenge website (Mueller, 2025).

3.5. Assessment method

Overall Ranking: To establish the challenge winner, we calculated a point-based significance ranking, using the statistical approach and recommended best practices outlined in the paper by Maier-Hein et al. (Maier-Hein et al., 2020). The final ranking was based on statistically significant outperformance of competing submissions. To avoid reliance on a single metric, rankings combined per-frame accuracy, quantifying overall correctness, F1-score to account for class imbalance; and edit score, which captures temporal coherence and the correct sequence of phases. F1-scores used the macro average configuration. All scores were first calculated for each test sample and then averaged across the entire test set.

The pseudocode for this approach is described below:

All metrics used in the challenge are calculated for each case (video) and averaged across all cases in the test set. We adopt a point-based ranking method supported by statistical analysis for robustness and to filter out small differences in scores between participants. We choose the significance level $\alpha = 5\%$. For each metric m_k , $k = 1, \dots, o$:

1. Determine performance m_k of each algorithm for each test case
2. Perform pairwise comparisons between the results for the metric m_k from two participants with the values using Wilcoxon signed rank test (with α)
3. Determine a significance score which equals the number of results (from other teams) performing significantly worse than the current algorithms results according to the test
4. Compute the overall ranking (shared ranks possible) for the metric m_k based on the scores. The highest score corresponds to the best result(s) from the submitted algorithm(s) (rank 1)

The final ranking over all metrics is computed by aggregating the significance scores by calculating their mean or median.

The ranking and the winner of the challenge were announced on the 31st of August 2025 via mail and on the official webpage of the SICS-155 Challenge (Mueller, 2025).

3.6. Process to distil insights from contributions

The results from the top four challenge participants showed improvements compared to our baseline algorithm by introducing new features, neural network architectures, and post-processing strategies, but did not systematically sample the resulting design space (Mueller et al., 2025). Therefore, after concluding the challenge, we investigated whether clear best practices could be distilled from all approaches and whether combining them would further improve results. We also evaluated whether the resulting architecture was specific to SICS or if it also provided state-of-the-art results for phase recognition in phacoemulsification surgery.

During the post-challenge phase, we evaluated various combinations of the teams' methods on the validation set. Then, we ran the best model from this comparison on the SICS-155 test set and compared it with the individual contributions. To determine the generalizability of the optimal architecture, we trained and evaluated it without additional hyperparameter optimization on the Cataract-101 dataset.

The results and comparison to prior results are described in the Results section below.

3.7. Software tools and statistical analysis

The Python programming language (version 3.11), the SciPy statistics package and the Matplotlib library were used to randomize acquired videos, facilitate training and inference of the DL networks, evaluate results and generate plots. DINOv3, CLIP and i3d features were extracted using the video_features library from Vladimir Iashin (v-iasin) available on GitHub (Iashin, 2020). This library uses PyTorch Image Models (timm) to extract embeddings from the underlying backbone. For this patch tokens are calculated by the frozen transformer which are then pooled to produce a 1-dimensional vector per video frame containing dense spatial information. For DINOv3, we considered two feature extraction choices, mean-pooled patch tokens and the CLS token. Linear probing with both performed comparably (around 40% frame accuracy). We decided to use the pooled patch features, in the hope that they capture more details of individual anatomical structures and surgical tools.

To rank the challenge participants, we used frame-wise accuracy, F1-score and segmental edit score to keep the protocol simple as required by the challenge review. The edit-score is defined using a normalized edit distance punishing insertions, deletions and replacements as described by Lea et al. (2016). For our post-challenge experiments, we used a more comprehensive set of metrics. We utilized frame-level metrics to get a general idea of the performance of the algorithm and phase-level evaluation metrics to get more robustness against class imbalances. For the first, we use accuracy, precision and recall and for the second group metrics consisted of phase-accuracy, the edit score and F1-scores for an overlap between prediction and ground truth of more than 10%, 25% and 50%. For simplicity, we calculate the average F1 score (Zhang et al., 2024, Lea et al., 2017; Lea et al., 2016):

$$F1_{avg} = \frac{1}{3} \times (F1 @ 0.10 + F1 @ 0.25 + F1 @ 0.5)$$

For the uncertainty analysis of our models, we retrain the model with 7 random seeds and report each metric as its mean across seeds with a 95% confidence interval. Given the small number of seeds, the interval is derived from the standard error using Student's t distribution rather than the normal approximation (Raschka, 2024; Steinbach et al., 2022). The resulting confidence intervals quantify variability due to repeated training, for a fixed choice of videos for evaluation.

4. Results

In the following, we present results for the SICS-155 challenge, analyze the methods that have been used by participants, and provide a

more detailed description of the most relevant ones.

4.1. Results of the SICS-155 challenge

Our challenge concluded on 30 August, and eight teams managed to submit running algorithms to our platform. The performance metrics accuracy, edit- and F1-score were calculated using the outlined point-based significance ranking approach (significance level $\alpha=5\%$) using their predictions and the private ground truth of the test set ($n=40$ videos). The resulting rankings are displayed in Table 1. Details have been removed and team anonymized when no method description was supplied in accordance with the BIAS guidelines.

The MIPLab team from the University of Calgary in Canada achieved the top results, with an accuracy of 85.88% and an edit-score of 89.00% that were significantly better than all other teams. Second place was shared by TeamBIT from the Beijing Institute of Technology in China; and the team AtlasVision from the University Mohammed VI Polytechnic Benguerir in Morocco.

Third place was awarded to the teams from the Fraunhofer HHI Institute in Berlin, Germany and Team HD from the Hang Zhou Dianzi University in China. The HHI team was only able to run inference at a temporal resolution of 5 FPS, so their score might under-represent their model's power. The remaining submissions could not generalize well to the test set.

4.2. Clustering of methods used by challenge participants

We reviewed all viable submissions to identify similarities and differences. All submissions focused on offline phase recognition, no online methods that could be used during a surgery have been submitted. Overall, participants exploited three main areas of improvement to boost phase recognition accuracy in the video processing pipeline, as summarized in Fig. 1C:

1. Teams experimented with different types of input video features. These features were usually extracted from the input video by fixed backbone networks and then used for training and inference by the main network.
2. Additionally, our participants improved the prediction architecture. All submissions, like our baseline, involved some kind of temporal modeling. While some teams built on the MS-TCN++ architecture that was used in our baseline, other teams explored LSTMs or transformer-based architectures.
3. Finally, some teams post-processed the predictions from the main network to improve temporal coherence. Here, confidence filtering and ensembles were considered.

This clustering allowed us to systematically analyze the results and extract further insights by combining the best-performing elements from all three areas in Section 5. In the next subsection, we will highlight some effective and innovative approaches to the intrinsic challenges of our dataset.

4.3. Highlighting specific approaches in the challenge

In this section, we present an overview of the methods used by participants and described in the five short papers whose authors are co-authors of this joint manuscript. The contribution of the second team sharing rank 2 will not be discussed because no short paper was submitted.

Confidence Filtering and VideoMAE (rank 1): The winning team (Nielsen et al.) used the MS-TCN++ implementation from our baseline implementation but replaced the I3D features with the more powerful VideoMAE features (Tong et al., 2022) and increased its capacity via hyperparameter optimization. Extraction speed of these features has been optimized by temporally downsampling the input video and then

Table 1

Results of the participants of the SICS-155 challenge in the final grading round.

Participant	Initials*	Rank**	Accuracy	Edit-score	Frame- F1-score	Method Description
MIPLab, CA	CN, PG, AG, NDF	1	85.88	89.00	81.05	MS-TCN++, Confidence Filter, VideoMAE features
Anonymized; details not reported	-	2	82.37	86.21	75.65	-
Team BIT, CN	YK, HJ, JD, HL	2	82.96	83.35	77.53	Ensemble MS-TCN++ / C2F-TCN, I3D/SD features
AtlasVision, MA	YEJ, YI, LM	2	81.88	84.74	75.76	Boundary-aware FACT, I3D features
Fraunhofer HHI, DE	LRK, ELW	3	74.72	41.41	66.06	ASFormer, Custom Odd-k-out features
Team HD, CN	DZ, JL, MH	3	70.64	48.70	61.93	Bidirectional LSTM + temp. ConvNet, I3D features
Anonymized; details not reported	-	4	48.24	7.74	36.93	-
Anonymized; details not reported	-	5	10.36	68.95	8.33	-

The participants' code is executed on a remote virtual machine on the private test set ($n = 40$ videos) and the results published on CodaBench. Details were removed from teams that did not supply a method description. *This column reports the initials from the participants of the team in correspondence to our co-author list.

**Ranking of the teams is established by the significance rating algorithm described in the methods section.

interpolating the calculated features for the target framerate of 15. A key technical innovation in their approach improved temporal consistency via confidence filtering; identifying frames that were not classified with sufficient confidence by the network and reassigning them to the phase of the temporally closest confident frame. This approach successfully suppresses brief, low-confidence label flips. As shown below, it improves results for both MS-TCN++ and the ASFormer architecture.

Because frame-level class probabilities were not well calibrated initially, this team calibrated confidence scores via temperature scaling. First, a single temperature parameter was optimized by minimizing the negative log-likelihood on the logits in the training set, following temperature scaling for neural calibration. Second, a per-class confidence threshold is calculated as the 25th percentile of the true-class softmax confidences, measured on probabilities in the training set. At inference, the predicted logits are divided by the temperature parameter before computing the predicted class and confidence at each frame, which serve as the basis for the above-described confidence filtering approach.

Feature and Model Ensembles (shared rank 2): Du et al. explored ensemble strategies for video features and temporal modeling to boost overall performance. They evaluated multiple feature types and finally decided to combine I3D and S3D features to capture different aspects of the data. They found that the combination with equal weights of two differently trained MS-TCN++ networks and one Coarse-to-Fine Temporal Convolutional Network (C2F-TCN) on the logit level yielded the best results. Accuracy improved by 0.79 percentage points and the edit-score by 4.47 percentage points compared to the baseline architecture.

Boundary-aware FACT (shared rank 2): Yasser et al. concentrated on identifying an architecture for better temporal modeling and adapting it to the problem posed by the SICS-155 challenge. Their approach was based on the Frame-Action Cross-Attention Temporal Model (FACT) published in 2024 by Lu et al. which combines a frame branch based on convolutions and an action branch with learned action tokens (Lu and Elhamifar, 2024). The team extended that framework by modifying the temporal smoothness loss to facilitate label changes at true phase transitions. To this end, they added a boundary head to FACT's frame branch, consisting of a parallel 1-dimensional convolutional layer followed by a sigmoid function, yielding a per-frame phase boundary probability $p_b(t) \in [0, 1]$ using the same features as the classifier. During training, the boundary head was supervised through a cross-entropy loss based on the ground truth $y_b(t) = 1$ ($y_t \neq y_{t-1}$). The boundary head was not used at inference. Its output only served as a boundary weight in a total variation loss that regularized frame-wise phase predictions p_t ,

$$L_{TV} = \frac{1}{T-1} \sum_{t=1}^{T-1} (1 - p_b(t))^\gamma \|\log p_{t+1} - \log p_t\|_2^2$$

This loss replaces FACT's uniform smoothing. The exponent $\gamma > 0$ controls how much even uncertain boundaries $p_b(t) < 1$ should relax the smoothness penalty. It is set to $\gamma=2$ via hyperparameter optimization. A schematic of their modified architecture can be seen in Fig. 3.

Odd-K-Out feature extraction (shared rank 3): Most participants utilized out-of-the-box pre-trained features like i3d or VideoMAE with good predictive performance. Renz-Kiefel et al. chose to instead learn their own representation, with the goal of maximizing its ability to discriminate between different phases. They formalized this as an odd-one-out prediction task, in which the network is presented with three frames, two belonging to the same surgical phase, the remaining one to a randomly selected different phase. A ResNet50 is then trained to recognize the odd-one out (Muttenthaler et al., 2023). During training, the images in one set are each passed through the network, and the resulting logits are averaged over the set. The purpose of this is to overcome the class imbalances that are posed by our dataset and to extract better features that can be used for the final step of the phase predictions using a network that models the temporal component of the surgical sequences. Even though the features learned with this approach performed less well than the VideoMAE features (accuracy of 73.92% versus 82.61%) in our experiments, combining them with the ASFormer architecture for the final phase predictions worked sufficiently well to achieve rank 3 (Yi et al., 2021).

Bi-LSTM and convolutions (shared rank 3): Zheng et al. used standard I3D features for the training of a bidirectional LSTM and a temporal ConvNet to combine long-range dependencies and local similarities. Fusion of the two networks occurs before a final fully connected layer and the cross-entropy loss. A latent space of 256 dimensions worked well for the LSTM and was determined together with the other hyperparameters via an exhaustive grid search. This approach worked well on the development set but suffered from over-segmentation, as can be seen from the discrepancy between accuracy (70.64%) and edit-score (48.70%).

5. A new state-of-the-art model based on the SICS-155 challenge results

Analyzing the results of our challenge, we identified three main areas of improvement: input features, network architectures, and post-processing strategies, as discussed in Section 4.2. However, the submissions represented a limited number of combinations of these approaches, which prevented us from assessing the individual relevance of each aspect. Therefore, this section reports additional experiments in which we systematically explored combinations of the most successful strategies after the challenge concluded.

5.1. Quantitative evaluation of methods

Three teams experimented with improving pre-computed input features (see Table 2) and achieved promising results with inflated-3D (i3d), VideoMAE features and a custom Odd-K-Out feature extractor. For completeness, we add experiments with the recently released DINOv3 features and CLIP features that achieved good results in other computer vision domains. All extracted features were used to train an MS-TCN++ model on the train subset, and this model was used to run

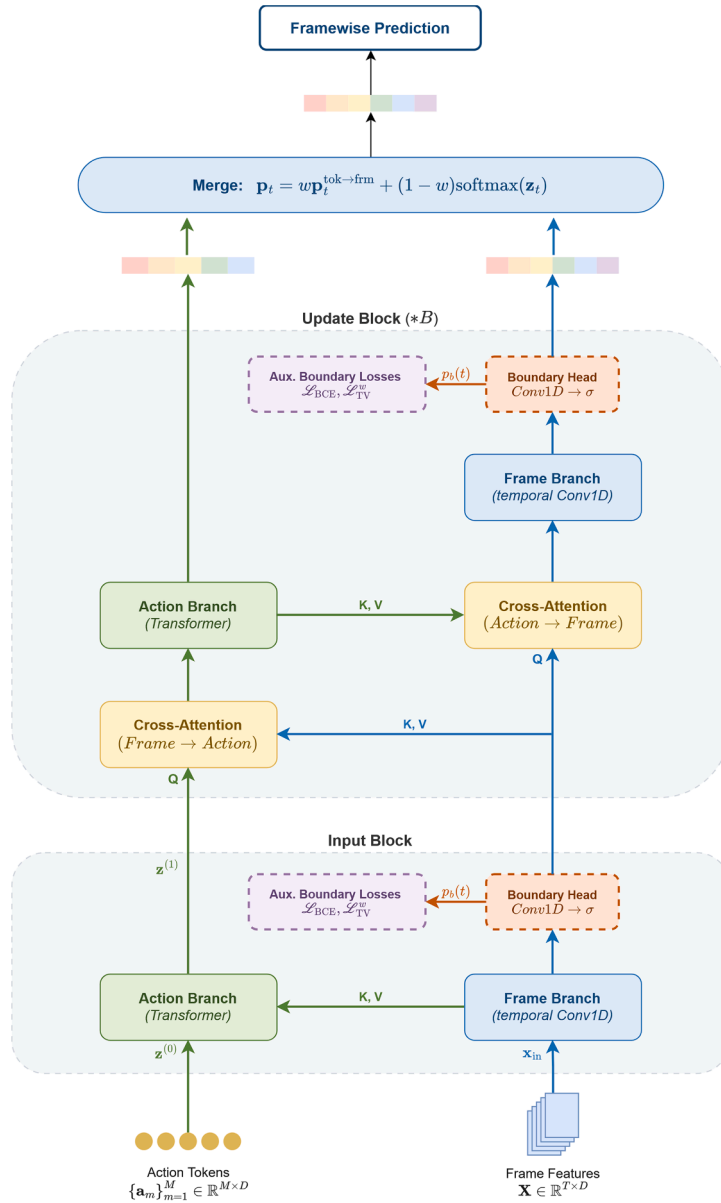


Fig. 3. Overview adapted from Yasser et al. showcasing the boundary-aware FACT architecture (Lu and Elhamifar, 2024). The addition of a CNN-based boundary head to the frame branch described in the original research paper that is only active in the training phase boosts predictive performance on the SICS-155 dataset and contributed to the second rank in our challenge.

Table 2

Evaluation of the effect of changing the pre-processed feature extraction on prediction quality.

Metric	i3d	CLIP	VideoMAE	DINOv3-B	DINOv3-L	Odd-K-Out
Accuracy	76.59 [±2.38]	68.82 [±2.42]	82.61 [±1.56]	77.89 [±0.93]	80.42 [±1.73]	73.92 [±1.23]
Edit-score	68.77 [±4.87]	71.36 [±4.23]	75.12 [±1.80]	75.88 [±3.04]	77.84 [±5.07]	43.73 [±1.74]
F1 @ 0.10	69.52 [±6.06]	67.64 [±3.34]	79.54 [±2.17]	78.82 [±2.05]	80.49 [±4.47]	54.29 [±1.91]
F1 @ 0.25	65.47 [±6.73]	63.32 [±2.91]	76.92 [±2.24]	75.07 [±2.39]	76.83 [±4.43]	49.50 [±1.93]
F1 @ 0.50	53.61 [±6.34]	49.43 [±3.35]	67.13 [±2.99]	61.10 [±2.61]	65.29 [±4.39]	38.89 [±2.12]
F1-avg	62.86 [±6.26]	60.13 [±3.13]	74.53 [±2.35]	71.67 [±2.47]	74.20 [±4.38]	47.56 [±1.94]
T _{compute} / 100 vids	23:10	9:10	5:48	12:23	15:40	6:07

After the extraction of the given feature type, they are used for training an MS-TCN++ model and calculating the performance metrics on the validation data. 95% confidence intervals (derived from seven training runs with different random seed initializations) are shown in square brackets. Compute time is given in the format HH:MM. Here the segmental F1-score with different thresholds 10%, 25% and 50% is reported.

predictions on the validation subset.

These experiments showed that VideoMAE and DINOv3-Large perform best in the context of our challenge with an accuracy of

82.61% versus 80.42%, an edit-score of 75.12% versus 77.84%, comparable F1-scores and overlapping 95% confidence intervals for the measured metrics. Comparing VideoMAE features to the i3d feature

baseline, segmental scores improve significantly with +6.02 points in accuracy and +6.35 points in the edit score. Processing 100 videos was significantly faster (5:48 versus 15:40 hours), using the VideoMAE feature extraction code from the winning team, due to the down-sampling and interpolation described in Section 4.3. DINOv3-Base was slightly faster, but also less accurate than the larger model variant, and still took more than twice as long as VideoMAE. The custom-trained Odd-K-Out feature processor performed worse with the MS-TCN++ architecture than frozen out-of-the-box feature extractors like DINOv3 or i3d.

In the next step, we evaluated different DL architectures that were utilized by our participants. Three teams used the MS-TCN++ model, one team used ASFormer and two teams variations of the FACT model. To make these approaches comparable, we trained all of them on i3d features. The resulting Table 3 shows that boundary-aware FACT as proposed by Yasser et al. (team AtlasVision) outperformed both MS-TCN++ and ASFormer, especially when looking at the edit-score (84.23% vs 68.77% and 80.47%). The performance of MS-TCN++ could be boosted to narrow the gap with FACT when using the Confidence Filtering approach proposed by Nielsen et al. (team MIPLab) which especially boosts edit-score and F1-score.

We combined confidence filtering also with the other architectures but found that it improved results only for MS-TCN++ and ASFormer (improvements of 12.47 and 3.52 percentage points in edit-score) but did not work well with the baseline or boundary-aware FACT model, due to less oversegmentation errors (further elaboration in the discussion).

Similarly, we found that the ranking of feature types depended on the model architecture. Therefore, we ran an exhaustive search on all combinations of input features, model architecture, and post-processing steps, including and excluding confidence filtering. The most important results of this search are reported below.

In a first step, to validate the benefit of the boundary-aware modification proposed by Yasser et al., we trained an unmodified version of FACT with i3D features as a baseline (see Table 3). On these features the modification had a mixed effect: frame-wise accuracy and edit-score increased (accuracy of 81.24% vs. 80.28% and edit score 84.23% vs. 84.14%), while segmental F1-scores decreased (F1-avg of 80.71 vs.

Table 3
Evaluation of various DL architectures on the prediction quality.

Metric	MS-TCN++ (Table 2)	ASFormer	FACT	Ba-FACT*	MS-TCN++ w/ Confidence Filter
Accuracy	76.59 [±2.38]	78.66 [±1.88]	80.28 [±0.81]	81.24 [±0.69]	74.63 [±5.64]
Edit-score	68.77 [±4.87]	80.47 [±2.63]	84.14 [±1.80]	84.23 [±1.98]	81.24 [±5.62]
F1 @ 0.10	69.52 [±6.06]	82.35 [±3.24]	86.40 [±1.55]	85.97 [±1.74]	83.80 [±3.28]
F1 @ 0.25	65.47 [±6.73]	78.72 [±3.36]	83.41 [±1.65]	82.75 [±1.57]	79.37 [±4.24]
F1 @ 0.50	53.61 [±6.34]	66.75 [±4.23]	74.52 [±2.35]	73.41 [±1.29]	66.09 [±6.34]
F1-avg	62.86 [±6.26]	75.94 [±3.57]	81.44 [±1.78]	80.71 [±1.45]	76.42 [±4.56]
T _{training} (hh:mm)	03:02	18:31	06:22	06:22	03:02
T _{inference} (in secs)	2.64	6.31	1.08	1.08	2.67

All models were trained on inflated 3D features. The last column shows the effect of the post-processing technique confidence filtering on the MS-TCN++ architecture. The best overall results are in bold font. 95% confidence intervals (derived from seven training runs with different random seed initializations) are shown in square brackets. *A modified version of the FACT architecture that is phase boundary aware has been used as proposed by the team AtlasVision in the previous subsection. Compute time for training is given in the format HH:MM and for inference in seconds. Here the segmental F1-score with different thresholds 10%, 25% and 50% is reported.

81.44%). We then re-evaluated the feature and architecture combinations that had performed best in our earlier analysis, assessing both the boundary-aware modification (plain vs. boundary-aware FACT with VideoMAE features) and the input feature type (VideoMAE vs. DINOv3-L with boundary-aware FACT) on the validation set. These results are reported in Table 4.

With VideoMAE features, the boundary-aware modification improved all six metrics over the plain FACT baseline, unlike the mixed effect seen with i3D features. The largest gains were in the segmental edit-score (87.59% vs. 85.22%) and F1 @ 0.10 score (89.59% vs. 87.64%), whereas the accuracy score improved only marginally (+0.14) and the F1-average by +1.6. All confidence intervals overlap, so these differences should be read as consistent trends.

Substituting VideoMAE with DINOv3-L features left validation performance essentially unchanged, but VideoMAE retained a small edge on the edit score (87.59% vs. 87.41%) and the F1-average score (85.71% vs. 85.59%). DINOv3 performs slightly better on the accuracy score: 85.52% vs. 85.72%. All differences fall well within the reported confidence intervals.

Since we saw no meaningful difference in the scores, and computing VideoMAE features took roughly a third of the time compared to DINOv3-L, we performed the remaining experiments using VideoMAE features. In particular, we trained the boundary-aware FACT with VideoMAE features on the combined training and validation subsets and ran inference on the challenge test set of 40 videos. Again, we estimated the uncertainty through 95% confidence intervals obtained by reseeding during training. This configuration outperformed all challenge submissions, with an accuracy of 87.98% [95%-confidence interval: ±2.06], a mean edit score of 90.80% [±1.47], and a mean F1-average of 89.56% [±1.92].

5.2. Validation on the Cataract-101 dataset

Finally, we examined whether the optimal model from our post-challenge experiments could be generalized to the Cataract-101 dataset. We treated this dataset as an external, completely independent test case, as it used a different variant of cataract surgery (phacoemulsification) and a different set of phases. Due to the similar performance between VideoMAE and DINOv3-L features in our initial evaluation (see Tables 2 and 4), we trained the boundary-aware FACT both with VideoMAE and with DINOv3-L features on the Cataract-101 dataset and compared results with those reported by other authors in the last five years, as shown in Table 5. We selected the subset of performance metrics by reviewing the given articles and choosing the metrics reported in almost all of them.

As shown in Table 5, our approach demonstrates strong cross-dataset generalization without tuning on the external dataset and seems to improve on the previously reported results (14.85% of videos in the test split) with an accuracy of 97.61%, a precision of 98.42% and a recall of 97.97%. We again observe small performance increases when using

Table 4
Validation set ablation study of our final model.

Metric	FACT /w VideoMAE	Ba-FACT /w VideoMAE	Ba-FACT /w DINOv3-L
Accuracy	85.38 [±0.72]	85.52 [±1.13]	85.72 [±0.99]
Edit-score	85.22 [±1.63]	87.59 [±1.67]	87.41 [±1.23]
F1 @ 0.10	87.64 [±1.57]	89.59 [±1.20]	89.98 [±0.58]
F1 @ 0.25	85.73 [±1.44]	87.22 [±1.36]	87.57 [±0.77]
F1 @ 0.50	78.96 [±1.63]	80.31 [±1.79]	79.21 [±1.19]
F1-avg	84.11 [±1.46]	85.71 [±1.37]	85.59 [±0.73]

Evaluated components include previously optimal features (VideoMAE, DINOv3-L) and FACT architectures (boundary-aware, plain). The best overall results are in bold font. 95% confidence intervals (derived from seven training runs with different random seed initializations) are shown in square brackets. Here the segmental F1-score with different thresholds 10%, 25% and 50% is reported.

Table 5

Comparing the results of our assembly of best practices from the SICS-155 challenge on the Cataract-101 dataset with phase recognition approaches by other authors in the last 5 years.

Author	%vids in test-set	Accuracy*	Precision	Recall	Method Description
AutoML (Touma et al., 2022)	11.25	96.00	81.60	77.10	CNNs, features not reported
MACBNet (Pan et al., 2026)	39.60	94.31	92.97	93.51	ConvNeXt features, Adaptive Temporal Convolution Network
GlobalLocal MS-TCN (Fang et al., 2022)	27.72	96.50	94.90	95.20	TS-Net to extract features, MS-TCN for prediction
GLSFormer (Shah et al., 2023)	39.60	92.91	90.04	89.45	ResNet features, gated temporal, shared spatial transformer encoder
SlowFast TMN (Zhang et al., 2024)	49.50	93.82	91.72	91.46	ResNet50 features, Slow-fast AS-Former
MS-TCN++, 2025 (Mueller et al., 2025)	14.85	91.30	97.80	91.20	I3D features, MS-TCN++ CNN-based model
BA-FACT with VideoMAE (ours)	14.85	97.61	98.42	97.97	
	49.50	93.36	95.27	93.32	VideoMAE features, boundary-aware FACT transformer
BA-FACT with DINOv3-L (ours)	49.50	92.12	94.60	92.10	DINOv3-L features, boundary-aware FACT transformer

The results reported for MS-TCN++ in 2025 were done by our group as an initial approach to establish a baseline for SICS phase recognition. %vids in test-set = Percentage of videos in the test set as reported by the author. *Frame-based accuracy-score.

VideoMAE features compared to DINOv3-L. Additionally, we calculated our previously established edit score (91.72%) and $F1_{avg}$ -score (92.70%).

For a fairer comparison with test set splits utilized by other authors, we also report the results of our model with the least amount of training data among the reported splits (51 training videos, 50 test videos). This made accuracy drop to 93.36% (precision 95.27% and recall 93.32%) which remains competitive with results of specialized architectures.

5.3. Implementation details of final architecture

The previously described boundary-aware FACT architecture was trained with the following configuration: Video features were extracted using the custom extraction code supplied by Nielsen et al. which internally utilizes the VideoMAE v2 implementation from Hugging Face, yielding a feature dimension of 1024. These features were used to train the modified FACT model of Yasser et al., which augments the input block with a boundary head and an additional boundary loss (see Section 4.3 for details).

The model was trained for 150 epochs using the Adam optimizer with a learning rate of $1e-4$, decayed once at epoch 80. The batch size

was set to one video, with all frames sampled from the input features. The architecture comprises one input block with 6 attention layers and three update blocks with 1 attention layer each; every block contains 10 convolutional layers in the frame branch. The action branch is initialized with 37 action tokens. A conditional masking ratio of 0.3 was applied. The total loss is a uniform combination of four terms: Frame Loss, Cross-Attention Loss, Action-Token Loss, and Temporal Smoothing Loss (replaced by the new boundary loss). Token-to-segment matching follows a one-to-one assignment strategy.

6. Discussion

Compared to our baseline on the smaller, manual small-incision cataract surgery dataset (SICS-105), the results from the SICS-155 challenge show significant improvements at every step of the surgical phase detection pipeline. Analysis of the validation set results shows that the accuracy and edit score improved by 8.93 and 18.82 percentage points, respectively, when comparing the original MS-TCN++ architecture to the best model identified during post-challenge experiments (boundary-aware FACT with VideoMAE features).

This result is especially significant for countries in the Global South, where SICS is the preferred surgical technique. Here, the automation of surgical analysis can help counteract limitations in training resources and improve quality control.

The final model narrows the gap to human expert agreement but does not close it. Averaged over seven training runs with different random seeds, the top-performing model achieved an accuracy of 87.98% [95%-confidence interval ± 2.06], an edit-score of 90.80% [± 1.47], and a segmental F1-score of 89.56% [± 1.92]. For reference, the inter-rater agreement of two ophthalmologists was 98.41% accuracy, 97.64% edit-score, and 96.54% F1-score, leaving a residual gap. The model therefore reproduces the expert phase sequence with high fidelity while still falling measurably short of human performance, which we regard as the practical ceiling for this task.

We also demonstrate that these strong results translate to phase recognition in phacoemulsification surgery and that our final boundary-aware FACT model shows strong cross-dataset generalization without additional tuning on an external dataset. The architecture is competitive and even surpasses some of the results reported by other authors in the literature. However, we note that a direct comparison of results is limited by the fact that no unique split into training and test set has been established for this dataset.

Results in Table 2 highlight that VideoMAE and DINOv3 features performed significantly better than other older features. We theorize that VideoMAE is well suited to our surgical videos because it encodes spatial and temporal dependencies by using a novel masking strategy. Tong et al. found that most videos have a very high temporal correlation. Simple frame masking is insufficient for training an autoencoder because there is enough information in the predecessor and successor frames to enable restoration without learning of semantic relations. Their team proposed tube masking, which removes an area in a frame and propagates the mask along the temporal axis. This method has an extreme masking ratio of 90%. This forces the self-supervised transformer network to reason over high-level semantics to recover these totally missing cubes during training (Tong et al., 2022). The resulting backbone is ideal for encoding the properties of a SICS surgery because large parts of the scene remain unchanged throughout the procedure. The surgical environment, ocular anatomy, and supporting instruments remain the same throughout the recording process.

In contrast, DINOv3 operates on a per-frame basis. In this approach, a student model must produce predictions that closely resemble the outputs of a teacher model. During self-supervised training, the teacher model uses a global view of an image, while the student model only has access to cropped versions of the same image. This iterative process, applied to a large set of natural images, teaches the student model to relate local image patches to the global context of the teacher model

(Siméoni et al., 2025). DINOv3 encodes spatial information in such detail that it compensates for the missing temporal information in our SICS use case. It is interesting to observe that both frame features (DINOv3) and video features with a temporal component (VideoMAE) can achieve good results on our dataset. This indicates that our various temporal network architectures can adapt well to missing, or additional, temporal information encoded in the input features. Overall, VideoMAE and DINOv3-L features produced very similar metrics in many of our experiments, while VideoMAE has substantially lower computational effort. We therefore recommend VideoMAE features for the final configuration, while noting that our experiments indicate that features and temporal architecture must be selected jointly rather than independently.

Further, our experiments showed that specialized transformer networks such as FACT outperform CNN-based approaches such as MS-TCN++ by a wide margin. This follows general trends in the computer vision community seen in recent years (Rückert et al., 2025).

However, this difference becomes less clear when examining the older ASFormer architecture, which cannot clearly outperform the temporal convolutional network. The underwhelming results in the SICS-155 dataset stand in contrast to improvements on other video datasets reported in the literature. We attribute a small part of this to configuration: adopting the hyperparameters of Ahmadi (2026), two decoders and 11 layers per block, yielded results more competitive with MS-TCN++ (ASFormer 78.66% accuracy / 80.47% edit-score; MS-TCN++ 76.59% / 68.77%), whereas replicating Zhang (2024) gave no meaningful gain (Ahmadi et al., 2026; Zhang et al., 2024). This does not fully account for the gap. As a sanity check, ASFormer trained on Cataract-101 performed substantially better (90.67% accuracy, 88.37% edit-score), suggesting that the higher complexity of SICS-155 (19 phases versus 10, and longer videos) is more difficult for this architecture. We also note that MS-TCN++ trains faster (around 3 versus 20 hours) and has lower inference cost.

We believe that the favorable performance of FACT is caused by multiple properties. Action tokens make it possible to encode long temporal dependencies better and with more stability, which is especially important in surgical videos where strong standardization leads to a consistent order of phases, e.g., wound suturing always happens at the end of a surgery. Lu et al. proved this property in their paper by showing that performance is more stable on longer videos compared to other architectures (Lu and Elhamifar, 2024). Additionally, the combination of a frame branch and the action branch brings together the best properties of CNN networks with strong local connectivity and transformer networks that enable the learning of complex relationships between tokens and features.

The confidence filtering approach was an important component of the winning submission but proved ineffective in combination with FACT: For MS-TCN++, the filter affected 4.16% of frames across 72.87% of phases, removing 47 phase segments in total on the validation set. For Ba-FACT, the corresponding figures were 0.58% of frames across 13.90% of phases, with only 2 segments removed. Consequently, filtering had almost no effect on validation metrics. This corresponds to the observation that the oversegmentation that the confidence filter was designed to suppress arises much less frequently in boundary-aware FACT (2.46 oversegmented phases per video compared to 4.4 with MS-TCN++).

Finally, our ablation isolates the contribution of the boundary-aware modification. With i3d features, boundary-aware FACT improved the edit-score by 0.09 points (84.14 to 84.23%) over the original FACT baseline. Also, frame-wise accuracy only showed minimal improvements (+0.96, 80.28 vs. 81.24) with clearly overlapping confidence intervals. With VideoMAE features the performance increase through the boundary-aware modification become more visible with an edit-score of 85.22% vs. 87.59%. Given the very low computational cost of this component, we include it in our final pipeline despite these somewhat limited gains.

Our extensive experiments revealed that input features, model architecture, and post-processing cannot be analyzed as independent components of a pipeline. This is clearly visible when observing the overlapping confidence intervals for most comparisons in Tables 2, Tables 3 and 4. The findings discussed should be interpreted as consistent trends rather than statistically established differences.

As expected, features that encode temporal context can compensate for shortcomings in temporal modeling. The winning entry by the MIPLab team demonstrated that relatively basic architectures, such as MS-TCN++, can achieve competitive accuracy when combined with modern features, such as VideoMAE, and effective post-processing. In general, selecting the correct features seems to impact performance as much as selecting the deep learning architecture (see Tables 2 and 3). We expect these insights to generalize to other surgical datasets beyond ophthalmology (e.g., Cholec80).

Looking ahead, the deployment of automated video analysis in clinical practice requires sufficient prediction quality and computational efficiency. In this regard, CNN-based approaches have an advantage over more powerful transformer-based architectures. However, the main computational bottleneck in our current pipeline is the feature extraction step, which accounts for more than 95% of computing time. But even ignoring the time needed for feature extraction, the 1 to 2 seconds necessary for model inference of our fastest architectures (see Table 3) must be optimized for online video analysis.

An important aspect for real-time applications that has not yet been considered in our current work is causality: all presented architectures ingest the entire video and therefore also account for information from future frames to predict the phase at a given point in time, which will not be possible in a system for real-time guidance. This limitation should be taken into consideration for modeling and during the training process, for example through the use of temporal masking as it is often done in language models (Boels et al., 2025). In future publications, we would investigate how to improve processing speed and how to incorporate causality into real-time predictions.

Our results demonstrate accurate offline phase recognition in SICS videos, which may be useful for workflow analysis, assisted quality control in high-volume surgical settings, and training support. Our work may lay the foundation for possible future intraoperative real-time guidance, complication detection, and autonomous surgical quality assessment. In addition, because recordings with complications requiring additional phases (e.g. vitrectomy) were excluded, the model has primarily been evaluated on routine SICS workflows rather than the full spectrum of complex cases.

7. Conclusions

Manual small-incision cataract surgery (SICS) is the predominant type of cataract surgery in the Global South, yet it has received minimal attention from the medical image analysis community. To address this, we published the novel SICS-155 dataset, which consists of video recordings of 155 SICS surgeries. We also organized an associated surgical phase recognition challenge as part of the MICCAI 2025 conference in Daejeon, South Korea.

The results of the challenge substantially improved on our recently published baseline (Mueller et al., 2025). Modern pre-trained feature extractors, such as VideoMAE and DINOv3, outperform older features, like inflated 3D features, as well as custom extractors, such as the Odd-K-Out approach. Furthermore, transformer architectures significantly outperformed CNN-based approaches. Follow-up experiments established a boundary-aware FACT architecture with VideoMAE features as a state-of-the-art approach for phase recognition in SICS and phacoemulsification. This achieved an accuracy of 87.98% [95%-CI: ± 2.06], an edit score of 90.80% [± 1.47], and a mean segmental F1-score of 89.56% [± 1.92] on the challenge dataset and competitive results on the Cataract-101 phacoemulsification public dataset that however depend on the exact data split, indicating that

transformer-based methods benefit from larger training sets.

Our top-performing model narrows the gap toward our established performance ceiling - the initial inter-rater agreement (accuracy 98.41%; edit-score 97.64%, F1-score 96.54%). These findings suggest that the model effectively addresses the primary challenges of SICS.

We plan to utilize these insights to facilitate the development of clinical training software for ophthalmologists in low- and middle-income countries. Furthermore, we hope that our dataset will encourage more research on SICS analysis within the broader scientific community. We are considering organizing follow-up challenges in the future that could include surgical quality assessment or complication detection.

Financial Disclosures

Author	Financial disclosures
Simon Müller	None
Bhuvan Sachdeva	Sankara Eye Foundation India, Bengaluru, India
Raphael Lechtenboehmer	Grants: Deutsche Ophthalmologische Gesellschaft, Platenstr. 1 80336 München; Baxter Healthcare Corporation Travel: Théa Pharma GmbH, Schillerstraße 3, 10625, Berlin
Singri Niharika Prasad	None
Frank G. Holz	Research Grant Support: Acucela, Allergan, Apellis, Bayer, Bioeq/Formycon, CenterVue, Ellex, Roche/Genentech, Geuder, Kanghong, NightStarx, Novartis, Optos, Zeiss; Consultant: Acucela, Aerie, Allergan, Apellis, Bayer, Boehringer-Ingelheim, ivericbio, Roche/Genentech, Geuder, Grayburg Vision, ivericbio, LinBioscience, Kanghong, Novartis, Pixium Vision, Oxurion, Stealth BioTherapeutics, Zeiss
Robert P. Finger	Consultant: Alimera, Apellis, Astellas, Bayer, Boehringer Ingelheim, Chiesi, OD-OS, Roche/Genentech Inc., Santen. Grant Support: Bayer, Biogen, Roche. Speaker: Alimera, Apellis, Bayer, Novartis, Roche
Kaushik Murali	None
Mohit Jain	Microsoft Research India
Maximilian W. M. Wintergerst	ASKIN & CO GmbH (R), Baxter Healthcare Corporation (F), Bayer AG (R), Berlin-Chemie AG (R, F), CenterVue SpA (F), Carl Zeiss Meditec (F), D-Eye Srl (F), DigiSight Technologies (R, F), Eyenuk, Inc. (F), Eyepress Fachmedien GmbH (R), glaucare GmbH (O, C), Heine Optotechnik GmbH (C, R, F), Heidelberg Engineering (R, F), Novartis Pharma GmbH (R, F), Optos (F), Pro Generika e.V. (R), Science Consulting in Diabetes GmbH (R)
Thomas Schultz	None
Nils D. Forkert	None
Yasser El Jarida	None
Youssef Iraqi	None
Loubna Mekouar	None
Christopher Nielsen	None
Patrick Gooi	None
Adrian Gooi	None
Nils D. Forkert	None
Yiran Kong	None
Zhihai Huang	None
Jiawei Du	None
Huiqi Li	None
Lasse Renz-Kiefel	Fraunhofer HHI
Eric L. Wisotzky	Fraunhofer HHI
Dong Zheng	None
Jiacheng Lin	None
Miao Hu	None
Yanwu Xu	None

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

The SICS-155 dataset generated and analyzed during the current study is available in the Zenodo repository, under the following URL:

<https://doi.org/10.5281/zenodo.15044589>. Currently the CC BY-NC (Attribution-NonCommercial) license applies to this data.

The dataset contains surgical phase ground truth, anonymized SICS videos, extracted i3D features, train/validation/test splits and mappings used by the deep neural network. Any additional data can be requested from the corresponding author on reasonable request.

Relevant hyperparameters are included in this published article (and its Supplementary Information files). Source code for the evaluation of the challenge and ranking is hosted on GitHub under the following URL: https://github.com/MedVisBonn/SICS155_challenge.

Declaration of generative AI utilization

During the preparation of this work the author(s) used Anthropic Claude Opus 4.8 in order to improve spelling, grammar and to suggest alternative sentence structures to improve readability. After using this tool, the author(s) reviewed and edited the content as needed and take (s) full responsibility for the content of the published article.

CRediT authorship contribution statement

Simon Mueller: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Bhuvan Sachdeva:** Writing – review & editing, Validation, Software, Formal analysis, Data curation. **Singri Niharika Prasad:** Writing – review & editing, Resources, Formal analysis, Data curation. **Raphael Lechtenboehmer:** Writing – review & editing, Investigation, Formal analysis, Data curation. **Frank G Holz:** Writing – review & editing, Supervision, Investigation, Formal analysis. **Robert P Finger:** Writing – review & editing, Supervision, Investigation, Formal analysis. **Christopher Nielsen:** Writing – review & editing, Software, Methodology. **Patrick Gooi:** Writing – review & editing, Software, Methodology. **Adrian Gooi:** Writing – review & editing, Software, Methodology. **Nils D. Forkert:** Writing – review & editing, Software, Methodology. **Yasser El Jarida:** Writing – review & editing, Software, Methodology. **Youssef Iraqi:** Writing – review & editing, Software, Methodology. **Loubna Mekouar:** Writing – review & editing, Software, Methodology. **Yiran Kong:** Writing – review & editing, Software, Methodology. **Zhihai Huang:** Writing – review & editing, Software, Methodology. **Jiawei Du:** Writing – review & editing, Software, Methodology. **Huiqi Li:** Writing – review & editing, Software, Methodology. **Lasse Renz-Kiefel:** Writing – review & editing, Software, Methodology. **Eric L. Wisotzky:** Writing – review & editing, Software, Methodology. **Dong Zheng:** Writing – review & editing, Software, Methodology. **Jiacheng Lin:** Writing – review & editing, Software, Methodology. **Miao Hu:** Writing – review & editing, Software, Methodology. **Yanwu Xu:** Project administration. **Mohit Jain:** Writing – review & editing, Supervision, Project administration, Investigation, Formal analysis. **Kaushik Murali:** Writing – review & editing, Supervision, Resources, Project administration, Data curation. **Maximilian W M Wintergerst:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Thomas Schultz:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Raphael Lechtenboehmer reports financial support was provided by German Society of Ophthalmology. Raphael Lechtenboehmer reports financial support was provided by Baxter Healthcare Corporation. Raphael Lechtenboehmer reports travel was provided by Théa Pharma. Frank G. Holz reports financial support was provided by Aerie Pharmaceuticals Inc. Frank G. Holz reports financial support was provided by

Apellis Pharmaceuticals, Inc. Frank G. Holz reports financial support was provided by Bayer Corporation. Frank G. Holz reports financial support was provided by Boehringer Ingelheim GmbH. Frank G. Holz reports financial support was provided by Roche Innovation Center Munich. Frank G. Holz reports financial support was provided by Geuder. Frank G. Holz reports financial support was provided by Chengdu Kanghong Biotechnology Inc. Frank G. Holz reports financial support was provided by NightstaRx Limited. Frank G. Holz reports financial support was provided by Novartis. Frank G. Holz reports financial support was provided by Optos. Frank G. Holz reports financial support was provided by Carl Zeiss AG. Robert P. Finger reports was provided by Alimera Sciences Ophthalmologie GmbH. Robert P. Finger reports financial support was provided by Apellis Pharmaceuticals, Inc. Robert P. Finger reports financial support was provided by Bayer AG. Robert P. Finger reports financial support was provided by Biogen. Robert P. Finger reports financial support was provided by Roche. Maximilian W. M. Wintergerst reports financial support was provided by ASKIN & CO GmbH. Maximilian W. M. Wintergerst reports financial support was provided by Baxter Healthcare Corporation. Maximilian W. M. Wintergerst reports financial support was provided by Bayer AG. Maximilian W. M. Wintergerst reports financial support was provided by Berlin-Chemie AG. Maximilian W. M. Wintergerst reports financial support was provided by Centervue SpA. Maximilian W. M. Wintergerst reports financial support was provided by Carl Zeiss Meditec Inc. Maximilian W. M. Wintergerst reports financial support was provided by DigiSight Technologies. Maximilian W. M. Wintergerst reports financial support was provided by Eyepress Fachmedien GmbH. Maximilian W. M. Wintergerst reports financial support was provided by GlauCare GmbH. Maximilian W. M. Wintergerst reports financial support was provided by Heine Optotechnologie. Maximilian W. M. Wintergerst reports financial support was provided by Heidelberg Engineering Inc. Maximilian W. M. Wintergerst reports financial support was provided by Novartis Pharma GmbH. Maximilian W. M. Wintergerst reports financial support was provided by Optos. Maximilian W. M. Wintergerst reports financial support was provided by Pro Generika eV. Maximilian W. M. Wintergerst reports financial support was provided by Science Consulting in Diabetes GmbH. Maximilian W. M. Wintergerst reports financial support was provided by Eyenuk. Bhuvan Sachdeva reports financial support was provided by Sankara Eye Care Institutions. Patrick Gooi reports financial support was provided by Cloudbreak Eye Care. Lassee Renz-Kiefel reports financial support was provided by Universitätsmedizin Rostock, Germany. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at [doi:10.1016/j.media.2026.104313](https://doi.org/10.1016/j.media.2026.104313).

References

- Ahmadi, M.J., Gandomi, I., Abdi, P., Mohammadi, S.F., Taslimi, A., Khodaparast, M., Hashemi, H., Tavakoli, M., H. D.T., 2026. Cataract-LMM large-scale multi-source multi-task benchmark for deep learning in surgical video analysis. *Sci. Data*. <https://doi.org/10.1038/s41597-026-07464-0>.
- Asbell, P.A., Dualan, I., Mindel, J., Brocks, D., Ahmad, M., Epstein, S., 2005. Age-related cataract. *Lancet* 365, 599–609. [https://doi.org/10.1016/S0140-6736\(05\)17911-2](https://doi.org/10.1016/S0140-6736(05)17911-2).
- Baltussen, R., Sylla, M., Mariotti, S.P., 2004. Cost-effectiveness analysis of cataract surgery: a global and regional analysis. *Bull. World Health Organ.* 82, 338–345. <https://www.ncbi.nlm.nih.gov/pubmed/15298224>.
- Bhagat, N., Nissirios, N., Potdevin, L., Chung, J., Lama, P., Zarbin, M.A., Fechtner, R., Guo, S., Chu, D., Langer, P., 2007. Complications in resident-performed phacoemulsification cataract surgery at New Jersey Medical School. *Br. J. Ophthalmol.* 91, 1315–1317. <https://doi.org/10.1136/bjo.2006.111971>.
- Boels, M., Liu, Y., Dasgupta, P., Granados, A., Ourselin, S., 2025. SWAG: long-term surgical workflow prediction with generative-based anticipation. *Int. J. Comput. Assist. Radiol. Surg.* <https://doi.org/10.1007/s11548-025-03452-8>.
- Carreira, J., Zisserman, A., 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 4724–4733. <https://doi.org/10.1109/cvpr.2017.502>.
- Dean, W.H., Murray, N.L., Buchan, J.C., Golnik, K., Kim, M.J., Burton, M.J., 2019. Ophthalmic simulated surgical competency assessment rubric for manual small-incision cataract surgery. *J. Cataract. Refract. Surg.* 45, 1252–1257. <https://doi.org/10.1016/j.jcrs.2019.04.010>.
- Fang, L., Mou, L., Gu, Y., Hu, Y., Chen, B., Chen, X., Wang, Y., Liu, J., Zhao, Y., 2022. Global-local multi-stage temporal convolutional network for cataract surgery phase recognition. *Biomed. Eng. Online* 21, 82. <https://doi.org/10.1186/s12938-022-01048-w>.
- Flaxman, S.R., Bourne, R.R.A., Resnikoff, S., Ackland, P., Braithwaite, T., Cicinelli, M.V., Das, A., Jonas, J.B., Keeffe, J., Kempen, J.H., Leasher, J., Limburg, H., Naidoo, K., Pesudovs, K., Silvester, A., Stevens, G.A., Tahhan, N., Wong, T.Y., Taylor, H.R., Vision Loss Expert Group of the Global Burden of Disease, S., 2017. Global causes of blindness and distance vision impairment 1990–2020: a systematic review and meta-analysis. *Lancet Glob. Health* 5, e1221–e1234. [https://doi.org/10.1016/S2214-109X\(17\)30393-5](https://doi.org/10.1016/S2214-109X(17)30393-5).
- Garcia Nespole, R., Yi, D., Cole, E., Valikodath, N., Luciano, C., Leiderman, Y.I., 2022. Evaluation of artificial intelligence-based intraoperative guidance tools for phacoemulsification cataract surgery. *JAMA Ophthalmol.* 140, 170–177. <https://doi.org/10.1001/jamaophthalmol.2021.5742>.
- Ghamsarian, N., El-Shabrawi, Y., Nasirihaghighi, S., Putzgruber-Adamitsch, D., Zinkernagel, M., Wolf, S., Schoeffmann, K., Sznitman, R., 2024. Cataract-1K dataset for deep-learning-assisted analysis of cataract surgery videos. *Sci. Data* 11, 373. <https://doi.org/10.1038/s41597-024-03193-4>.
- Goel, R., Shah, S., Malik, K.P.S., Sontakke, R., Golhait, P., Gaonker, T., 2022. Complications of manual small-incision cataract surgery. *Indian J. Ophthalmol.* 70, 3803–3811. <https://doi.org/10.4103/ijo.IJO.1812.22>.
- Golnik, K.C., Haripriya, A., Beaver, H., Gauba, V., Lee, A.G., Mayorga, E., Palis, G., Saleh, G.M., 2011. Cataract surgery skill assessment. *Ophthalmology*. 118, e2092. <https://doi.org/10.1016/j.ophtha.2011.06.039>.
- Grammatikopoulou, M., Flouty, E., Kadkhodamohammadi, A., Queller, G., Chow, A., Nehme, J., Luengo, I., Stoyanov, D., 2021. CaDIS: Cataract dataset for surgical RGB-image segmentation. *Med. Image Anal.* 71, 102053. <https://doi.org/10.1016/j.media.2021.102053>.
- Gurnani, B., K. K., 2023. Manual Small Incision Cataract Surgery. *StatPearls*. <https://www.ncbi.nlm.nih.gov/books/NBK582123/>.
- Hickman, M.S., Dean, W.H., Puri, L., Singh, S., Siegel, R., Patel, D., 2022. Ophthalmic telesurgery with a low-cost smartphone video system for surgeon self-reflection and remote synchronous consultation: a qualitative and quantitative study. *Telemed. Rep.* 3, 30–37. <https://doi.org/10.1089/tmr.2021.0037>.
- Hu, M., Xia, P., Wang, L., Yan, S., Tang, F., Xu, Z., Luo, Y., Song, K., Leitner, J., Cheng, X., Cheng, J., Liu, C., Zhou, K., Ge, Z., 2025. OphNet: a large-scale video benchmark for ophthalmic surgical workflow understanding. In: Leonadis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (Eds.), *Computer Vision – ECCV 2024*. Springer Nature Switzerland, Cham, pp. 481–500.
- Iashin, V., 2020. Video Features. *Github*. https://github.com/v-iashin/video_features.
- Jaggernath, J., Gogate, P., Moodley, V., Naidoo, K.S., 2014. Comparison of cataract surgery techniques: safety, efficacy, and cost-effectiveness. *Eur. J. Ophthalmol.* 24, 520–526. <https://doi.org/10.5301/ejo.5000413>.
- Joshua Ehrlich, B.H.F., Patel, Alpa S., Tripathy, Koushik, 2024. Manual Small Incision Cataract Surgery. *Eyewiki American Academy of Ophthalmology*. https://eyewiki.aaao.org/Manual_Small_Incision_Cataract_Surgery.
- Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D., 2017. Temporal convolutional networks for action segmentation and detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1003–1012. <https://doi.org/10.1109/cvpr.2017.113>.
- Lea, C., Vidal, R., Hager, G., 2016. Learning convolutional action primitives for fine-grained action recognition. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 1642–1649. <https://doi.org/10.1109/icra.2016.7487305>.
- Lindfield, R., Polack, S., Wadud, Z., Choudhury, K.A., Rashid, A.K., Kuper, H., 2008. Causes of poor outcome after cataract surgery in Satkhira district. *Bangladesh. Eye (L.)* 22, 1054–1056. <https://doi.org/10.1038/sj.eye.6702836>.
- Liu, Y.C., Wilkins, M., Kim, T., Malyugin, B., Mehta, J.S., 2017. Cataracts. *Lancet* 390, 600–612. [https://doi.org/10.1016/S0140-6736\(17\)30544-5](https://doi.org/10.1016/S0140-6736(17)30544-5).
- Lu, Z., Elhamifar, E., 2024. FACT: frame-action cross-attention temporal modeling for efficient action segmentation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18175–18185. <https://doi.org/10.1109/CVPR52733.2024.01721>.
- Mahmoud, O., Zhang, H., Matton, N., Mian, S.I., Tannen, B., Nallasamy, N., 2024. CatStep: automated cataract surgical phase classification and boundary segmentation leveraging inflated 3D-convolutional neural network architectures and BigCat. *Ophthalmol. Sci.* 4, 100405. <https://doi.org/10.1016/j.xops.2023.100405>.
- Maier-Hein, L., Reinke, A., Kozubek, M., Martel, A.L., Arbel, T., Eisenmann, M., Hanbury, A., Jannin, P., Müller, H., Onogur, S., Saez-Rodriguez, J., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., 2020. BIAS: transparent reporting of biomedical image analysis challenges. *Med. Image Anal.* 66, 101796. <https://doi.org/10.1016/j.media.2020.101796>.
- Mueller, S., 2025. Webpage of the SICS-155 Challenge. *Github Bonn, Germany*. Welcome to the Phase Recognition in Small Incision Cataract Surgery Videos (SICS-155) challenge, a Satellite Event of the MICCAI 2025 conference in Daejeon, Republic of Korea! <https://medvisb Bonn.github.io/>.
- Mueller, S., Jain, M., Sachdeva, B., Shah, P.N., Holz, F.G., Finger, R.P., Murali, K., Wintergerst, M.W.M., Schultz, T., 2024. Artificial intelligence in cataract surgery: a

- systematic review. *Transl. Vis. Sci. Technol.* 13, 20. <https://doi.org/10.1167/tvst.13.4.20>.
- Mueller, S., Sachdeva, B., Prasad, S.N., Lechtenboehmer, R., Holz, F.G., Finger, R.P., Kaushik, M., Jain, M., Wintergerst, M.W.M., Schultz, T., 2025. Phase recognition in manual small-incision cataract surgery with MS-TCN + + on the novel SICS-105 dataset. *Sci. Rep.* 15, 16886. <https://doi.org/10.1038/s41598-025-00303-z>.
- Muralikrishnan, R., Venkatesh, R., Prajna, N.V., Frick, K.D., 2004. Economic cost of cataract surgery procedures in an established eye care centre in Southern India. *Ophthalmic. Epidemiol.* 11, 369–380. <https://doi.org/10.1080/09286580490888762>.
- Murthy, G., Gupta, S.K., John, N., Vashist, P., 2008. Current status of cataract blindness and Vision 2020: the right to sight initiative in India. *Indian J. Ophthalmol.* 56, 489–494. <https://doi.org/10.4103/0301-4738.42774>.
- Muttenhaler, L., Dippel, J., Linhardt, L., Vandermeulen, R.A., Kornblith, S., 2023. Human alignment of neural network representations. *ICLR 2023*. Arxiv, San Diego. <https://doi.org/10.48550/arXiv.2211.01201>.
- Nizami, A.A., Gurnani, B., Gulani, A.C., 2022. Cataract. *StatPearls*. <https://pubmed.ncbi.nlm.nih.gov/30969521/>.
- Oladigbolu, K.K., Rafindadi, A.L., Mahmud-Ajeigbe, A.F., Chinda, D., Pam, V., Samaila, E., 2014. Outcome of cataract surgery in rural areas of Kaduna State, Nigeria. *Ann. Afr. Med.* 13, 25–29. <https://doi.org/10.4103/1596-3519.126943>.
- Pan, X., Zhang, Z., Chen, H., Wang, H., He, X., 2026. MACBNet: A multi-stage adaptive category balancing network for addressing class imbalance in surgical phase recognition. *Biomed. Signal. Process. Control* 113. <https://doi.org/10.1016/j.bspc.2025.109052>.
- Primus, M.J., Putzgruber-Adamitsch, D., Taschwer, M., Münzer, B., El-Shabrawi, Y., Böszörményi, L., Schoeffmann, K., 2018. Frame-based classification of operation phases in cataract surgery videos. In: Schoeffmann, K., Chalidabhongse, T.H., Ngo, C. W., Aramvith, S., O'Connor, N.E., Ho, Y.-S., Gabbouj, M., Elgammal, A. (Eds.), *MultiMedia Modeling*. Springer International Publishing, Cham, pp. 241–253.
- Rampat, R., Deshmukh, R., Chen, X., Ting, D.S.W., Said, D.G., Dua, H.S., Ting, D.J., 2021. Artificial intelligence in cornea, refractive surgery, and cataract: basic principles, clinical applications, and future directions. *Asia Pac. J. Ophthalmol. (Phila)* 10, 268–281. <https://doi.org/10.1097/APO.0000000000000394>.
- Raschka, S., 2024. Chapter 25: Confidence Intervals, Machine Learning Q and AI: 30 Essential Questions and Answers on Machine Learning and AI, 1st ed. No Starch Press, San Francisco, p. 264.
- Rossi, T., Romano, M.R., Iannetta, D., Romano, V., Gualdi, L., D'Agostino, I., Ripandelli, G., 2021. Cataract surgery practice patterns worldwide: a survey. *BMJ Open*. *Ophthalmol.* 6, e000464. <https://doi.org/10.1136/bmjophth-2020-000464>.
- Rückert, T., Rauber, D., Maerkl, R., Klausmann, L., Yildiran, S.R., Gutbrod, M., Nunes, D. W., Moreno, A.F., Luengo, I., Stoyanov, D., Toussaint, N., Cho, E., Kim, H.B., Choo, O.S., Kim, K.Y., Kim, S.T., Arantes, G., Song, K., Zhu, J., Xiong, J., Lin, T., Kikuchi, S., Matsuzaki, H., Kouno, A., Manesco, J.R.R., Papa, J.P., Choi, T.-M., Jeong, T.K., Park, J., Alabi, O., Wei, M., Vercauteren, T., Wu, R., Xu, M., an, W., Bai, L., Ren, H., Yamahli, A., Hennighausen, J., Maier-Hein, L., Kondo, S., Kasai, S., Hirasawa, K., Yang, S., Wang, Y., Chen, H., Rodríguez, S., Aparicio, N., Manrique, L., Lyons, J.C., Hosie, O., Ayobi, N., Arbeláez, P., Li, Y., Khalil, Y.A., Nasirihaghghi, S., Speidel, S., Rückert, D., Feussner, H., Wilhelm, D., Palm, C., 2025. Comparative validation of surgical phase recognition, instrument keypoint estimation, and instrument instance segmentation in endoscopy: Results of the PhaKIR 2024 challenge. *Med. Image Anal.* <https://doi.org/10.1016/j.media.2026.103945>.
- Sachdeva, B., Akash, N., Ashraf, T., Mueller, S., Schultz, T., Wintergerst, M.W.M., Jain, M., Singri, N., Kaushik, M., 2025. Phase-Informed Tool Segmentation for Manual Small-Incision Cataract Surgery, MICCAI 2025. Springer Nature Switzerland, South Korea, pp. 446–455. https://doi.org/10.1007/978-3-032-05114-1_43.
- Sarker, I.H., 2021. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN. Comput. Sci.* 2, 420. <https://doi.org/10.1007/s42979-021-00815-1>.
- Schoeffmann, K., Taschwer, M., Sarny, S., Münzer, B., Primus, M.J., Putzgruber, D., 2018. Cataract-101: video dataset of 101 cataract surgeries. In: *Proceedings of the 9th ACM Multimedia Systems Conference*. ACM, Amsterdam, Netherlands, pp. 421–425. <https://doi.org/10.1145/3204949.3208137>.
- Shah, N.A., Sikder, S., Vedula, S.S., Patel, V.M., 2023. GLSFormer: gated - long, short sequence transformer for step recognition in surgical videos. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (Eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Springer Nature Switzerland, Cham, pp. 386–396. https://doi.org/10.1007/978-3-031-43996-4_37.
- Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., 2025. DINOV3. *arXiv.preprint.arXiv:2508.10104*. <https://doi.org/10.48550/arXiv.2508.10104>.
- Singh, K., Misbah, A., Saluja, P., Singh, A.K., 2017. Review of manual small-incision cataract surgery. *Indian J. Ophthalmol.* 65, 1281–1288. <https://doi.org/10.4103/ijo.IJO.863.17>.
- Spencer, M., 2006. Phaco vs. small-incision. *Ophthalmology*. 113, 353. <https://doi.org/10.1016/j.ophtha.2005.10.038>.
- Steinbach, P., Gernhardt, F., Tanveer, M., Schmerler, S., Starke, S., 2022. Machine learning state-of-the-art with uncertainties. In: *International Conference on Learning Representations (ICLR) 2022*. ICLR, Virtual. <https://iclr.cc/virtual/2022/7327>.
- Tabin, G., Chen, M., Espandar, L., 2008. Cataract surgery for the developing world. *Curr. Opin. Ophthalmol.* 19, 55–59. <https://doi.org/10.1097/ICU.0b013e3282f154bd>.
- Thia, B.C., Wong, N.J., Sheth, S.J., 2019. Video recording in ophthalmic surgery. *Surv. Ophthalmol.* 64, 570–578. <https://doi.org/10.1016/j.survophthal.2019.01.005>.
- Tong, Z., Song, Y., Wang, J., Wang, L., 2022. VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Neural Information Processing Systems Foundation, Inc. (NeurIPS). New Orleans, LA, USA, 732. <https://doi.org/10.52202/068431-0732> p. Article.
- Touma, S., Antaki, F., Duval, R., 2022. Development of a code-free machine learning model for the classification of cataract surgery phases. *Sci. Rep.* 12, 2398. <https://doi.org/10.1038/s41598-022-06127-5>.
- Valikodath, N.G., Cole, E., Ting, D.S.W., Campbell, J.P., Pasquale, L.R., Chiang, M.F., Chan, R.V.P., 2021. Impact of artificial intelligence on medical education in ophthalmology. *Transl. Vis. Sci. Amp; Technol.* 10, 14. <https://doi.org/10.1167/tvst.10.7.14>.
- Ward, T.M., Fer, D.M., Ban, Y., Rosman, G., Meireles, O.R., Hashimoto, D.A., 2021. Challenges in surgical video annotation. *Comput. Assist. Surg. (Abingdon)* 26, 58–68. <https://doi.org/10.1080/24699322.2021.1937320>.
- Yi, F., Wen, H., Jiang, T., 2021. ASFormer: transformer for action segmentation. In: *Proceedings of the British Machine Vision Conference 2021*. British Machine Vision Association, United Kingdom (UK). <https://doi.org/10.5244/c.35.49> virtual.
- Zhang, B., Sarhan, M.H., Goel, B., Petculescu, S., Ghanem, A., 2024. SF-TMN: SlowFast temporal modeling network for surgical phase recognition. *Int. J. Comput. Assist. Radiol. Surg.* 19, 871–880. <https://doi.org/10.1007/s11548-024-03095-1>.