OphNet: A Large-Scale Video Benchmark for Ophthalmic Surgical Workflow Understanding

(ECCV 2024)

➢ Background

Baret et al.[1] showed networks, like Inflated 3D ConvNet (I3D) that utilize spatiotemporal convolutions, require a relatively extensive dataset for effective training. In their study, the model achieves an accuracy exceeding 80% when trained on 100 videos, with a progressive improvement as the sample size surpasses 700. However, the highly efficient and rapidly evolving deep learning technologies for surgical workflow analysis are currently limited by the following shortcomings in current video benchmarks:

(1) Small-Scale: the majority of surgical video datasets contain no more than 100 videos. The small size of these datasets can lead to underfitting or overfitting of the model, and can also impact the model's ability to generalize;

(2) Limited Categories of Surgeries and Phases: they often only include a single type of surgical label, such as *'cataract surgery'* and a few phases, which does not reflect the real clinical surgical environment;

(3) Single-Boundary Annotation: they only annotate designated phases in the videos, ignoring the continuity of various phases in ophthalmic surgery;

(4) Uniform Domain: the videos are meticulously collected, and while this ensures video quality, the uniform style is not conducive to testing the model's domain generalization ability.

[1] Omri Bar, Daniel Neimark, Maya Zohar, Gregory D. Hager, Ross Girshick, Gerald M. Fried, Tamir Wolf, and Dotan Asselmann. Impact of data on generalization of ai for surgical intelligence applications. Scientific Reports, 10(1):22208, 2020.

➢ OphNet Overview

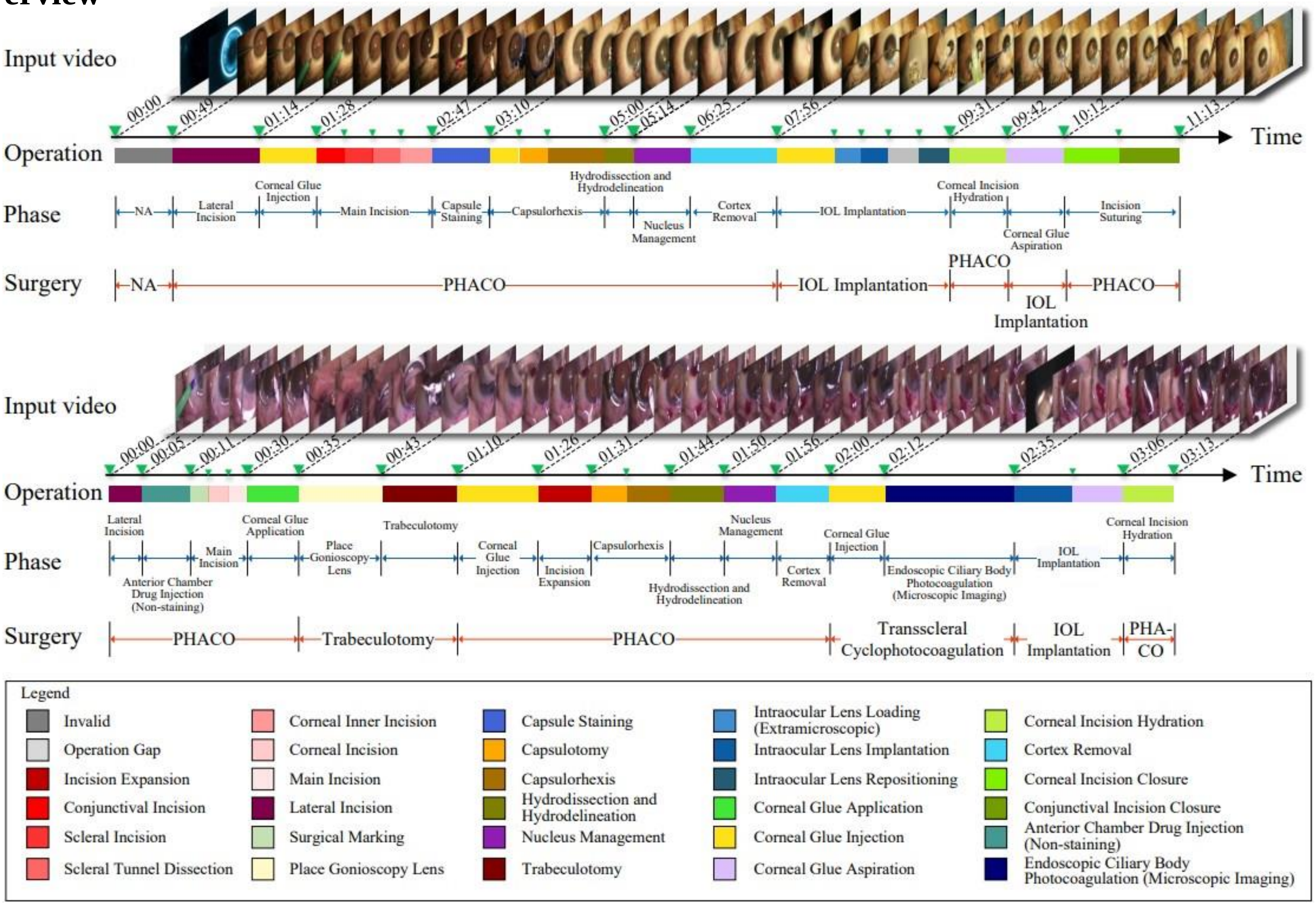

Figure 1. The examples of surgical and phase boundary annotations. The figure shows two surgical videos, PHACO + IOL implantation and ECCE (Extracapsular Cataract Extraction) + IOL implantation. For each frame marked in color, we provide temporal boundary annotations at both the surgery and phase levels.

➢ OphNet Statistics

Protocol	Datasets	No. of Videos	No. of Action Segments	No. of Surgery Categories	No. of Action Categories	Total Duration	Multi-Surgery Presence Recognition	Phase Recognition	Phase Localization	Phase Prediction
		Dataset Properties					Tasks			
Endo&Lap	Cholec120 [45]	120	-	1	7	76.2h	✗	✓	✗	✗
	SurgicalActions160 [54]	160	160	1	16	0.2h	✗	✓	✗	✗
	HeiCo [39, 50]	30	-	3	14	2.8h	✗	✓	✓	✓
	EndoVis 2021 [67]	33	250	1	7	22.0h	✗	✓	✗	✗
	PitVis [3]	25	287	1	17	33.3h	✗	✓	✗	✗
	CholecT50 [46]	50	-	1	10	44.7h	✗	✓	✓	✓
	AutoLaparo [70]	21	300	1	7	23.1h	✗	✓	✓	✓
OphScope	LensID [22]	100	2,440	1	2	11.7h	✗	✓	✗	✗
	Cataract-101 [55]	101	1,266	1	10	14.0h	✗	✓	✓	✓
	CatRelDet [23]	21	2,400	1	4	2.0h	✗	✓	✗	✗
	CATARACTS [4]	50	1,536	1	19	20.0h	✗	✓	✓	✓
	Cataract-1K [21]	1,000	931	1	12	118.7h	✗	✓	✓	✓
	OphNet(Ours)	**2,278**	**9,795**	**66**	**150**	**284.8h**	✓	✓	✓	✓

Table 1. The statistics comparison among existing workflow analysis datasets and our OphNet. Compared to other datasets, OphNet focuses on more comprehensive coverage of various surgery, phase and operation categories, collects a large number of videos, totaling 204.8 hours, and also enables a variety of recognition, localization and prediction tasks. OphNet demonstrates considerable competitiveness in both its scale and the richness of its labels. Endo\&Lap denotes the endoscopic and laparoscopic protocol, OphScope denotes the ophthalmic microscope protocol. We choose the latest version for comparison in cases where datasets have multiple supplementary updates. For instance, Cholec120, Cholec80, m2cai-workflow and LapChole form one series, whereas CholecT50, CholecT45, and CholecT40 comprise another series. We have excluded the following scenarios from our comparison: (1) non-open-source datasets such as Bypass170, ESD, Yu's, etc.; (2) a superset of multiple open-source or non-open-source datasets, like Cholec207, etc.; (3) datasets employed for lesion, anatomy, and instrument classification and segmentation, such as SUN-SEG, CVC-ClinicDB, ROBUST-MIS, Mesejo's, Cata7, etc., anomaly detection such as PolypDiag (from Hyper-Kvasir and LDPolypVideo), Kvasir-Capsule, etc., and other datasets not dedicated to workflow analysis. It's worth mentioning that even in comparison with the above datasets, OphNet demonstrates considerable competitiveness in both its scale and the richness of its labels.

➢ OphNet Statistics

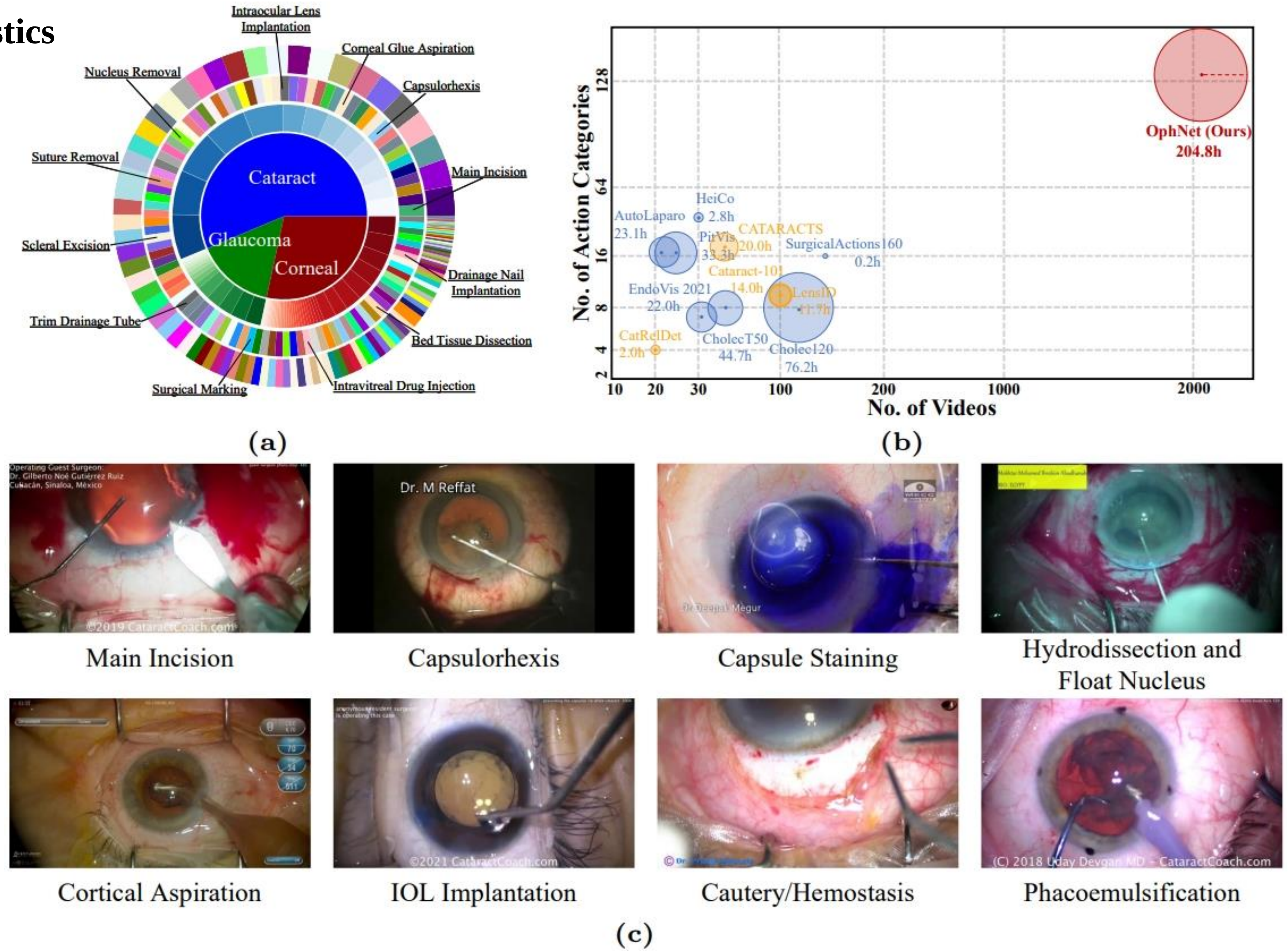

Figure 2. OphNet's composition, comparison with other datasets for the same task, and some phase examples: (a) an overview of the composition ratios at the levels of surgery, phase, and operation; (b) comparison among existing open-source laparoscopic & endoscopic, and ophthalmic microscope workflow analysis video datasets and our OphNet. OphNet stands as the largest real-world video dataset for ophthalmic surgical workflow understanding, featuring the highest number of videos, longest duration, and diverse categories of surgeries and phases; (c) eight phase examples in OphNet.

➢ Baselines

Table 2. Per-class Top-1 and Top-5 accuracy (%) for the primary surgery presence recognition on untrimmed videos and phase recognition on trimmed videos. * denotes the initialization from the model pre-trained on Kinetics 400. For the two CLIP models, we chose ViT-B/16 as the backbone and compared the performance of two different input frame numbers, 16 and 32. The best performance for each split has been highlighted in bold.

Baselines	Phase Classification								Operation Classification							
	Cataract		Glaucoma		Cornea		All		Cataract		Glaucoma		Cornea		All	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
I3D [9]	27.2	55.7	24.1	57.5	18.9	52.1	25.7	58.2	26.8	54.9	23.5	56.0	18.0	51.2	25.0	57.1
SlowFast [19]	26.5	56.5	23.1	56.5	24.2	49.1	26.7	60.1	25.8	55.2	22.9	45.9	23.5	48.5	26.0	59.0
X3D [36]	27.0	58.3	21.0	55.5	21.8	28.5	26.6	62.3	26.4	47.2	20.5	44.6	21.3	27.8	26.1	61.5
MViT V2 [18]	26.2	54.9	21.0	53.4	26.0	46.8	27.0	59.8	25.9	43.8	20.5	42.7	25.5	45.9	26.5	58.9
I3D*	29.5	68.9	22.1	58.6	26.0	50.9	30.2	71.2	28.8	67.5	21.8	47.9	25.7	50.3	29.5	60.0
SlowFast*	30.6	72.3	25.2	54.7	30.7	59.8	31.7	61.8	29.9	71.1	24.8	43.9	29.5	58.7	30.5	60.9
X3D*	27.2	72.9	22.1	59.6	30.7	61.5	33.5	63.2	26.5	71.8	21.7	48.8	29.9	60.2	32.8	62.1
MViT V2*	34.2	76.5	23.3	52.0	38.4	65.1	28.3	60.2	33.5	75.2	22.8	41.5	37.9	64.0	27.8	59.5
X-CLIP$_{16}$ [49]	68.3	92.2	47.3	89.8	53.0	77.4	63.4	85.3	67.5	91.0	46.5	78.9	82.2	76.1	62.5	84.0
X-CLIP$_{32}$	69.1	**94.0**	48.7	81.7	54.8	80.4	62.7	85.8	68.0	93.0	**47.9**	**80.5**	84.0	79.5	62.0	84.7
ViFi-CLIP$_{16}$ [57]	**75.9**	93.7	40.4	85.4	66.6	81.6	66.1	**88.4**	74.5	92.5	42.8	74.5	**85.0**	80.5	**65.0**	**87.5**
ViFi-CLIP$_{32}$	73.0	92.9	**49.6**	**82.7**	**57.7**	**81.6**	**68.4**	87.2	**75.1**	**93.8**	43.2	80.2	83.7	85.2	64.8	86.5

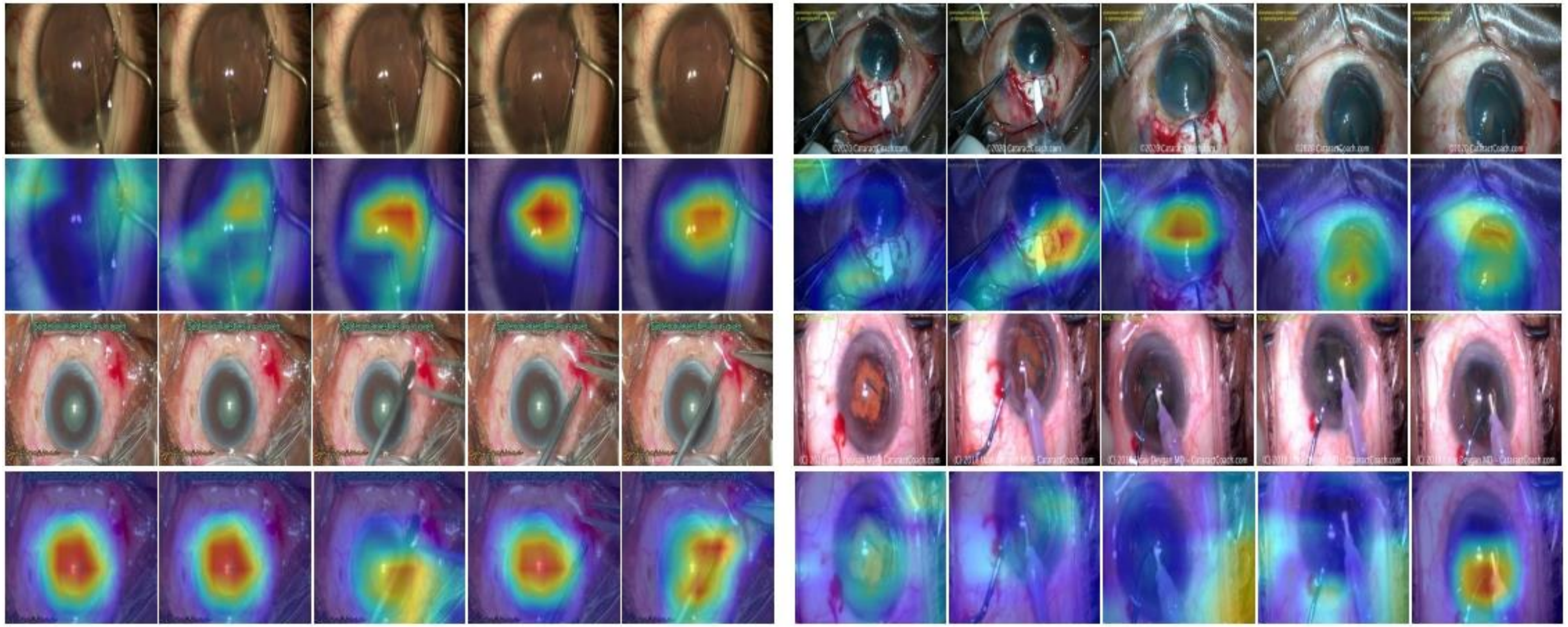

Figure 3. Attention map visualizations of ViFi-CLIP on four examples from OphNet test set in the phase recognition task.

➢ Baselines

Baselines	Backbones	mAP (%)				
		0.1	0.3	0.5	0.7	Avg.
ActionFormer [89]	CSN [73]	53.7	50.1	40.6	24.5	42.5
	SwinViviT [41]	59.3	54.7	43.3	26.3	46.4
	SlowFast [19]	60.0	55.9	45.1	26.0	47.5
TriDet [64]	CSN	56.1	53.0	43.1	29.4	46.2
	SwinViviT	61.0	**57.1**	**47.1**	**33.1**	**50.4**
	SlowFast	**61.3**	56.0	45.6	30.4	48.6

Table 3. The results for phase detection. ActionFormer and TriDet are state-of-the-art models for human action detection tasks, and we use three different backbones for feature extraction and report mAP at the IoU thresholds of [0.1:0.2:0.9]. Average mAP is computed by averaging different IoU thresholds. The best performance for each split has been highlighted in bold.

Baselines	Top-1 Acc. (%)				
	0.1	0.3	0.5	0.7	Avg.
I3D [9]	26.5	42.2	49.8	51.3	47.3
SlowFast [19]	25.4	42.6	48.9	52.2	47.2
MViT V2 [18]	25.6	43.7	49.3	52.3	47.5
I3D*	27.3	43.5	50.1	51.4	47.6
SlowFast*	27.5	43.2	49.9	**52.3**	47.8
MViT V2*	**27.8**	**43.8**	**50.5**	51.7	**48.2**

Table 4. The results for phase anticipation. We report top-1 accuracy at the observation ratios [0.1:0.2:0.9]. Average top-1 accuracy is computed by averaging different observation ratios. The best performance for each split has been highlighted in bold.

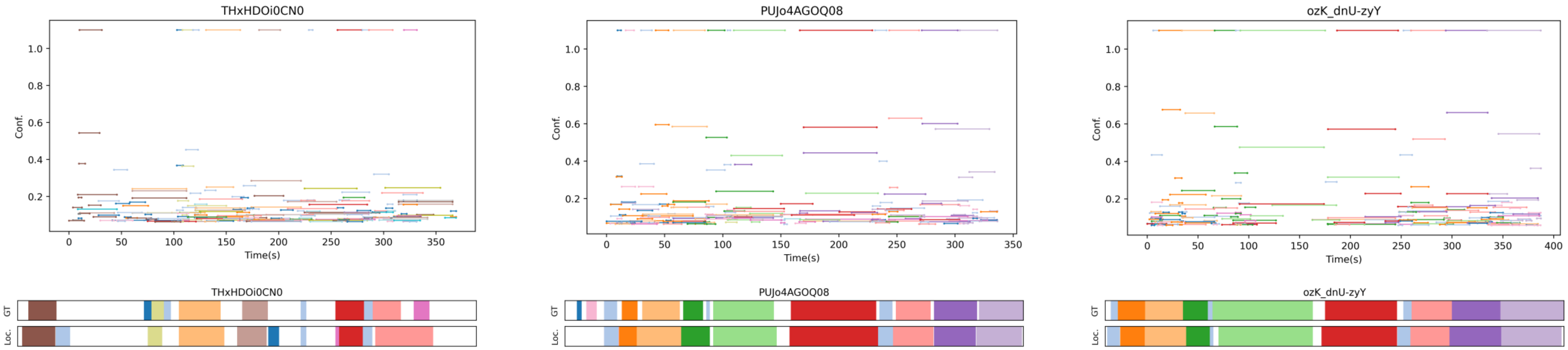

Figure 4. Phase localization visualization of TriDet. From top to bottom: visualization of groundtruth and results of all confidence phases.

➢ Interface

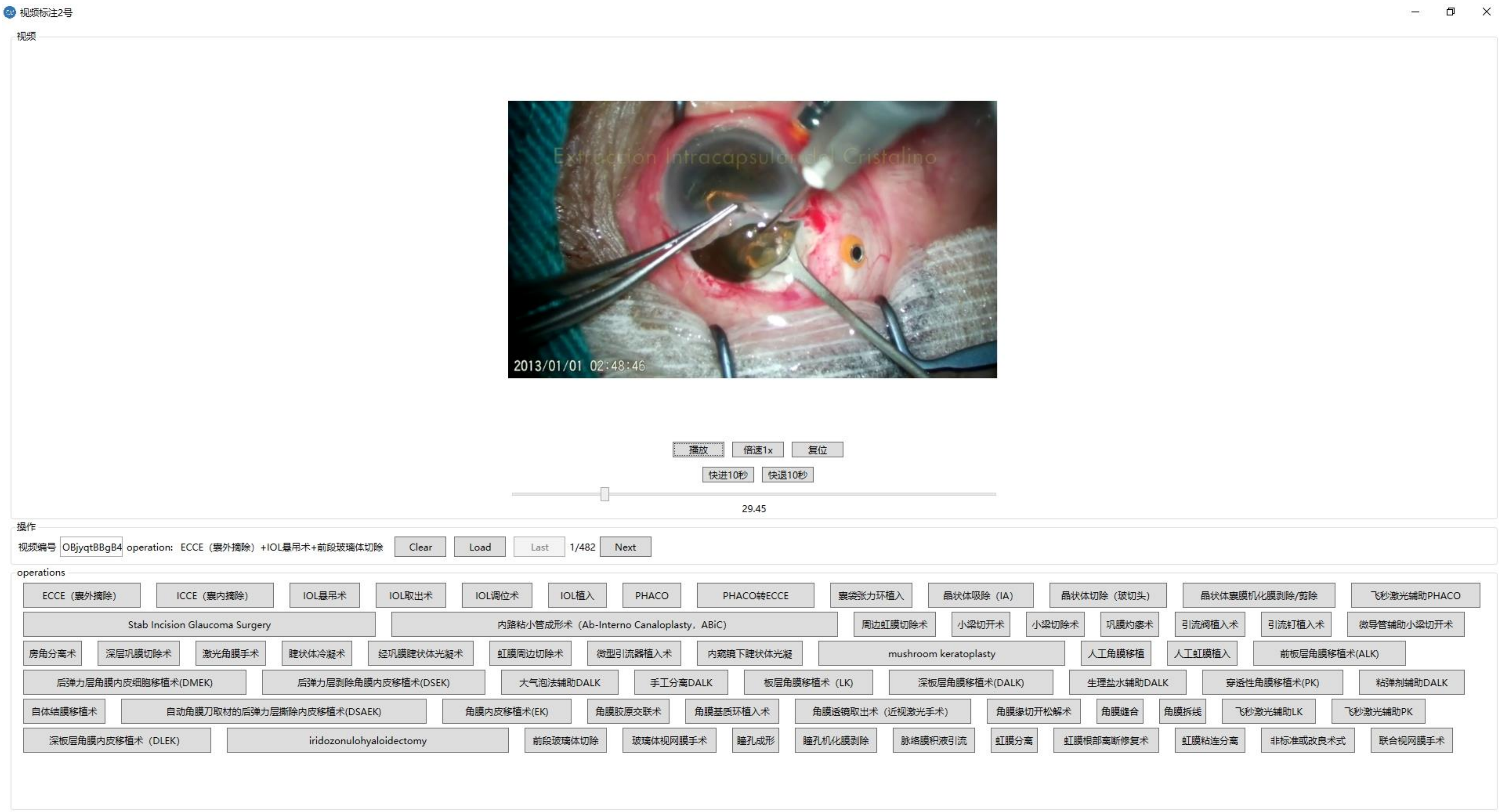

Figure 5. Video filtering and surgery classification annotation interface

➢ Interface

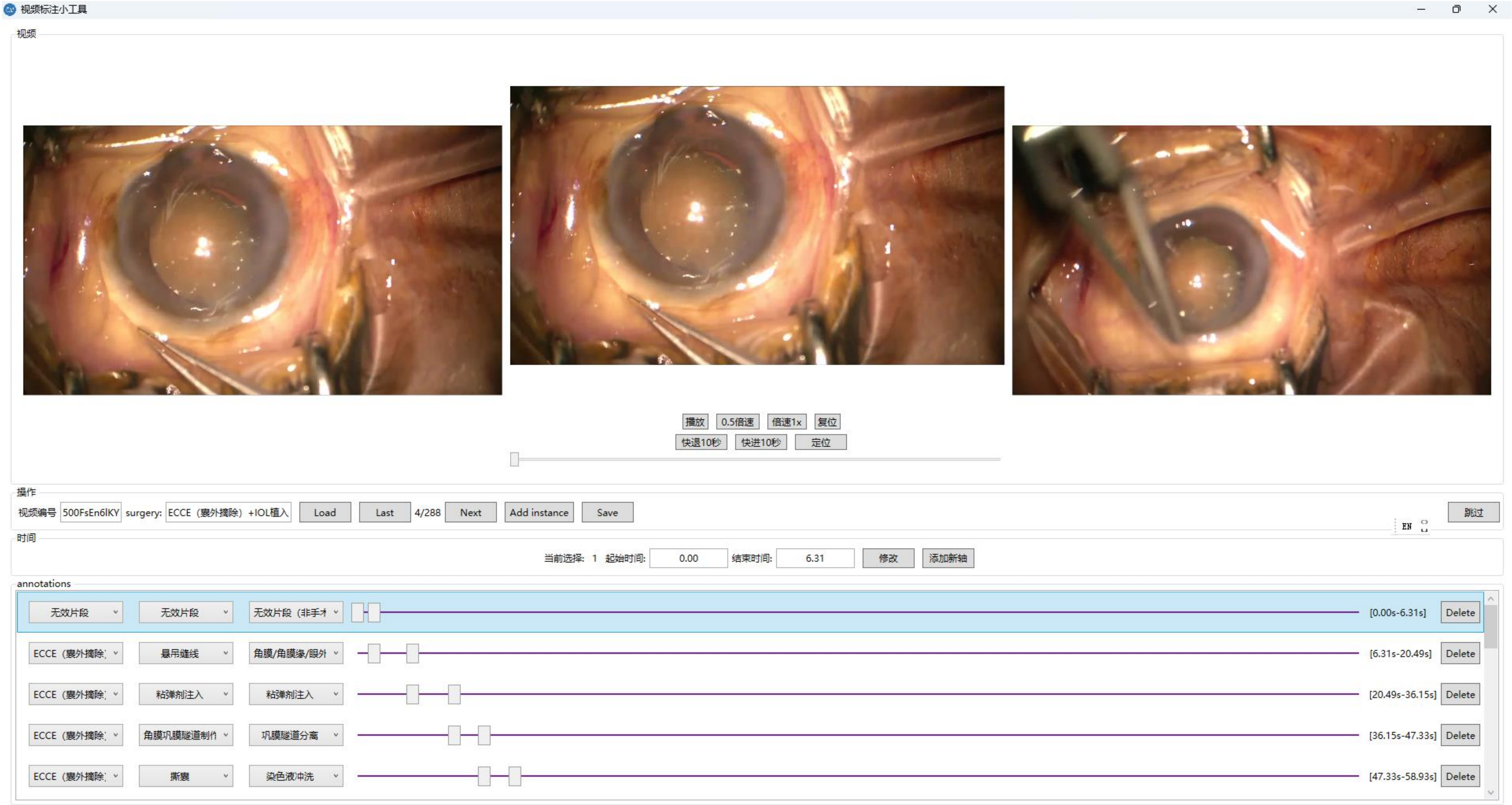

Figure 6. Hierarchical temporal localization annotation interface for surgery, phase, and operation