OphCLIP: Hierarchical Retrieval-Augmented Learning for Ophthalmic Surgical Video-Language Pretraining

Supplementary Material

7. OphVL Dataset

Fig. 6 illustrates the clip-text pair samples we constructed. Through our data processing pipeline, OphVL achieves high-quality modality alignment between ophthalmic surgery videos and descriptive texts.

8. Experiments

8.1. Implementation Details.

Hyper-parameter		Value
Epochs		60
Clip-level Pretraining	Batch Size	120
	Image Size	224
	# of Frames	8
	Text Length	77
Clip-level Pretraining	Batch Size	140
	Image Size	224
	# of Frames	8
	# of Retrieved Videos	1
	Text Length	77
Optimization	Learning Rate	8e-5
	Scheduler	Cosine
	Optimizer	Adam
	Momentum	0.9
Loss Function	Temperature	0.1
	Weight of $L_{\text{clip}}^{\text{vv}}$	0.5
	Weight of $L_{\text{clip}}^{\text{vl}}$	0.5

Table 6. Hyper-parameter details.

Architecture. We use the CLIP-like architecture [48] with two branches, i.e., visual and textual encoders. We use the ResNet-50 as the visual encoder from the ImageNet initialization. We apply BioClinicalBert [19] as the textual encoder, which is pretrained on the clinical notes. Then we apply the average pooling at the visual features to generate the visual embeddings. We apply a linear projection layer at the end of Bert model's [*CLS*] token to generate textual embeddings. We use 768 as the dimensionality of the embedding space.

Pretraining Setups. In total, we use 8 RTX-4090 with 24 GB and train for 2 days. We first perform clip-level pretraining for 40 epochs and then apply the hierarchical pretraining strategy, which alternatively trains with 3 epochs of clip-level video-text pairs, followed by 2 epochs of video-level video-text pairs. We use a batch of 120/140 for the clip- and video-level pretraining, respectively. More hyper-parameter details can be found in Tab. 6.

8.2. Evaluation Setup.

We evaluate the representation ability of our OphCLIP using two types of downstream tasks: surgical phase recognition and surgical tool recognition. Additionally, we conduct zero-shot evaluation and linear probing to assess the model's multi-modal alignment and visual representation capabilities. Tables 11-19 list the specific label names we used for the downstream validation datasets. The labels for the OphNet dataset can be found in the online table: https://docs.google.com/spreadsheets/d/1p5lURkth587-lxYwd6eOSmSxPpvIqvyuOKW-4B49PT0/edit?gid=0#gid=0

Surgical Phase Recognition. This task evaluates the model's understanding of surgical scenes by classifying video frames into predefined surgical phases. It requires the model to identify instruments, anatomical structures, and their interactions by extracting meaningful visual patterns. To focus on multi-modal representation learning, we exclude temporal modeling and analyze frame-level understanding instead.

Surgical Tool Recognition. This task tests the model's ability to detect and classify surgical instruments within video frames. By analyzing visual features like shape, texture, and contextual cues, the model demonstrates object-level understanding without reliance on broader workflow context. We assess its robustness in identifying tools despite variations in orientation, scale, or occlusion, emphasizing the quality of learned visual representations.

Instrument Label	Textual Prompt
Capsulorhexis Forceps	This video shows capsulorhexis forceps.
Capsulorhexis Cystotome	This video shows capsulorhexis cystotome.
Katena Forceps	This video shows katena forceps.
Irrigation-Aspiration	This video shows irrigation aspiration.
Slit Knife	This video shows slit knife.
Phacoemulsification Tip	This video shows phacoemulsification tip.
Spatula	This video shows spatula.
Gauge	This video shows gauge.
Lens Injector	This video shows lens injector.
Incision Knife	This video shows incision knife.

Table 7. Textual prompts for each instrument label in the Cataract-1K dataset.

Zero-shot Evaluation. To perform frame-wise classification tasks for surgical phase and tool recognition, we construct textual prompts tailored to the class labels. For phase recognition, we address their high-level definitions by breaking them down into essential components such as

Instrument Label	**Precision**	**Recall**	**F1**	**Support**
Capsulorhexis Forceps	6.1	100.0	11.5	100
Capsulorhexis Cystotome	4.8	100.0	9.1	85
Katena Forceps	1.6	100.0	3.1	28
Irrigation-Aspiration	25.4	100.0	40.5	451
Slit Knife	1.6	100.0	3.1	28
Phacoemulsification Tip	30.7	100.0	46.9	545
Spatula	40.3	100.0	57.4	716
Gauge	24.0	100.0	38.7	426
Lens Injector	3.7	100.0	7.2	66
Incision Knife	1.2	100.0	2.4	22
Macro Avg.	13.9	100.0	22.0	2475

Table 8. Detailed instrument recognition performance of CLIP [48], SLIP [39], and LaCLIP[9] on Cat-21 dataset in terms of each class label.

phase, instrument, medication, and goal. These are referred to as keyword-only prompts as shown in Tab. 10. Additionally, we leverage Large Language Models (LLMs) to generate caption-only prompts, which are detailed descriptive sentences that incorporate relevant surgical instruments, anatomical structures, and events for each phase. These prompts help align the textual domain of pretraining with the downstream task corpus. For tool recognition, we create human-like descriptive sentences to minimize the textual domain gap, ensuring better alignment between pretraining and downstream corpus, as shown in Tab. 7. This approach facilitates robust zero-shot performance by bridging differences in textual contexts.

Linear-Probing Evaluation. For linear-probing, we freeze the visual encoder and train a linear classifier on the extracted features. No image augmentations are applied during training. The linear classifier is implemented as a linear Support Vector Machine (SVM) with a "linear" kernel. We fit the model on the training and validation sets, then evaluate its performance on a separate test set. For few-shot linear-probing, we use a *k*-percentage shot approach, tailored for surgical video data. Specifically, we sample 10% of the videos from the training set, ensuring no data leakage while maintaining a balanced number of samples across classes. This setup allows for a fair evaluation of the model's generalization with limited supervision.

8.3. More Ablation Experiments

Tab. 9 presents additional results of ablation experiments on the Cataract-101 [57] and CatRelDet [11] datasets.

9. Limitation

Data Bias. The OphVL dataset is sourced from YouTube, showcasing diverse styles, resolutions, and screen elements. This diversity enhances the evaluation of a model's generalization ability but may also impact its effectiveness and performance. Some videos in the dataset contain subtitles, watermarks, or additional video windows. Furthermore, regional variability introduces discrepancies in surgical descriptions, such as differences in surgical standards, nomenclature, and definitions influenced by cultural or demographic factors. These characteristics in OphVL reflect the complexity of real-world surgical environments, where ophthalmic microscopes may inherently display various windows or parameters during recording. While these factors pose challenges, they also present opportunities to develop models that are better equipped to handle such diversity.

Model	OphVL	KB	Prompt	Cataract-101 [57]	CatRelDet [11]
CLIP [48]	×	×	Caption	10.0 / 3.3	15.3 / 11.9
CLIP [48]	✓	×	Caption	36.2 / 25.5	26.7 / 23.7
OphCLIP	✓	×	Caption	37.1 / 31.9	33.6 / 35.4
OphCLIP	✓	×	Mix	31.9 / 28.4	**34.5 / 36.1**
OphCLIP	✓	✓	Caption	**41.1 / 34.7**	33.6 / 35.3
OphCLIP	✓	✓	Mix	39.3 / 33.7	32.6 / 34.2

Table 9. Ablation study of OphCLIP with various components: OphVL (use of the OphVL pretraining dataset), KB (knowledge base with silent videos), and Prompt (descriptive phase prompts: Caption vs. Mix, which includes additional keywords in the captions). We report Accuracy / F1-score in this table.

Downstream Task Limitation. The zero-shot downstream evaluation datasets for OphCLIP are sourced from publicly available datasets, leveraging their high-quality characteristics and ensuring fair comparisons. However, due to the limited diversity of these datasets—most of which primarily focus on phase recognition and instrument classification in ophthalmology—it is challenging to validate the model on a broader range of vision-language understanding tasks, such as lesion identification or anomaly detection. While the Cataract-1K dataset includes annotations for two types of anomalies, lens rotation and pupil reaction, it does not provide frame-level annotations for these cases.

Phase Label	Caption Only Prompt	Keyword Only Prompt

Incision	A diamond or steel keratome blade is used to create a small, self-sealing incision in the cornea, providing access to the anterior chamber of the eye. This incision allows the introduction of surgical instruments while maintaining intraocular pressure.	Phase: Initial access;
		Instrument: Diamond or steel blade;
		Medication: None;
		Goal: Create an entry point into the anterior chamber.

Viscoelastic	A viscoelastic agent, such as sodium hyaluronate, is injected into the anterior chamber using a cannula. This agent maintains space, protects the corneal endothelium, and stabilizes the anterior chamber during the surgery.	Phase: Chamber stabilization;
		Instrument: Syringe or cannula;
		Medication: Ophthalmic Viscoelastic Device (OVD);
		Goal: Maintain anterior chamber depth and protect corneal endothelium.

Capsulorhexis	Using capsulorhexis forceps or a cystotome, the surgeon creates a circular tear in the anterior lens capsule. This opening allows access to the underlying cataractous lens, preparing it for removal.	Phase: Capsule opening;
		Instrument: Forceps or cystotome;
		Medication: None (Viscoelastic used for support);
		Goal: Create a circular opening in the anterior lens capsule.

Hydrodissection	Balanced salt solution (BSS) is injected with a cannula between the lens capsule and the lens cortex, separating the cataract from the capsule. This ensures that the lens material can be removed more easily during phacoemulsification.	Phase: Lens loosening;
		Instrument: Cannula;
		Medication: Balanced Salt Solution (BSS);
		Goal: Separate the lens cortex from the capsule for easy extraction.

Phacoemulsification	A phacoemulsification handpiece with an ultrasonic probe is inserted into the eye to emulsify the cataract into tiny fragments. These fragments are simultaneously aspirated, removing the clouded lens while protecting surrounding structures.	Phase: Lens removal; Instrument: Phacoemulsification handpiece; Medication: Balanced Salt Solution (BSS) for cooling and irrigation; Goal: Break up and emulsify the cataract for extraction.

Irrigation/Aspiration	A dual-function irrigation and aspiration (I/A) handpiece is used to remove any remaining lens material and fluid from the capsular bag and anterior chamber. The procedure ensures the capsular bag is clear for lens implantation.	Phase: Lens material removal;
		Instrument: Irrigation/Aspiration handpiece;
		Medication: Balanced Salt Solution (BSS);
		Goal: Remove remaining lens fragments from the capsular bag.

Capsule Pulishing	A polishing tip or I/A tool is used to gently remove residual epithelial cells from the inner surface of the posterior capsule, minimizing the risk of posterior capsule opacification (secondary cataract formation).	Phase: Capsule cleaning;
		Instrument: Polishing tip or Irrigation/Aspiration tool;
		Medication: None;
		Goal: Remove residual lens epithelial cells to reduce posterior capsule opacification.

Lens Implantation	An intraocular lens (IOL) is loaded into an injector and inserted through the corneal incision. It is placed within the capsular bag to replace the natural lens and restore the patient's vision.	Phase: Lens insertion;
		Instrument: Intraocular lens (IOL) injector;
		Medication: None;
		Goal: Insert the artificial intraocular lens into the capsular bag.

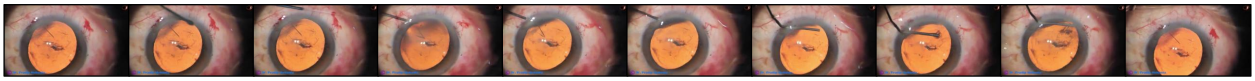
Lens positioning	Using fine-tipped instruments, the surgeon carefully adjusts the position of the IOL within the capsular bag to ensure proper centration and stability, optimizing visual outcomes.	Phase: Lens alignment;
		Instrument: Manipulating hook or forceps;
		Medication: None;
		Goal: Ensure the intraocular lens is correctly positioned and centered.

Viscoelastic_Suction	The viscoelastic agent is aspirated from the anterior chamber using the I/A handpiece to prevent postoperative pressure spikes and ensure a clear visual axis.	Phase: Viscoelastic removal;
		Instrument: Irrigation/Aspiration handpiece;
		Medication: None;
		Goal: Remove any remaining viscoelastic agents from the anterior chamber.

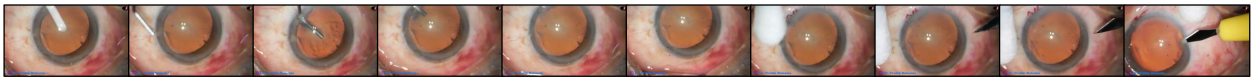
Anterior_Chamber Flushing	The anterior chamber is flushed with balanced salt solution (BSS) to remove any remaining debris or blood. This final rinse ensures that the chamber is clear and that the incision site is clean.	Phase: Final chamber cleaning;
		Instrument: Irrigation/Aspiration handpiece;
		Medication: Balanced Salt Solution (BSS);
		Goal: Ensure the anterior chamber is clear of any debris or substances.

Tonifying/Antibiotics	A pupil-constricting agent, such as acetylcholine, may be injected to stabilize intraocular pressure. Following this, an antibiotic such as moxifloxacin is administered to prevent infection, and sometimes corticosteroids are used to reduce inflammation.	Phase: Final stabilization and protection;
		Instrument: Syringe or cannula;
		Medication: Acetylcholine (for pupil constriction) and moxifloxacin (antibiotic);
		Goal: Stabilize intraocular pressure and prevent infection.

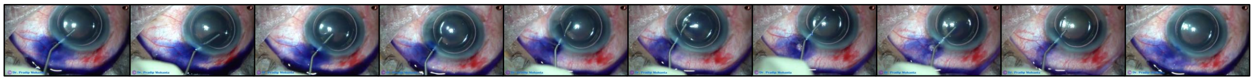
Table 10. **Prompt example.** Caption-only and keyword-only prompts for each phase label in the Cataract-1K dataset, respectively.



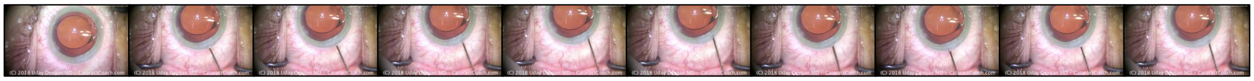
This video details the hydrodissection phase of the surgery, utilizing a 27-gauge cannula.



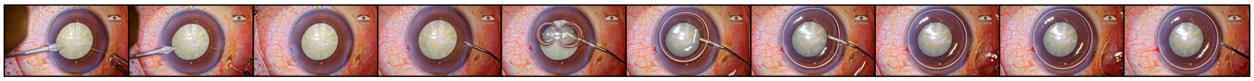
This video demonstrates a surgical procedure beginning with the main incision, followed by the filling of the anterior chamber with 2% hydroxypropyl methylcellulose. Subsequently, a side port incision is created at the 2 o'clock position.



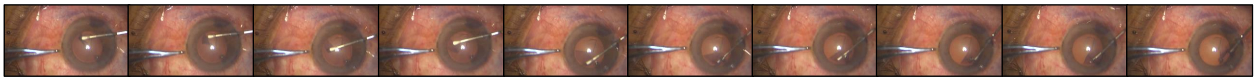
This video demonstrates the staining of the anterior capsule with trypan blue dye under an air bubble.



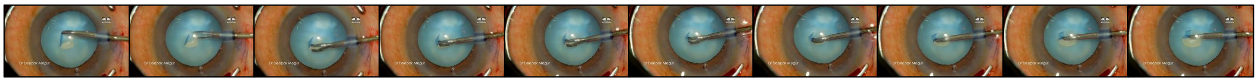
This video describes the incision sealing phase of a surgical procedure. Minimal hydration is used to ensure the incision is appropriately sealed without inducing prolonged astigmatism. It confirms the lens is in good position with overlap of the optic before sealing the incision to complete the procedure.



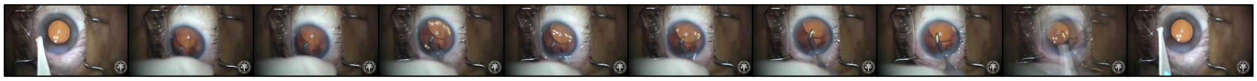
This video demonstrates the surgical steps for performing capsulorhexis in an eye procedure. The surgeon carefully evaluates the eye to determine the optimal approach for creating a continuous circular capsular opening, integral to successful cataract surgery. Specialized instruments, such as microforceps or a cystotome, are utilized to achieve precision. Understanding the anatomy of the anterior capsule and surrounding structures is crucial to avoid complications like anterior capsule tears.



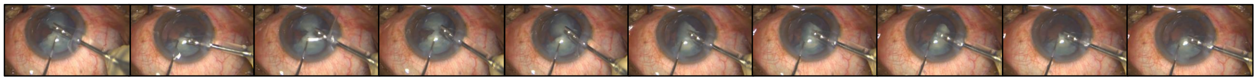
This video demonstrates a surgical procedure where the surgical phases begin with making incisions. Due to a smaller pupil size, Trypan Blue is utilized to assist in creating a 5 mm continuous curvilinear capsulorhexis.



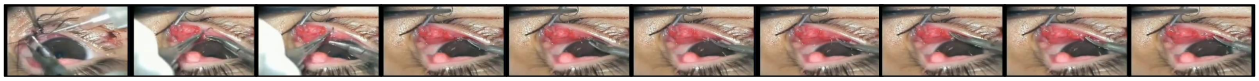
This video demonstrates the use of micro forceps through the side port during the procedure to minimize any escape of ophthalmic viscosurgical device (OVD).



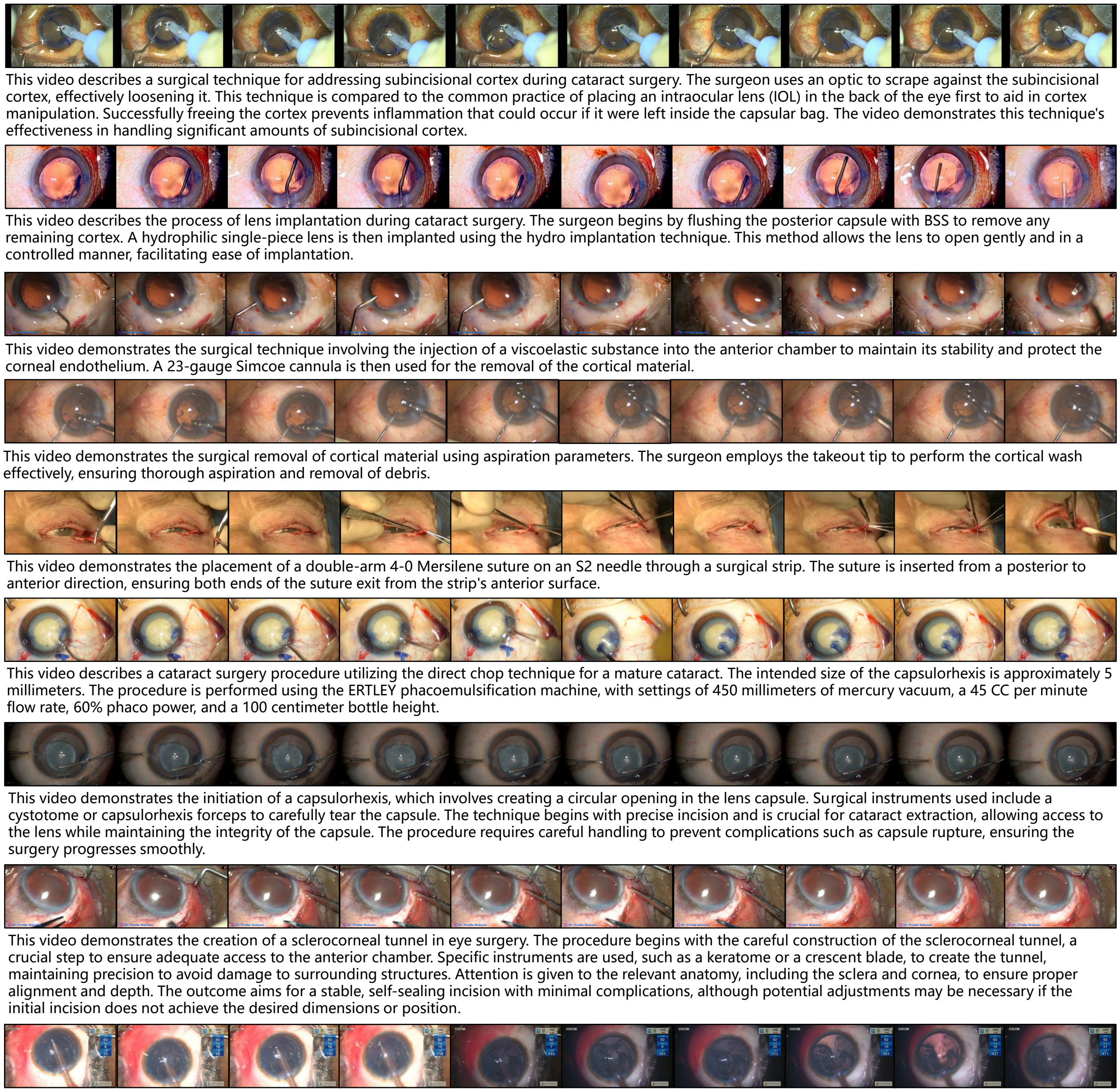
This video demonstrates a surgical procedure focusing on lens implantation. It highlights the use of instruments with specific design features, such as the intraocular instrument with a bend at the tip, to assist in rotating and positioning the lens accurately within the eye. This technique is crucial for achieving the desired anatomical alignment and optimizing surgical outcomes.



This video details a cataract surgery using phacoemulsification. Initially, a low phaco power is set at 50 percent, but less is used during the procedure. The surgeon emulsifies and aspirates the final lens fragment. Due to a small capsulorhexis, bimanual techniques are employed to effectively navigate deeper into the lens capsule.



This video demonstrates the conjunctival repair using buried sutures to prevent corneal irritation. The exposed tube end is fixated to prevent extrusion, utilizing an innovative horizontal mattress suture technique. The suture secures the tube by taking a bite through the lower portion of the tube, then directing the pull upward and medially. This ensures stability, preventing extrusion. The suture is tied over a bolster fashioned from a polythene tube. The video concludes with the closure and canalicular repair, described as highly successful.



This video describes a surgical technique for addressing subincisional cortex during cataract surgery. The surgeon uses an optic to scrape against the subincisional cortex, effectively loosening it. This technique is compared to the common practice of placing an intraocular lens (IOL) in the back of the eye first to aid in cortex manipulation. Successfully freeing the cortex prevents inflammation that could occur if it were left inside the capsular bag. The video demonstrates this technique's effectiveness in handling significant amounts of subincisional cortex.

This video describes the process of lens implantation during cataract surgery. The surgeon begins by flushing the posterior capsule with BSS to remove any remaining cortex. A hydrophilic single-piece lens is then implanted using the hydro implantation technique. This method allows the lens to open gently and in a controlled manner, facilitating ease of implantation.

This video demonstrates the surgical technique involving the injection of a viscoelastic substance into the anterior chamber to maintain its stability and protect the corneal endothelium. A 23-gauge Simcoe cannula is then used for the removal of the cortical material.

This video demonstrates the surgical removal of cortical material using aspiration parameters. The surgeon employs the takeout tip to perform the cortical wash effectively, ensuring thorough aspiration and removal of debris.

This video demonstrates the placement of a double-arm 4-0 Mersilene suture on an S2 needle through a surgical strip. The suture is inserted from a posterior to anterior direction, ensuring both ends of the suture exit from the strip's anterior surface.

This video describes a cataract surgery procedure utilizing the direct chop technique for a mature cataract. The intended size of the capsulorhexis is approximately 5 millimeters. The procedure is performed using the ERTLEY phacoemulsification machine, with settings of 450 millimeters of mercury vacuum, a 45 CC per minute flow rate, 60% phaco power, and a 100 centimeter bottle height.

This video demonstrates the initiation of a capsulorhexis, which involves creating a circular opening in the lens capsule. Surgical instruments used include a cystotome or capsulorhexis forceps to carefully tear the capsule. The technique begins with precise incision and is crucial for cataract extraction, allowing access to the lens while maintaining the integrity of the capsule. The procedure requires careful handling to prevent complications such as capsule rupture, ensuring the surgery progresses smoothly.

This video demonstrates the creation of a sclerocorneal tunnel in eye surgery. The procedure begins with the careful construction of the sclerocorneal tunnel, a crucial step to ensure adequate access to the anterior chamber. Specific instruments are used, such as a keratome or a crescent blade, to create the tunnel, maintaining precision to avoid damage to surrounding structures. Attention is given to the relevant anatomy, including the sclera and cornea, to ensure proper alignment and depth. The outcome aims for a stable, self-sealing incision with minimal complications, although potential adjustments may be necessary if the initial incision does not achieve the desired dimensions or position.

This video demonstrates the use of a coaxial irrigation and aspiration probe to aspirate the cortical bowl during surgery.

Figure 6. **Some examples of clip-text pairs from OphVL.**

ID	Phase
0	Antibiotikum
1	Hydrodissektion
2	Incision
3	Irrigation-Aspiration
4	Kapselpolishing
5	Linsenimplantation
6	Phako
7	Rhexis
8	Tonisieren
9	Visco-Absaugung
10	Viscoelasticum
11	not_initialized

Table 11. Phase labels of the Cat-21 dataset.

ID	Phase
0	Incision
1	Viscoelastic
2	Capsulorhexis
3	Hydrodissection
4	Phacoemulsification
5	Irrigation/Aspiration
6	Capsule Pulishing
7	Lens Implantation
8	Lens positioning
9	Viscoelastic_Suction
10	Anterior_Chamber Flushing
11	Tonifying/Antibiotics

Table 12. Phase labels of the Cataract-1K dataset.

ID	Phase
0	Incision
1	Viscous agent injection
2	Rhexis
3	Hydrodissection
4	Phacoemulsificiation
5	Irrigation and aspiration
6	Capsule polishing
7	Lens implant setting-up
8	Viscous agent removal
9	Tonifying and antibiotics

Table 13. Phase labels of the Cataract-101 dataset.

ID	Phase
0	Implantation
1	Irrigation_Aspiration and Visc_Suction
2	Phacoemulsification
3	Rhexis
4	Rest

Table 14. Phase labels of the Catreldet dataset.

ID	Phase
0	Linsenimplantation
1	Linsenimplantation_before
2	Linsenimplantation_after

Table 15. Phase labels of the LensID dataset.

ID	Instrument
0	spatula
1	27 gauge cannula
2	slit knife
3	phaco tip
4	capsulorhexis forceps
5	cartridge
6	I/A handpiece
7	cannula
8	katena forceps
9	eye retractors
10	angled incision knife

Table 16. Instrument labels of the CatInstSeg dataset.

ID	Instrument
0	Capsulorhexis Forceps
1	Capsulorhexis Cystotome
2	Katena Forceps
3	Irrigation-Aspiration
4	Slit Knife
5	Phacoemulsification Tip
6	Spatula
7	Gauge
8	Lens Injector
9	Incision Knife

Table 17. Instrument labels of the Cataract-1K dataset.

ID	Instrument
0	I/A Handpiece
1	Marker
2	Rycroft Cannula Handle
3	Eye Retractors
4	Cotton
5	Secondary Knife Handle
6	Surgical Tape
7	Troutman Forceps
8	Hydrodissection Cannula Handle
9	Vitrectomy Handpiece
10	Iris Hooks
11	Rycroft Cannula
12	Lens Injector
13	Secondary Knife
14	Mendez Ring
15	Primary Knife
16	Capsulorhexis Cystotome
17	I/A Handpiece Handle
18	Micromanipulator
19	Charleux Cannula
20	Phacoemulsifier Handpiece
21	Viscoelastic Cannula
22	Capsulorhexis Forceps
23	Phacoemulsifier Handpiece Handle
24	Lens Injector Handle
25	background
26	Hydrodissection Cannula
27	Capsulorhexis Cystotome Handle
28	Needle Holder
29	Suture Needle
30	Bonn Forceps
31	Primary Knife Handle

Table 18. Fine-grained instrument labels of the CaDIS dataset (CaDIS-F).

ID	Instrument
0	I/A Handpiece
1	Cap. Forceps
2	Eye Retractors
3	Lens Injector
4	Tissue Forceps
5	Surgical Tape
6	Ph. Handpiece
7	Cannula
8	Secondary Knife
9	Cap. Cystotome
10	Primary Knife
11	Micromanipulator

Table 19. Coarse-grained instrument labels of the CaDIS dataset (CaDIS-C).