OPHTH-500K: CURATING UNSTRUCTURED VIDEO STREAMS FOR SCALING MULTIMODAL OPHTHALMOLOGY FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale vision-language learning in ophthalmology remains constrained by the limited availability of multimodal data with clinically descriptive supervision. We introduce Ophth-500K, a large-scale multimodal ophthalmology dataset constructed from 14,769 hours of expert-led videos. Our automated data engine identifies clinically relevant ophthalmic content, aligns visual segments with expert narration, and converts continuous video streams into structured image, text, and audio samples. Ophth-500K contains 467,697 aligned multimodal samples spanning Optical Coherence Tomography, Color Fundus Photography, and Ultra Widefield imaging. Its source content covers 116 languages and more than 100 countries and regions, while UMLS-based analysis identifies 1,885 anatomical concepts, 689 pathologic findings, and 3,379 disease concepts. To evaluate the learning value of the curated supervision, we train OphthCLIP using a standard contrastive vision-language objective. OphthCLIP substantially improves cross-modal retrieval over general-purpose, medical, and ophthalmology-specific baselines, achieving 32.6% Recall@1 for image-to-text retrieval and 33.2% for text-to-image retrieval. These results demonstrate the value of expert ophthalmology videos as a scalable source of structured multimodal supervision.

1 INTRODUCTION

Large-scale multimodal pretraining has become increasingly important in medical AI Moor et al. (2023); Tu et al. (2024); Zhang et al. (2023); Wu et al. (2025), where language supervision can provide semantic information beyond conventional categorical labels Zhang et al. (2023); Silva-Rodriguez et al. (2025); Du et al. (2024). This is particularly relevant to ophthalmology, where a single retinal image may contain multiple anatomical structures, pathological findings, and diagnostic cues that cannot be fully represented by a disease label alone Chia et al. (2024); Zhou et al. (2023); Shi et al. (2025). Large ophthalmic repositories such as INSIGHT at Moorfields and UK Biobank provide substantial imaging scale, but access is often controlled, and their supervision is not designed for large-scale image-text learning INSIGHT. (2024); Chua et al. (2019); Khan et al. (2021). Public resources are more accessible but often remain focused on specific modalities or tasks, including classification, segmentation, and synthetic image generation Gulshan et al. (2016); de Vente et al. (2024); Li et al. (2024); Shang et al. (2023). Consequently, clinically descriptive image-text resources remain scarce, limiting the development of ophthalmology-specific vision-language models that can learn richer associations between visual findings and clinical concepts Shi et al. (2025); Silva-Rodriguez et al. (2025); Da Soh et al. (2025).

A largely untapped source of such supervision already exists in expert ophthalmology videos. For instance, educational lectures, imaging demonstrations, and clinical presentations routinely show retinal images while specialists describe the corresponding anatomy, pathology, and diagnostic findings Ikezogwo et al. (2023). These videos therefore contain naturally paired visual and linguistic clinical information at a scale that is difficult to obtain through manual annotation. However, this supervision is not directly usable for model training. Relevant ophthalmic images may appear only briefly or occupy a small portion of the frame, expert narration may precede or follow the corresponding visual evidence, and the surrounding content may include presenters, slides, text overlays, or unrelated discussion Ikezogwo et al. (2023). The key challenge is therefore to recover reliable

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and structured multimodal supervision from these unstructured video streams at scale Yang et al. (2025).

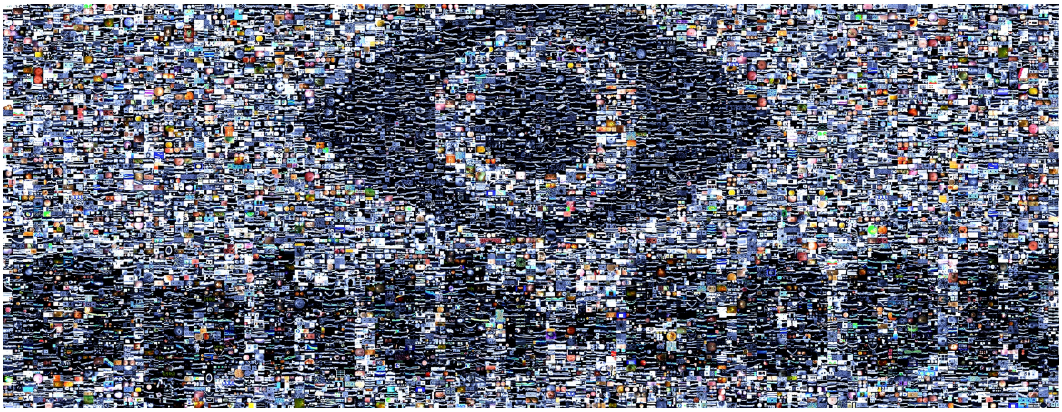

Figure 1: Overview of the Ophth-500K. The figure has been processed for privacy protection.

We address this challenge by introducing Ophth-500K, a large-scale multimodal ophthalmology dataset constructed from unstructured expert-led videos, as demonstrated in Figure 1. We develop an automated data engine that identifies clinically relevant visual content, aligns ophthalmic images with expert narration, and transforms noisy video streams into structured image, text, and audio supervision. Using this pipeline, we process 14,769 hours of video from 29,465 source videos and obtain 467,697 aligned multimodal samples. Ophth-500K covers three major ophthalmic imaging modalities, including Optical Coherence Tomography (OCT), Color Fundus Photography (CFP), and Ultra Widefield (UWF) imaging. Its source content spans 116 languages and more than 100 countries and regions. Unified Medical Language System (UMLS)-based analysis further identifies 1,885 anatomical concepts, 689 pathologic functions, and 3,379 disease concepts, demonstrating broad clinical and semantic coverage.

To evaluate the learning value of the automatically curated supervision, we train OphthCLIP with a standard contrastive vision-language objective. We evaluate cross-modal alignment on image-to-text and text-to-image retrieval tasks and compare against general-purpose, medical, and ophthalmology-specific vision-language models. OphthCLIP consistently outperforms all evaluated baselines. The best variant achieves 32.6% Recall@1 for image-to-text retrieval and 33.2% for text-to-image retrieval, while the strongest baseline achieves 7.2% and 5.2%, respectively. These results show that the curated multimodal supervision provides strong domain-specific training signals for ophthalmic vision-language learning.

Our main contributions are:

- Ophth-500K, a large-scale multimodal ophthalmology resource containing 467,697 aligned image, text, and audio samples.
- An automated data engine that recovers structured clinical supervision from unstructured expert ophthalmology videos.
- A controlled vision-language evaluation showing strong cross-modal alignment after training on Ophth-500K.

2 Related Work

Ophthalmology Datasets Existing ophthalmology datasets vary substantially in scale, accessibility, modality coverage, and annotation type Khan et al. (2021). As summarized in Table 1, large clinical repositories such as INSIGHT at Moorfields INSIGHT. (2024), AlzEye Wagner et al. (2022), I-ODA Mojab et al. (2021), and UK Biobank Chua et al. (2019) contain extensive real-world ophthalmic imaging and linked clinical information. Their scale has enabled major advances in retinal representation learning and clinical prediction, but access is typically controlled, and the available

supervision is not designed primarily for large-scale image-text pretraining. Public datasets improve accessibility, yet many remain focused on specific modalities or tasks. EyePACS Gulshan et al. (2016) and AIROGS de Vente et al. (2024) primarily support disease classification from CFP, OCTA 500 Li et al. (2024) focuses on OCT and OCTA segmentation; and SynFundus-1M Shang et al. (2023) provides large-scale synthetic fundus images with structured annotations. More recent resources such as FFA IR provide richer clinical annotations but remain centered on a specific imaging modality Li et al. (2025). Consequently, publicly available ophthalmic resources still provide limited support for large-scale multimodal learning across diverse imaging settings and clinically descriptive text.

Dataset	Scale	Modality	Instance	Access	Annotation Level
INSIGHT Moorfields INSIGHT. (2024)	31,000,000	CFP, OCT, SLO, FA, AF, etc.	Natural	Limited	Unknown
AlzEye Wagner et al. (2022)	6,000,000	OCT, etc.	Natural	Private	Unknown
I-ODA Mojab et al. (2021)	3,500,000	CFP, OCT, HVF, etc.	Natural	Private	EHR
SynFundus-1M Shang et al. (2023)	1,000,000	CFP	Synthetic	Public	Classification
IDHea Chopra et al. (2025)	870,000	CFP, OCT	Natural	Limited	EHR
UK Biobank Chua et al. (2019)	500,000	CFP, OCT, etc.	Natural	Limited	Quality Grading
OCTA-500 Li et al. (2024)	360,000	OCT, OCTA	Natural	Public	Segmentation
AIROGS de Vente et al. (2024)	110,000	CFP	Natural	Public	Classification
EyePACS Gulshan et al. (2016)	88,000	CFP	Natural	Public	Classification
Ophth-500K (Ours)	500,000	CFP, OCT, UWF	Natural	Public	Audio, Text

Table 1: Summary of Ophthalmic Datasets

Ophthalmology Foundation Models Foundation models have rapidly expanded across medical AI, from generalist multimodal systems to domain-specific models for biomedical vision-language learning, radiology, and medical image segmentation Moor et al. (2023); Tu et al. (2024); Zhang et al. (2023); Wu et al. (2025); Ma et al. (2024). Ophthalmology has followed the same transition from task-specific supervised learning toward large-scale pretraining on retinal images and multimodal clinical data Chia et al. (2024). Image-based models such as RETFound learn transferable retinal representations from large ophthalmic image collections and have demonstrated strong performance across ocular and systemic disease prediction tasks Zhou et al. (2023). Subsequent studies have shown that RETFound representations can also transfer to real-world community screening settings Zhang et al. (2024). More recent multimodal models incorporate textual supervision to connect retinal findings with clinical concepts. Ophthalmology-specific vision-language models such as EyeCLIP and FLAIR support image-text retrieval, zero-shot recognition, and representation learning from clinical descriptions Shi et al. (2025); Silva-Rodriguez et al. (2025). Recent ophthalmic AI has also explored label-efficient and zero-shot generalization to previously unseen disease patterns Pan et al. (2026). Beyond ophthalmology, BiomedCLIP, Med PaLM M, and RadFM demonstrate the value of large-scale multimodal pretraining across biomedical and radiological domains Zhang et al. (2023); Tu et al. (2024); Wu et al. (2025). Despite this progress, ophthalmology vision-language learning remains constrained by the limited availability of large and diverse image-text resources. Existing models often rely on institution-specific data, disease labels, or relatively narrow collections of clinical descriptions, which restrict the scale and diversity of multimodal supervision Khan et al. (2021); Da Soh et al. (2025); Chia et al. (2024). These limitations motivate scalable approaches for constructing ophthalmology-specific image-text resources from broader clinical and educational sources.

Web Scale Biomedical Data Curation Recent work has shown that large-scale biomedical image-text data can be curated from publicly available literature and expert-generated content. BiomedCLIP Zhang et al. (2023) and BIOMEDICA Lozano et al. (2025) mine biomedical image-text pairs from scientific literature, while Quilt 1M Ikezogwo et al. (2023) extracts supervision from expert histopathology videos. RadFM further combines web-scale radiology data with existing medical datasets for multimodal model training Wu et al. (2025). These studies demonstrate the value of public biomedical content as a scalable source of multimodal supervision. Ophth-500K extends this direction to ophthalmology, where continuous expert videos additionally require temporal alignment, clinical image extraction, and narration refinement before forming structured training samples.

3 OPHTH-500K DATA ENGINE

Our data engine converts continuous ophthalmology videos into structured multimodal training samples Ikezogwo et al. (2023); Yang et al. (2025). Each source video passes through three stages. As illustrated in Figure 2, we first identify videos that contain clinically relevant ophthalmic content. Second, we locate diagnostic visual segments and align them with nearby expert narration. Third, we extract the relevant image regions, remove sensitive information, and refine the associated narration into concise clinical descriptions. The final output is an aligned image, text, and audio sample.

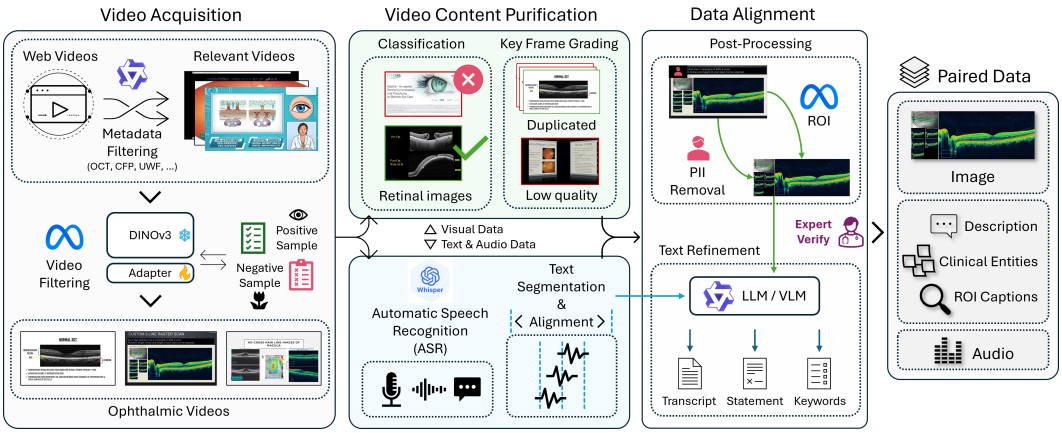

Figure 2: Overview of the Ophth-500K data construction pipeline. The pipeline first identifies videos that contain relevant ophthalmic imaging. It then locates stable clinical images and associates them with the corresponding expert narration. Finally, the visual content is refined, sensitive information is removed, and the narration is converted into a concise clinical description. Each resulting sample contains an ophthalmic image, its clinical description, and the associated audio.

3.1 OPHTHALMIC VIDEO ACQUISITION

Video Retrieval. We first construct a pool of candidate videos using ophthalmology-specific search queries. The queries cover major imaging modalities, common ophthalmic diseases, and professional sources such as clinical lectures, symposia, and imaging workshops. Video metadata, including titles, descriptions, tags, duration, and channel information, is collected for subsequent screening.

Metadata Screening. We then apply semantic filtering to remove videos with limited ophthalmic or clinical relevance. A language model examines the available metadata and retains videos that are likely to contain clinically informative content Bai et al. (2025). This stage removes general or unrelated material while preserving uncommon cases that may still provide useful ophthalmic supervision.

Clinical Image Verification. Metadata alone does not guarantee that a video contains usable ophthalmic images. We therefore perform a second screening step on sampled video frames. A classifier trained on frozen DINOv3 features identifies frames that contain retinal or other relevant ophthalmic imagery Siméoni et al. (2025). Videos with no high-confidence ophthalmic frames are discarded. The remaining videos form the source corpus for subsequent temporal alignment and multimodal sample construction.

3.2 CLINICAL CONTENT PURIFICATION AND DATA ALIGNMENT

Since a selected video may still contain long periods without useful ophthalmic content, we identify short clinical segments in which relevant ophthalmic imagery remains visually consistent.

Clinical Image Selection. We first sample frames from each video and apply the ophthalmic image classifier introduced in Section 3.1. Frames with low clinical relevance are removed, and the remaining frames serve as candidate visual anchors. We then group temporally adjacent candidate frames

according to their visual similarity. DINOv3 embeddings are extracted for the retained frames, and cosine similarity is computed between consecutive embeddings. A substantial decrease in similarity indicates a visual transition and separates two neighboring segments. Within each segment, the frame with the highest classifier confidence is selected as the representative image. Each resulting segment therefore corresponds to a temporally localized portion of the video with stable ophthalmic visual content.

Expert Narration Alignment. We next recover the narration associated with each visual segment. The audio track is transcribed with Faster-Whisper (Large-v3), and sentence-level timestamps are retained Radford et al. (2022). Expert narration does not always coincide exactly with the displayed image, since a finding may be described shortly before or after it appears on screen. We therefore collect transcript content within a temporal window around each clinical segment. The corresponding audio clip is retained together with the transcript for subsequent multimodal sample construction.

The output of this stage is a set of clinical segments, each containing a representative ophthalmic image, an aligned audio clip, and its associated transcript.

3.3 Multimodal Sample Construction

The clinical segments produced in Section 3.2 may still contain irrelevant visual regions, sensitive information, and noisy narration. We therefore refine each segment into a compact multimodal training sample with clinically relevant visual content and concise textual supervision.

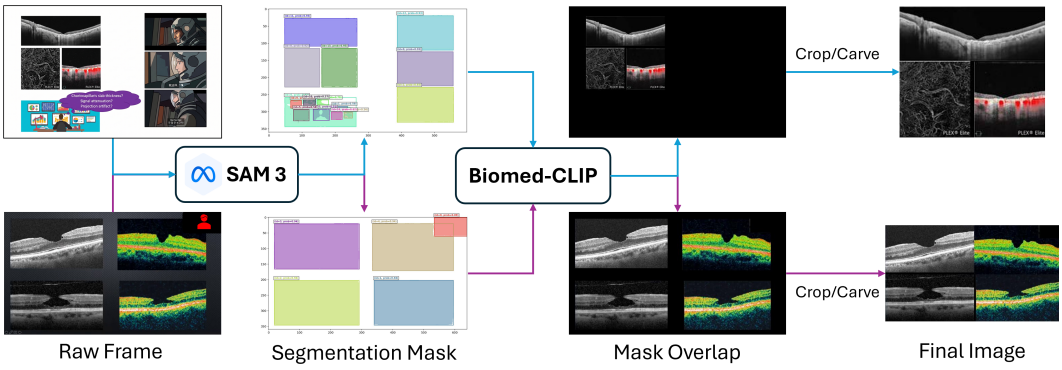

Figure 3: Clinical image extraction. Candidate regions are first identified with a segmentation model and filtered using a medical vision-language model. Irrelevant regions and empty background areas are then removed to produce the final clinical image.

Clinical Image Extraction. We first use Segment Anything 3 (SAM 3) to generate candidate ophthalmic regions from each representative frame Carion et al. (2025). Geometric filtering removes large or nested proposals that are unlikely to correspond to useful image panels. BiomedCLIP is then applied to retain regions with relevant ophthalmic content Zhang et al. (2023). If no valid region is detected, the full frame is evaluated as a fallback. Rejected regions and empty background areas are removed, producing a compact image that preserves the clinically relevant visual information, as shown in Figure 3. We next apply privacy filtering to the retained visual content. A dedicated detector identifies personally identifiable regions, including faces and patient-related information Ultralytics (2023); Kuprashevich et al. (2024), as shown in Figure 2. Detected regions are masked before the sample enters the final dataset.

Clinical Description Refinement. Finally, we refine the aligned transcript into a concise clinical description. The transcript may contain conversational language, repetition, or statements that do not describe the visible ophthalmic content. We use Qwen3 VL 8B Bai et al. (2025) to refine the aligned transcript using the associated image as visual context, converting the narration into a short declarative statement focused on visible anatomy, pathology, and diagnostic findings, as shown in Figure 4. The original aligned audio clip is retained together with the refined image and text. The final output is an aligned image, text, and audio sample used to construct Ophth-500K.



Figure 4: Selected Ophth-500K samples. Each sample contains a refined ophthalmic image, a concise clinical description derived from the associated expert narration, and extracted clinical concepts.

4 OPHTH-500K

4.1 DATASET COMPOSITION

Ophth-500K contains 467,697 aligned image, text, and audio samples extracted from 29,465 expert ophthalmology videos with a total duration of 14,769 hours, as summarized in Table 2. Each sample contains a clinically relevant ophthalmic image, a refined clinical description, and the corresponding audio segment.

The dataset covers three major ophthalmic imaging modalities. Optical Coherence Tomography accounts for 55.62% of the samples, Color Fundus Photography for 23.26%, and Ultra Widefield imaging for 21.12%. These modalities provide complementary views of retinal anatomy and pathology, including cross-sectional retinal structures, conventional fundus appearance, and peripheral retinal findings. The clinical descriptions contain an average of 14.75 words and a median of 15 words. Each sample contains an average of 3.36 UMLS-mapped clinical concepts Bodenreider (2004).

Metric	Value
Dataset Scale	
Total Videos	29,465
Total Duration (hours)	14,769
Image-Text Pairs	467,697
Clinical Description	
Mean Length (words)	14.75
Median Length (words)	15
Avg. UMLS Concepts per Sample	3.36
Modality Distribution	
OCT	55.62%
CFP	23.26%
UWF	21.12%

Table 2: Summary of Ophth-500K.

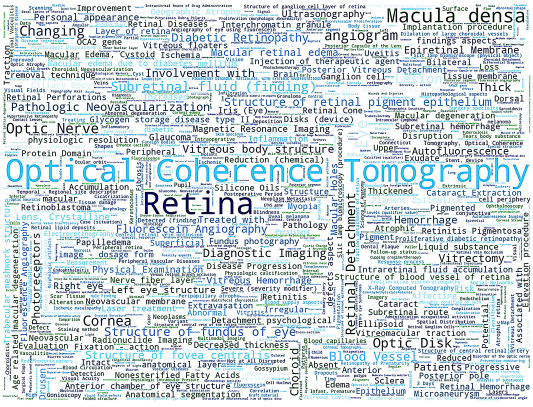

Figure 5: Top 1000 concepts.

4.2 Clinical and Source Diversity

The source videos span 116 languages and more than 100 countries and regions. India and the United States contribute the largest subsets, with additional representation from the United Kingdom, Germany, Brazil, Pakistan, and many other regions, as shown in Figure 6. The source corpus includes major linguistic groups such as Indo Aryan, Romance, and Arabic languages and covers content from Asia, Australia, North and South America, Europe, Africa, and the Middle East. Language and geographic source statistics are derived from the available video and channel metadata collected during acquisition. These statistics demonstrate broad geographic and linguistic diversity in the source content.

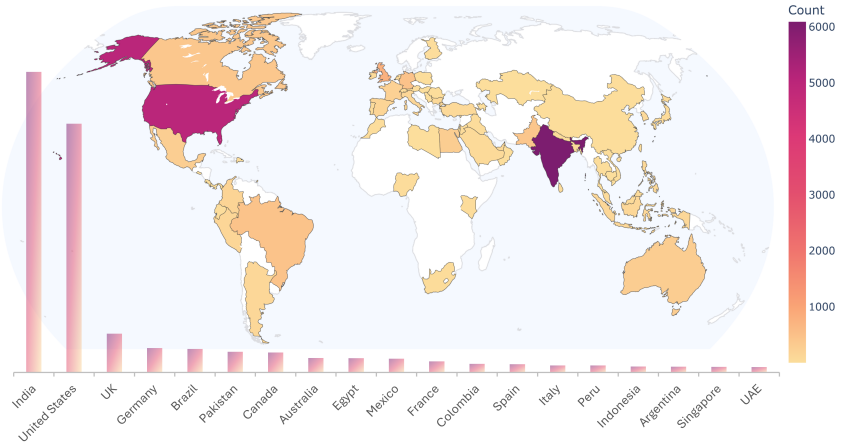

Figure 6: Geographic distributions of Ophth-500K.

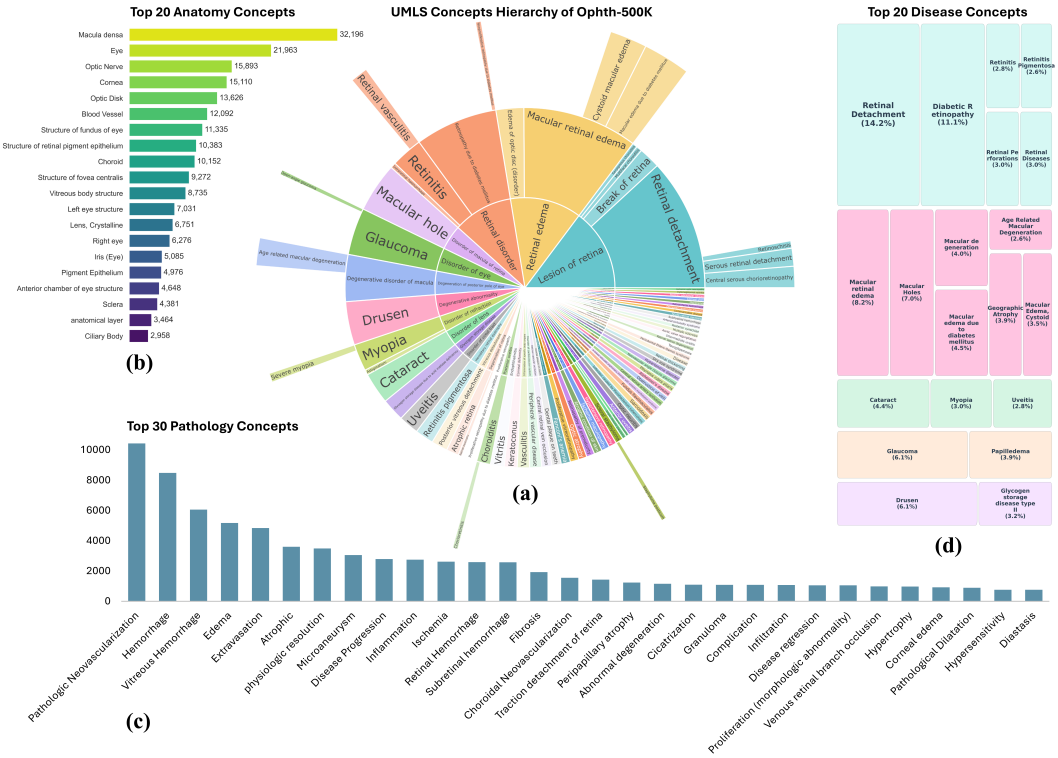

Figure 7: Medical entities analysis of Ophth-500K. (a) Hierarchy map of clinical concepts according to the UMLS dictionary Bodenreider (2004); (b) Top 20 anatomy concepts; (c) Distribution of pathologic functions and clinical mechanisms; (d) Statistics of ophthalmic diseases.

To summarize the clinical content of Ophth-500K, we use the Unified Medical Language System (UMLS) Bodenreider (2004) to standardize the medical terms appearing in the clinical descriptions. Different expressions that refer to the same clinical concept are grouped under a common Concept Unique Identifier (CUI), which reduces variation caused by synonyms and different wording across videos. We then organize the standardized concepts into three major categories, anatomy, pathology, and disease, following the UMLS and SNOMED CT Lee et al. (2014) frameworks. This organization provides a consistent basis for analyzing the clinical coverage of the dataset, as shown in Figure 7 (a).

Anatomy. Ophth-500K contains 1,885 unique anatomical concepts, with the distribution shown in Figure 7 (b). Posterior segment structures are strongly represented, including the retina, macula, optic nerve, and optic disc. The dataset also includes major anterior segment structures such as the cornea, lens, and iris. This distribution reflects broad anatomical coverage across the principal structures represented in ophthalmic imaging.

Pathology. We identify 689 unique pathologic findings, as shown in Figure 7 (c). Neovascular and hemorrhagic findings are among the most frequent, including pathologic neovascularization and multiple forms of retinal hemorrhage. Other commonly represented findings include edema, inflammation, and ischemia. Together, these concepts capture a broad range of vascular, inflammatory, and degenerative processes encountered in ophthalmic imaging.

Disease. Ophth-500K contains 3,379 unique disease concepts. As shown in Figure 7 (d), retinal and macular diseases are strongly represented, including retinal detachment, diabetic retinopathy, and macular edema. The dataset also includes optic nerve and anterior segment diseases such as glaucoma and cataract, covering a broad range of conditions commonly encountered in ophthalmic practice.

5 OPHTHCLIP

5.1 TRAINING WITH OPHTH-500K

We train OphthCLIP on Ophth-500K using the standard CLIP contrastive learning objective. Each training sample consists of an ophthalmic image I_i and its paired clinical description T_i. The image encoder f_θ and text encoder g_ϕ map the two modalities into a shared embedding space, followed by ℓ_2 normalization.

For a batch of size N, we optimize the symmetric InfoNCE loss

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{N} \left(-\log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{N} \exp(s_{ij}/\tau)} - \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^{N} \exp(s_{ji}/\tau)} \right), \tag{1}$$

where s_{ij} denotes the cosine similarity between image I_i and clinical description T_j, and τ is a learnable temperature.

We train four OphthCLIP variants with increasing visual model capacity: XS with ViT B/32, S with ViT B/16, L with ViT L/14, and H with ViT H/14. All variants use the same CLIP text encoder with a maximum context length of 77 tokens and share the same training objective. The visual backbone is the only model component that changes across the four variants.

5.2 CROSS MODAL RETRIEVAL

We evaluate OphthCLIP using image-to-text and text-to-image retrieval. Given an image or clinical description as a query, the model retrieves the most similar samples from the validation set. A retrieval is considered correct when the paired sample appears within the top K results. We report Recall@1, Recall@5, and Recall@10.

Validation Set. We construct a validation set of 500 image-text pairs from Ophth-500K. The samples are selected to preserve variation in imaging modality, clinical content, and source characteristics across the dataset. All samples originating from a selected validation video are excluded from training, ensuring separation between the training and validation sets.

Model	Config	Image-to-Text (%)			Text-to-Image (%)		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP	ViT-B/32	1.40	6.20	9.20	0.60	2.20	4.60
CLIP	ViT-B/16	4.60	10.20	14.20	0.80	3.80	6.20
CLIP	ViT-L/14	6.40	13.00	16.40	2.20	5.80	8.80
CLIP	ViT-H/14	5.80	12.40	16.80	1.80	4.40	6.40
SigLIP	ViT-SO400M/14	2.00	7.20	11.20	1.40	3.60	4.40
SigLIP	ViT-L/16	5.00	12.80	17.00	2.00	5.60	8.80
EyeCLIP	ViT-B/32	0.20	1.20	2.20	0.20	0.60	1.60
RET-CLIP	ViT-B/16	0.60	2.80	4.80	0.80	2.00	5.60
FLAIR	ResNet-50/Clinical-BERT	0.80	2.40	4.00	0.60	2.80	5.40
PMC-CLIP	ViT-L/14	3.00	10.60	16.80	2.00	9.60	13.60
BiomedCLIP	ViT-B/16	4.00	9.60	16.60	1.80	4.00	7.80
UniMed-CLIP	ViT-L/14	3.80	9.00	13.60	1.40	4.60	8.00
BMC-CLIP	ViT-L/14	7.20	18.00	24.00	5.20	9.40	13.60
OphthCLIP - XS	**ViT-B/32**	**22.20**	**43.20**	**56.20**	**20.40**	**41.20**	**56.40**
OphthCLIP - S	**ViT-B/16**	**22.60**	**46.40**	**58.00**	**21.60**	**46.20**	**57.20**
OphthCLIP - L	**ViT-L/14**	**32.60**	**60.20**	**70.20**	**33.20**	**60.40**	**70.00**
OphthCLIP - H	**ViT-H/14**	**30.60**	**55.20**	**67.00**	**31.20**	**55.20**	**65.60**

Table 3: Cross-modal retrieval results on the ophthalmic dataset. We report Recall@K (R@1, R@5, R@10) for both image-to-text and text-to-image retrieval. General, ophthalmology-specific, and medical vision-language models are compared. Our results are highlighted in bold text.

Results. We compare OphthCLIP with three groups of existing vision-language models: general-purpose models, including CLIP Radford et al. (2021) and SigLIP Zhai et al. (2023); ophthalmology-specific models, including EyeCLIP Shi et al. (2025), RET CLIP Du et al. (2024), and FLAIR Silva-Rodriguez et al. (2025); and general medical vision-language models, including PMC CLIP Lin et al. (2023), BiomedCLIP Zhang et al. (2023), UniMed CLIP Khattak et al. (2024), and BMC CLIP Lozano et al. (2025). As shown in Table 3, all OphthCLIP variants outperform the evaluated baselines across both retrieval directions and all Recall@K metrics. OphthCLIP L achieves the strongest overall performance, with an R@1 of 32.60% for image-to-text retrieval and 33.20% for text-to-image retrieval. The strongest baseline, BMC CLIP, achieves 7.20% and 5.20%, respectively. The performance gap is particularly pronounced at R@1, indicating substantially stronger correspondence between ophthalmic images and clinical descriptions after training on Ophth-500K. General medical vision-language models generally achieve higher retrieval performance than the evaluated ophthalmology-specific baselines, although substantial differences remain across individual models. Among the OphthCLIP variants, performance improves from XS to L but does not increase further with the H backbone. This result indicates that retrieval performance does not scale monotonically with visual backbone size under the current training setting.

6 Conclusion

We introduced Ophth-500K, a large-scale multimodal ophthalmology dataset containing 467,697 aligned image, text, and audio samples curated from expert ophthalmology videos. The dataset covers major ophthalmic imaging modalities and provides broad geographic, linguistic, anatomical, pathological, and disease coverage. We also presented an automated data engine that converts unstructured ophthalmology videos into structured multimodal supervision through clinical image selection, narration alignment, image refinement, privacy protection, and clinical description refinement. To evaluate the learning value of the resulting data, we trained OphthCLIP with a standard contrastive vision-language objective. OphthCLIP substantially improves cross-modal retrieval compared with general-purpose, medical, and ophthalmology-specific vision-language models. These results demonstrate that expert ophthalmology videos can serve as a scalable source of multimodal supervision for ophthalmic vision-language learning.

AI USE STATEMENT

Generative AI tools were used to assist with data cleaning and reformatting during dataset curation, including the processing and refinement of textual metadata. Generative AI tools were also used to assist with software development, literature search, and language polishing of the manuscript. All AI-assisted outputs incorporated into the dataset, code, and manuscript were reviewed and verified by the authors. Generative AI tools were not used to formulate scientific claims or draw conclusions from the experimental results. The authors take full responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

Ophth-500K was curated from publicly accessible online video sources and does not involve prospective data collection or direct interaction with human participants. During dataset construction, we designed the curation pipeline to retain ophthalmic imaging content and associated educational or clinical descriptions while excluding unrelated content and personally identifiable information. We acknowledge that large-scale curation of publicly available medical content raises considerations regarding privacy, data provenance, copyright, and potential downstream misuse. We therefore document the provenance and processing procedures of the dataset and establish release and usage conditions intended to support research use while respecting applicable source restrictions. Ophth-500K is intended for research on ophthalmic representation learning and multimodal foundation models and is not intended for direct clinical diagnosis or autonomous clinical decision-making.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the Ophth-500K data collection, filtering, modality classification, image-text pair construction, and concept extraction pipelines in the main paper and supplementary materials. Dataset statistics, processing criteria, and experimental configurations are reported to facilitate independent reproduction and evaluation. Code and the releasable components of Ophth-500K will be made publicly available upon publication, subject to the data distribution and licensing conditions described in the paper.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Mark A Chia, Fares Antaki, Yukun Zhou, Angus W Turner, Aaron Y Lee, and Pearse A Keane. Foundation models in ophthalmology. *British Journal of Ophthalmology*, 108(10):1341–1348, 2024.

Reena Chopra, Anya Guzman, Yi Sing Hsiao, Christopher Lee, Julia Jolly, Juho Uotila, Nessa Pantfoerder, Ian Douglas, Anthony P Khawaja, Pearse A Keane, et al. Datasheet for the idhea primary eye care dataset: A real-world ocular imaging resource for research. *medRxiv*, pp. 2025–04, 2025.

Sharon Yu Lin Chua, Dhanes Thomas, Naomi Allen, Andrew Lotery, Parul Desai, Praveen Patel, Zaynah Muthy, Cathie Sudlow, Tunde Peto, Peng Tee Khaw, et al. Cohort profile: design and methods in the eye and vision consortium of uk biobank. *BMJ open*, 9(2):e025077, 2019.

Zhi Da Soh, Yang Bai, Kai Yu, Yang Zhou, Xiaofeng Lei, Sahil Thakur, Zann Lee, Lee Ching Linette Phang, Qingsheng Peng, Can Can Xue, et al. An integrated language-vision foundation model for conversational diagnostics and triaging in primary eye care. *Cell Reports Medicine*, 6(12), 2025.

Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, Adrian Galdran, Miguel Ángel González Ballester, Gustavo Carneiro, R. G. Devika, Hrishikesh Panikkasseril Sethumadhavan, Densen Puthussery, Hong Liu, Zekang Yang, Satoshi Kondo, Satoshi Kasai, Edward Wang, Ashritha Durvasula, Jónathan Heras, Miguel Ángel Zapata, Teresa Araújo, Guilherme Aresta, Hrvoje Bogunović, Mustafa Arikan, Yeong Chan Lee, Hyun Bin Cho, Yoon Ho Choi, Abdul Qayyum, Imran Razzak, Bram van Ginneken, Hans G. Lemij, and Clara I. Sánchez. Airogs: Artificial intelligence for robust glaucoma screening challenge. *IEEE Transactions on Medical Imaging*, 43(1):542–557, 2024. doi: 10.1109/TMI.2023.3313786.

Jiawei Du, Jia Guo, Weihang Zhang, Shengzhu Yang, Hanruo Liu, Huiqi Li, and Ningli Wang. Retclip: A retinal image foundation model pre-trained with clinical diagnostic reports. In *International conference on medical image computing and computer-assisted intervention*, pp. 709–719. Springer, 2024.

Google. Youtube data api v3. https://developers.google.com/youtube/v3, 2026.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017, 2023.

INSIGHT. Insight dataset summaries. https://www.insight.hdrhub.org/insight-dataset-summaries, 2024. Accessed: 2026-04-16.

Saad M Khan, Xiaoxuan Liu, Siddharth Nath, Edward Korot, Livia Faes, Siegfried K Wagner, Pearse A Keane, Neil J Sebire, Matthew J Burton, and Alastair K Denniston. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*, 3(1):e51–e66, 2021.

Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. *arXiv preprint arXiv:2412.10372*, 2024.

Maksim Kuprashevich, Grigorii Alekseenko, and Irina Tolstykh. Beyond specialization: Assessing the capabilities of mllms in age and gender estimation. 2024.

Dennis Lee, Nicolette de Keizer, Francis Lau, and Ronald Cornet. Literature review of snomed ct use. *Journal of the American Medical Informatics Association*, 21(e1):e11–e19, 2014.

Mingchao Li, Kun Huang, Qiuzhuo Xu, Jiadong Yang, Yuhan Zhang, Zexuan Ji, Keren Xie, Songtao Yuan, Qinghuai Liu, and Qiang Chen. Octa-500: a retinal dataset for optical coherence tomography angiography study. *Medical image analysis*, 93:103092, 2024.

Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Tengfei Liu, Cong Wang, xin chen, zhong liu, Caineng Pan, Mengke Li, yingfeng zheng, Yizhi Liu, Flora Salim, Karin Verspoor, Xiaodan Liang, and Xiaojun Chang. Ffa-ir: Towards an explainable and reliable medical report generation benchmark. *PhysioNet*, January 2025. doi: 10.13026/f7w3-gm74. URL https://doi.org/10.13026/f7w3-gm74. Version 1.1.0.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 525–536. Springer, 2023.

Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, et al. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19724–19735, 2025.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

Nooshin Mojab, Vahid Noroozi, Abdullah Aleem, Manoj P. Nallabothula, Joseph Baker, Dimitri T. Azar, Mark Rosenblatt, R. V. Paul Chan, Darvin Yi, Philip S. Yu, and Joelle A. Hallak. I-oda, real-world multi-modal longitudinal data for ophthalmic applications. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) - HEALTHINF*, pp. 566–574. INSTICC, SciTePress, 2021. ISBN 978-989-758-490-9. doi: 10.5220/0010311405660574.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Chengchang Pan, Yudian Wang, Yixuan Jiang, Yan Su, Minwen Liao, Yao Lu, Weizhen Li, Yujing Huang, Yuexin Luo, Xuejiao Zhang, et al. A generalizable eye disease detection method based on zero-shot learning. *Communications Medicine*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Fangxin Shang, Jie Fu, Yehui Yang, Haifeng Huang, Junwei Liu, and Lei Ma. Synfundus-1m: a high-quality million-scale synthetic fundus images dataset with fifteen types of annotation. *arXiv preprint arXiv:2312.00377*, 2023.

Danli Shi, Weiyi Zhang, Jiancheng Yang, Siyu Huang, Xiaolan Chen, Pusheng Xu, Kai Jin, Shan Lin, Jin Wei, Mayinuer Yusufu, et al. A multimodal visual–language foundation model for computational ophthalmology. *npj digital medicine*, 8(1):381, 2025.

Julio Silva-Rodriguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Ultralytics. Ultralytics yolov8. https://github.com/ultralytics/ultralytics, 2023.

Siegfried Karl Wagner, Fintan Hughes, Mario Cortina-Borja, Nikolas Pontikos, Robbert Struyven, Xiaoxuan Liu, Hugh Montgomery, Daniel C Alexander, Eric Topol, Steffen Erhard Petersen, et al. Alzeye: longitudinal record-level linkage of ophthalmic imaging and hospital admissions of 353 157 patients in london, uk. *BMJ open*, 12(3):e058552, 2022.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. Synthetic continued pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Juzhao Zhang, Senlin Lin, Tianhao Cheng, Yi Xu, Lina Lu, Jiangnan He, Tao Yu, Yajun Peng, Yuejie Zhang, Haidong Zou, et al. Retfound-enhanced community-based fundus disease screening: real-world evidence and decision curve analysis. *NPJ digital medicine*, 7(1):108, 2024.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A DATA ENGINE: DETAILS, ALGORITHMS

Metadata of videos. To establish reliable data provenance prior to processing the raw video streams, we retrieve metadata for all candidate videos via a standard open-source video platform Google (2026). By extracting specific video- and channel-level attributes (e.g., textual descriptions, engagement statistics, and channel background), we construct heuristic indicators of both source authority and clinical relevance. This preliminary metadata filtering enables the exclusion of uninformative content, ensuring that the computationally intensive downstream alignment focuses exclusively on professional, expert-led materials. The core metadata fields integrated into our curation pipeline are detailed in Table 4.

Table 4: Metadata attributes of interest harvested.

Category	Attribute of Interest	Clinical & Technical Significance
Semantic Context	`title` & `description`	Provides dense semantic priors, enabling the identification of patient history and disease pathology before video processing.
	`tags`	Domain-specific identifiers used to semantically verify the presence of key clinical modalities (e.g., OCT, UWF).
	`duration`	Ensures sufficient temporal depth for extracting meaningful clinical narratives.
Quality Proxies	`viewCount` / `likeCount`	Leveraged as heuristic indicators of educational impact and community validation.
	`caption`	Prerequisite for our ASR-based distillation pipeline to ensure high-fidelity transcription.
	`license`	Validates compliance with open-access distribution and research use.
Source Authority	`channelTitle`	Facilitates the identification of professional medical societies and academic institutions.
	`subscriberCount`	Reflects the established reach and credibility of the source within the medical community.

Visual Anchor Extraction Details. We detail the core algorithmic implementation in Algorithm 1. After isolating candidate diagnostic frames via the DINOv3 backbone, we establish precise temporal visual anchors by enforcing a strict cosine similarity threshold of $\tau_s = 0.95$ between adjacent feature embeddings, grouping them into contiguous clinical segments. Finally, to ensure the downstream semantic distillation model receives sufficient auditory context without abrupt truncation, the corresponding audio tracks are explicitly extracted around these visual anchors with a temporal padding of $\delta = \pm 3.0$ seconds.

ROI Extraction Details. We provide additional details of the ROI extraction pipeline described in the main text. The full procedure is summarized in Algorithm 2, including candidate region generation using SAM3 Carion et al. (2025), geometric filtering, semantic gating with BiomedCLIP Zhang et al. (2023), and the fallback strategy when no valid regions are retained.

B OPHTHCLIP: IMPLEMENTATION DETAILS

We provide the detailed implementation configurations for our experiments as follows:

Training Configurations. Models are trained for 5 epochs using the AdamW optimizer with a learning rate of 5×10^{-6} and a weight decay of 0.05. We incorporate a linear warm-up period of 500 steps followed by a cosine learning rate decay schedule. To facilitate stable convergence, we employ mixed-precision training (amp_bfloat16). The batch size is set to 256 per GPU, with local contrastive loss enabled to leverage distributed training across 4 NVIDIA A100 (80G) GPUs.

Algorithm 1 Video Refinement and Multimodal Keyframe Extraction

Require: Video $\mathcal{V}$, thresholds τ_p, τ_s, temporal buffer δ
Ensure: Keyframes $\mathcal{K}$, aligned audio segments $\mathcal{S}_{audio}$, metadata $\mathcal{M}$

procedure PROCESSVIDEO($\mathcal{V}$)
  $\mathcal{M} \leftarrow$ extract video metadata
  extract audio track from $\mathcal{V}$
  $\mathcal{X} \leftarrow$ temporally subsample $\mathcal{V}$
  $\mathcal{F}_{valid}, \mathcal{Z}_{valid}, \mathcal{P}_{valid} \leftarrow \emptyset, \emptyset, \emptyset$
  $\mathcal{K}, \mathcal{S}_{audio} \leftarrow \emptyset, \emptyset$
  for each batch $\mathcal{X}_b \subset \mathcal{X}$ **do**
    $\mathbf{z} \leftarrow$ extract DINO features from $\mathcal{X}_b$
    $\mathbf{p} \leftarrow$ compute ophthalmic classifier scores
    $\mathcal{I} \leftarrow \{i \mid p_i > \tau_p\}$
    append $\mathcal{X}_b[\mathcal{I}], \mathbf{z}[\mathcal{I}], \mathbf{p}[\mathcal{I}]$ to $\mathcal{F}_{valid}, \mathcal{Z}_{valid}, \mathcal{P}_{valid}$
  end for
  $\mathcal{C}_{sim} \leftarrow$ cosine similarity of adjacent features in $\mathcal{Z}_{valid}$
  $\mathcal{S}_{temp} \leftarrow$ group consecutive frames where $C_{sim} \geq \tau_s$
  for each segment $\mathcal{S} \in \mathcal{S}_{temp}$ **do**
    $k^* \leftarrow \arg\max_{x_i \in \mathcal{S}} p_i$
    $\mathcal{K} \leftarrow \mathcal{K} \cup \{k^*\}$
    $(t_s, t_e) \leftarrow$ temporal boundaries of $\mathcal{S}$
    $a \leftarrow$ extract audio $[t_s - \delta,\ t_e + \delta]$
    $\mathcal{S}_{audio} \leftarrow \mathcal{S}_{audio} \cup \{a\}$
  end for
  update $\mathcal{M}$ with timestamps and segment statistics
  return $\mathcal{K}, \mathcal{S}_{audio}, \mathcal{M}$
end procedure

Algorithm 2 ROI Extraction with Semantic Gating and Spatial Compaction

Require: Representative frame I, SAM3 ϕ_{sam}, BiomedCLIP ϕ_{clip}
Ensure: Valid crops $\mathcal{C}_{valid}$, compact image I_{final}

procedure EXTRACTROI(I)
  $\mathcal{B}_{raw} \leftarrow \phi_{sam}(I)$
  $\mathcal{B}_{ref} \leftarrow$ remove large and nested boxes from $\mathcal{B}_{raw}$
  $\mathcal{C}_{crops} \leftarrow \{\text{Crop}(I, b) \mid b \in \mathcal{B}_{ref}\}$
  $M_{keep} \leftarrow \text{SemanticGate}(\phi_{clip}, \mathcal{C}_{crops})$
  if $\mathcal{B}_{ref} = \emptyset$ **or** no region in M_{keep} is retained **then**
    $m_{full} \leftarrow \text{SemanticGate}(\phi_{clip}, \{I\})$
    if m_{full} is retained **then**
      $\mathcal{C}_{valid} \leftarrow \{I\}$
      $I_{final} \leftarrow \text{compact}(I)$
    else
      return $\emptyset, \emptyset$
    end if
  else
    $\mathcal{C}_{valid} \leftarrow$ select crops retained by M_{keep}
    $I_{mask} \leftarrow$ mask rejected regions in I
    $I_{final} \leftarrow \text{compact}(I_{mask})$
  end if
  return $\mathcal{C}_{valid}, I_{final}$
end procedure