A171 (Wed-PM)

MICCAI 2025

Ophora: A Large-Scale Data-Driven Text-Guided Ophthalmic Surgical Video Generation Model

Wei Li[1,2*], Ming Hu[2,4], Guoan Wang[5], Lihao Liu[2], Kaijing Zhou[6], Junzhi Ning[2,8], Xin Guo[7], Zongyuan Ge[4], Lixu Gu[1], and Junjun He[2,3*]

[1] Shanghai Jiao Tong University, China, [2] Shanghai Artificial Intelligence Laboratory, China, [3] Shanghai Innovation Institute, China, [4] Monash University, Australia, [5] East China Normal University, China, [6] Eye Hospital, Wenzhou Medical University, China, [7] Shanghai Academy of Artificial Intelligence for Science, China, [8] Imperial College London, UK.

* liwei2022@sjtu.edu.cn *hejunjun@pjlab.org.cn

Background

In ophthalmic surgery, procedural videos serve as critical visual data for training AI systems that interpret surgical scenes and predict subsequent actions, potentially enhancing outcomes when integrated with robotics. However, acquiring large-scale, annotated surgical videos is challenging due to privacy concerns and annotation costs. To address this, **generating synthetic surgical videos based on surgeon instructions** has emerged as a promising solution, with **two major challenges**:

- Current Text-to-video (T2V) models for surgical videos rely heavily on coarse phase labels that lack fine-grained surgical detail. This limits their ability to accurately depict complex interactions between tools and anatomy.
- Existing T2V models using image-based backbones and temporal mixing layers often suffer from frame inconsistency due to insufficient modeling of spatial-temporal dynamics in surgical procedures.

Contributions

We propose a novel text-guided ophthalmic surgical video generation model, named **Ophora**, that can generate realistic and reliable ophthalmic videos following natural language instructions:

- ☐ We construct **Ophora-160K**, a large-scale ophthalmic video-instruction dataset, using a **comprehensive curation pipeline** to ensure video–instruction correspondence and visual quality;
- ☐ We propose a **progressive video-instruction tuning** strategy to adapt general T2V models for ophthalmic surgical video generation;
- ☐ We demonstrate the effectiveness of Ophora on both video realism and its utility in downstream ophthalmic surgical workflow understanding tasks.

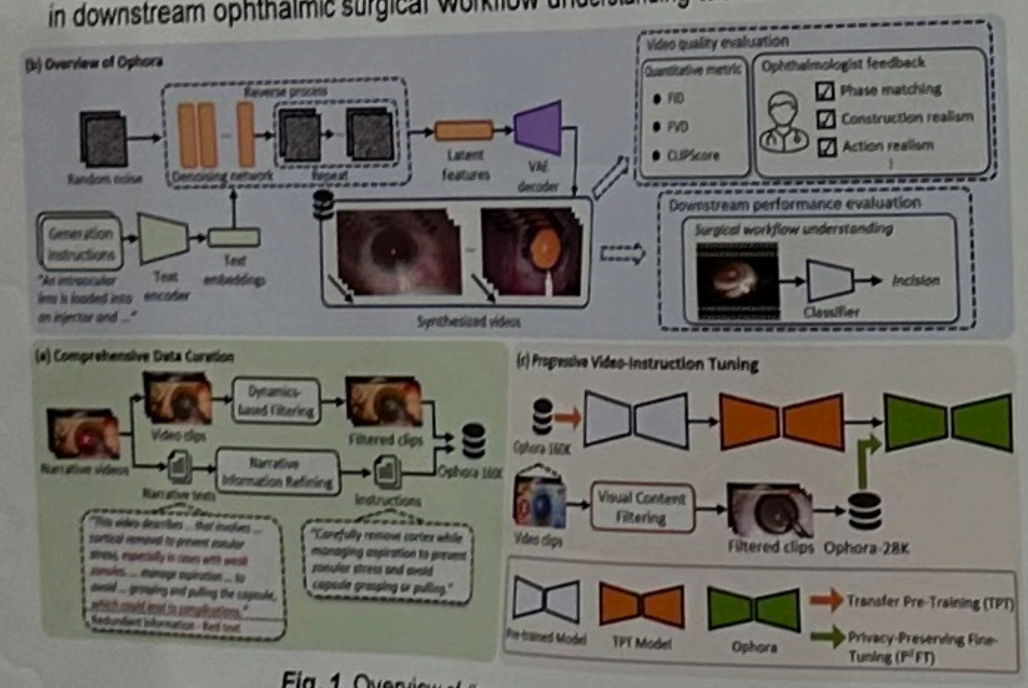

Fig. 1. Overview of the proposed Ophora.

Comprehensive Data Curation

To construct Ophora-160K, we propose a data curation pipeline that refines noisy video-caption pairs from OphVL. We employ a powerful LLM (e.g., **Qwen2.5-72B**) to remove redundant narrative content and convert captions into concise generation instructions (Narrative Information Refining).

Additionally, we filter video clips with abnormal temporal dynamics by analyzing keyframe density using PySceneDetect (Dynamics-based Filtering). Further filtering based on resolution ensures the quality of the final **160K video-instruction dataset**.

Overview of Ophora

Ophora builds upon **CogVideoX-2b**, a latent diffusion model comprising a **3D VAE**, a **T5 text encoder**, and a **transformer-based denoising network**. Video frames are compressed into latent embeddings, while input text is encoded into text embeddings. During training, Gaussian noise is added to video embeddings, and the model learns to denoise them conditioned on text. The training objective minimizes the MSE between predicted and true noise, following the standard diffusion pipeline.

$$L_{\text{diff}} = \mathbb{E}_{t,\epsilon_t,(z^v,z^\phi)\sim\mathcal{D}}\left[\|\epsilon - \epsilon_\theta([z_t^v, z^\phi], t)\|_2^2\right]$$

Progressive Video-Instruction Tuning

Ophora adopts a two-stage training strategy to preserve privacy while transferring spatial-temporal knowledge from natural videos. In the transfer pre-training stage, the model is continually trained on Ophora-160K by updating only the denoising network, while keeping the text encoder and VAE frozen. In the privacy-preserving fine-tuning stage, we detect and filter out videos with sensitive content using a large vision-language model (e.g., **Qwen2.5-VL-72B**), leading to **Ophora-28K**. This dataset is used to fine-tune the model to avoid generating sensitive videos without compromising previously learned knowledge.

Results

We evaluate the quality of synthesized videos from the Ophora.

Model	Dataset setting: OphVL [12]	Dataset setting: Ophora-160K	Metric: FID ↓	Metric: FVD ↓	Metric: CS ↑
Endora [17]			167.75	1433.29	-
Endora (w/ Ophora-160K)		✓	60.50	990.30	-
Bora [22]			138.30	1761.42	12.68
Bora (w/ Ophora-160K)		✓	49.74	604.20	32.02
CogVideoX-2b [27]			138.20	871.16	5.87
CogVideoX-2b (w/ OphVL)	✓		61.48	532.47	33.65
Ophora (TPT-only)		✓	42.16	441.09	37.03
Ophora		✓	**33.72**	**276.96**	**39.19**

Table 1. Comparison of synthesized video quality across different models based on quantitative metrics. **Bold** font denotes the best performance for each metric, and '-' indicates that CLIPScore (CS) was not calculated for this model.

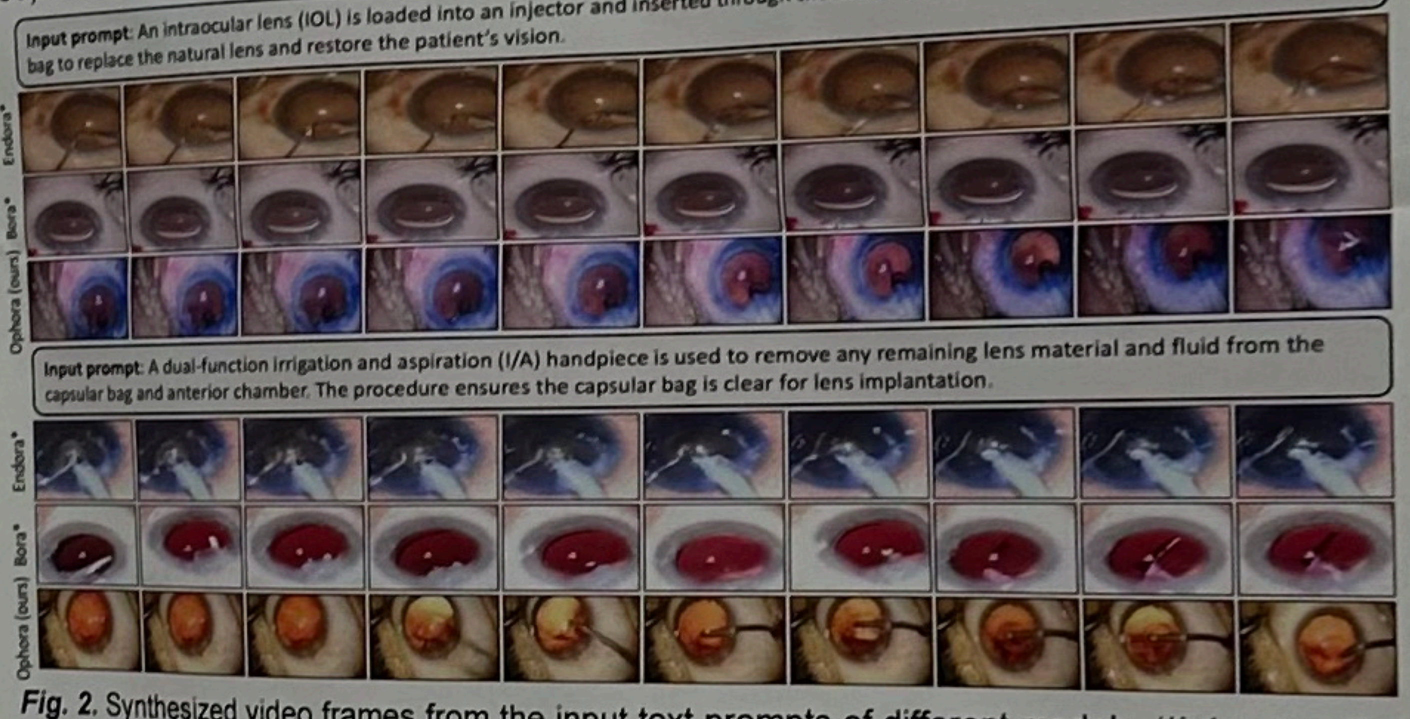

Fig. 2. Synthesized video frames from the input text prompts of different models. '*' denotes that this model was fine-tuned on the proposed Ophora-160K.

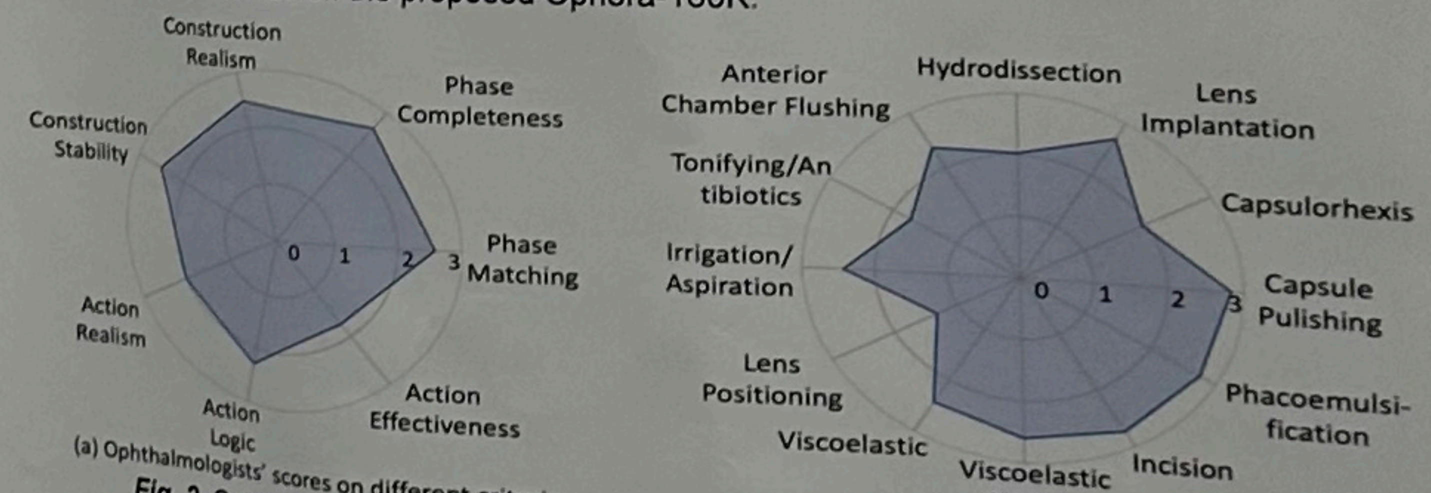

Fig. 3. Ophthalmologists' scores on different criteria (a) and surgical phases (b).

Downstream Task Performance

Training data	Classifier	Val Phase Top-1	Val Phase Top-5	Val Operation Top-1	Val Operation Top-5	Test Phase Top-1	Test Phase Top-5	Test Operation Top-1	Test Operation Top-5
Source	SlowFast [5]	34.55	70.24	26.93	65.11	37.04	72.69	27.21	67.26
	MViTv2 [18]	36.24	72.54	27.89	65.86	37.92	74.31	28.56	68.32
Source + Bora	SlowFast	36.35	72.88	28.58	67.52	39.24	74.94	28.88	69.53
	MViTv2	37.43	73.62	29.59	68.34	39.26	75.76	30.44	70.32
Source + Ophora	SlowFast	38.55	73.23	30.81	69.59	41.05	77.43	31.10	72.01
	MViTv2	**40.15**	**76.52**	**32.80**	**70.28**	**42.24**	**78.56**	**33.62**	**73.27**

Table 2. Comparison of the top-1 and top-5 accuracy of two classifiers on the validation and test sets of OphNet, including phase and operation-based classification tasks, under three training data configurations. **Bold** denotes the best performance for each split.

Please feel free to contact me!
I'm currently interning with the GMAI group at Shanghai AI Lab. Welcome to join us!

My WeChat

GMAI-github

Team Leader - Junjun He

Paper Link

Google Scholar