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
gulixu@sjtu.edu.cn
[2] Shanghai Artificial Intelligence Laboratory, China
hejunjun@pjlab.org.cn
[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

Abstract. In ophthalmic surgery, developing an AI system capable of interpreting surgical videos and predicting subsequent operations requires numerous ophthalmic surgical videos with high-quality annotations, which are difficult to collect due to privacy concerns and labor consumption. Text-guided video generation (T2V) emerges as a promising solution to overcome this issue by generating ophthalmic surgical videos based on surgeon instructions. In this paper, we present Ophora, a pioneering model that can generate ophthalmic surgical videos following natural language instructions. To construct Ophora, we first propose a Comprehensive Data Curation pipeline to convert narrative ophthalmic surgical videos into a large-scale, high-quality dataset comprising over 160K video-instruction pairs, Ophora-160K. Then, we propose a Progressive Video-Instruction Tuning scheme to transfer rich spatial-temporal knowledge from a T2V model pre-trained on natural video-text datasets for privacy-preserved ophthalmic surgical video generation based on Ophora-160K. Experiments on video quality evaluation via quantitative analysis and ophthalmologist feedback demonstrate that Ophora can generate realistic and reliable ophthalmic surgical videos based on surgeon instructions. We also validate the capability of Ophora for empowering downstream tasks of ophthalmic surgical workflow understanding. Code is available at https://github.com/mar-cry/Ophora.

Keywords: Ophthalmic Surgery · Video Generation · Transfer Learning · Instruction Tuning.

* Equal contribution
† Corresponding Author

1 Introduction

In ophthalmic surgery, surgical scenes are often recorded as videos [23]. AI systems capable of interpreting ophthalmic surgical videos and predicting subsequent operations hold the potential to improve procedural performance and reduce postoperative infections, especially when integrated with surgical robotics [25,2,3,7]. However, developing such systems requires numerous surgical videos with high-quality annotations, which are difficult to acquire due to privacy concerns [20] and labor consumption [15]. To overcome this issue, generating surgical videos based on surgeon requirements emerges as a promising solution.

Text-guided video generation (T2V) [19,27] provides a feasible method for generating ophthalmic surgical videos from natural language instructions. Current T2V work relies on surgical phase labels [4,14], which typically lack detailed descriptions for accurate surgical video generation. *As a result, these models struggle to capture fine-grained actions and intricate interactions between instruments and anatomical structures* [16]. Additionally, existing T2V research has also explored utilizing pre-trained Text-to-Image models with temporal mixing layers [17,22], which capture spatial and temporal correspondences separately. *However, this approach fails to holistically consider the complex spatial-temporal relationships inherent in surgical videos, leading to frame inconsistency* [27].

To address the first challenge, our primary motivation is to curate a large-scale, high-quality collection of surgical video-text pairs. Inspired by the properties of open-source narrative surgical videos [12,28], which provide detailed descriptions, including surgical phases, instruments, and medications from professional surgeons, we aim to collect similar videos from the internet and convert them into video-text pairs. To address the second challenge, we focus on directly modeling spatial-temporal correspondence for improved frame consistency. Inspired by the success of transfer learning [29] and acknowledging the significant scale gap between natural and surgical videos [28], we propose to transfer the spatial-temporal knowledge from a T2V model pre-trained on large-scale natural video-text datasets to guide the generation of surgical videos. During the transfer learning process, we also address the challenge of privacy preservation by ensuring that the generated videos exclude sensitive visual content, such as subtitles and watermarks, that are unrelated to the surgical process.

In this paper, 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. To achieve this, we first propose a *comprehensive data curation* pipeline to construct a large-scale, high-quality video-instruction dataset, *Ophora-160K*, which includes over 160K video clips paired with generation instructions. The pipeline involves eliminating irrelevant narrative information and filtering clips with extreme dynamics. Next, we propose a *progressive video-instruction tuning* approach to develop Ophora from a T2V model that learns spatial-temporal knowledge from natural videos. Specifically, we conduct transfer pre-training on this T2V model to transfer its knowledge for ophthalmic video generation using Ophora-160K. Then, we fine-tune the model on clips without sensitive information for preserving privacy. We conduct

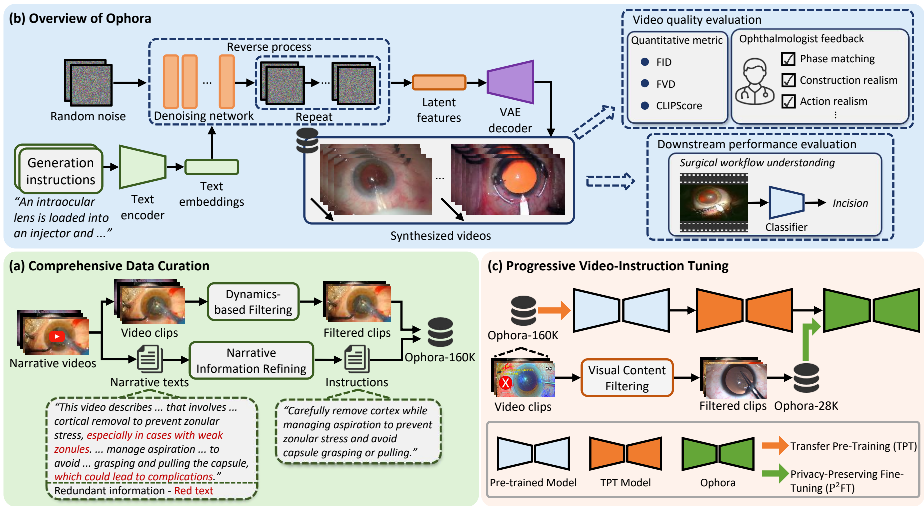

Fig. 1. Illustration of proposed Ophora that can generate ophthalmic surgical videos from instructions. Specifically, we propose a Comprehensive Data Curation pipeline to construct a large-scale, high-quality video-instruction dataset, Ophora-160K, from narrative videos (Sec. 2.1). We introduce a T2V model pre-trained on natural video-text pairs (Sec. 2.2) and leverage Progressive Video-Instruction Tuning to transfer spatial-temporal knowledge from the pre-trained model for ophthalmic video generation while preserving privacy using Ophora-160K (Sec. 2.3). We evaluate the capabilities of Ophora by assessing synthesized video quality and downstream performance.

experiments to evaluate Ophora by assessing quality of synthesized videos and potential for downstream ophthalmic surgical workflow understanding task [11].

2 Method

In this section, we first construct a large-scale, high-quality video-text dataset from narrative videos (Sec. 2.1). Then, we introduce a text-guided video generation (T2V) model pre-trained on natural video-text datasets as the generator backbone (Sec. 2.2) and apply the proposed Progressive Video-Instruction Tuning to transfer the spatial-temporal knowledge for ophthalmic video generation while preserving privacy (Sec. 2.3). Fig. 1 illustrates the overall framework.

2.1 Comprehensive Data Curation

OphVL [12] collects large-scale ophthalmic surgical narrative videos from the Internet and converts them into video clip-caption pairs. However, these captions contain redundant narrative information, e.g., educational content, that indirectly corresponds to the visual content and affects the instruction-following capabilities of a T2V model [16]. Besides, these videos exhibit variable temporal dynamics, e.g., intense camera shaking or nearly static content due to differences in surgeon experience, leading to temporal incoherence [13]. Therefore, we propose a comprehensive data curation pipeline to refine OphVL from textual and visual sides for video generation, as shown in Fig. 1(a).

Narrative Information Refining. Given the remarkable text understanding capabilities of existing large language models (LLMs), we employ a powerful open-source LLM, Qwen2.5-72B [26], to remove the irrelevant information from the captions and transform them into generation instructions. We manually select 10 captions and carefully annotate the redundant information within them as examples. Then, we leverage these examples to instruct the LLM to refine the remaining captions to enhance the accuracy and stability of the LLM's outputs.

Dynamics-based Filtering. We design a simple rule to filter clips with poor temporal dynamics quality. We utilize PySceneDetect toolkit to extract keyframes from a clip at timestamps where significant visual changes occur. Clips with an excessive or insufficient number of keyframes are filtered based on predefined thresholds. We empirically set the upper and lower thresholds to 100 and 2.

Ophora-160K. We further filter out low-resolution clips (below 720×480) and ultimately construct a large-scale video-instruction dataset, Ophora-160K, comprising approximately 160K clip-level video-instruction pairs.

2.2 Overview of Ophora

The backbone of Ophora is based on CogVideoX-2b [27], a latent diffusion model with strong capability of generating natural videos from text, shown in Fig. 1(b). It consists of a 3D Variational Autoencoder (VAE), a T5 text encoder [21], and a transformer-based denoising network. Given a video $x \in \mathbb{R}^{F\times H\times W\times C}$, where F, H, W, C denote the number of video frames, height, width, and channel number, respectively, the video is compressed into latent space using the VAE encoder. The latent representations are then patchified and flattened into D-dimensional vision embeddings z^v of length $\frac{F}{q}\cdot\frac{H}{p}\cdot\frac{W}{p}$, where q, p denote the temporal and spatial compression rates. Meanwhile, an input text is encoded into text embeddings z^ϕ via the T5 encoder. Then, random Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is added to z^v, i.e., $z_t^v = \sqrt{\bar{\alpha}_t} z^v + \sqrt{1-\bar{\alpha}_t}\epsilon$ with a noise level α_t at a random timestep $t \in [1:T]$, where T denotes the total diffusion steps, during the diffusion process [10]. Subsequently, z_t^v and z^ϕ are concatenated into a sequence and fed into the denoising network ϵ_θ to predict the Gaussian noise ϵ. The model is trained through the optimization objective of diffusion models:

$$L_{\text{diff}} = \mathbb{E}_{t,\epsilon_t,(z^v,z^\phi)\sim\mathcal{D}}\Big[\|\epsilon - \epsilon_\theta([z_t^v, z^\phi], t)\|_2^2\Big], \tag{1}$$

where $\mathcal{D}$ denotes the training dataset and $[,]$ is concatenation operation. After training, we sample random noise and use ϵ_θ in reverse process to iteratively generate latent vision embeddings from the noise given a text input. Subsequently, the vision embeddings are reconstructed to a video via the VAE decoder.

2.3 Progressive Video-Instruction Tuning

A T2V model pre-trained on large-scale natural videos provides reliable prior spatial-temporal knowledge for generating ophthalmic surgical videos. We aim to transfer such knowledge for privacy-preserved ophthalmic video generation. However, most clips in Ophora-160K contain sensitive visual information, such as subtitles or watermarks, prompting us to adopt a two-stage training approach, where we first introduce ophthalmic knowledge into the model and then guide it to generate videos without such sensitive information, as shown in Fig. 1(c).

Transfer Pre-Training. We utilize the entire Ophora-160K dataset to conduct continual pre-training on the backbone, focusing solely on training the denoising network based on the Eq. 1 while keeping T5 and the VAE frozen. We uniformly divide $[1 : T]$ into N sub-intervals when training the model on N GPUs and sample t from each sub-interval on each GPU to improve training efficiency.

Privacy-Preserving Fine-tuning.

We employ a powerful large vision-language model (LVLM), Qwen2.5-VL-72B [1], to detect whether a video contains sensitive information. We sample frames at a rate of 1 FPS between the first and last frames of each video and input them into the LVLM to detect the presence of sensitive information. We filter out videos in which at least one frame contains such sensitive information, resulting in Ophora-28K, a privacy-preserved dataset comprising over 28K video-instruction pairs. Then, we fine-tune the continual pre-trained model using Ophora-28K to enhance privacy while avoiding overwriting the previously learned spatial-temporal knowledge.

3 Experiments

Dataset. Ophora-160K contains 162,185 video clip-instruction pairs extracted from 9,819 narrative videos of ophthalmic surgery. The average duration of all clips is 5.54 seconds. For data pre-processing, each clip was resized to a resolution of 720×480 and uniformly sampled to 49 frames. Clips with fewer than 49 frames were padded with all-zero frames until they reached 49 frames. The dataset was split into 80% and 20% for training and testing, respectively. We constructed Ophora-28K, which contains 28,175 clip-instruction pairs from the training set.

Implementation. For transfer pre-training, we employ AdamW optimizer with learning rate 1×10^{-4}, batch size 128, and iteration number 65000 for sufficient knowledge transferring. For privacy-preserving fine-tuning, we employ the same optimizer with learning rate 5×10^{-5}, batch size 128, and 4500 iterations to avoid knowledge overwriting. All experiments are conducted on A100 GPUs.

Quantitative Analysis. We evaluated the quality of synthesized videos generated by our model, **Ophora**, against existing state-of-the-art surgical video generation models: **Endora** [17] and **Bora** [22] on the test set. We employ three metrics: Fréchet Inception Distance (FID) [9], Fréchet Video Distance (FVD) [24], and CLIPScore (CS) [8] to evaluate the realism and video-text consistency of the synthesized videos. The CS is calculated with the coefficient $\omega = 100$

based on OphCLIP [12], which aligns ophthalmic clips with narrative texts. We also fine-tune these models using Ophora-160K for comparison.

As presented in Table 1, Ophora achieved the best performance across all metrics compared to Endora and Bora. Ophora can generate ophthalmic videos with high visual fidelity, as measured by FID and FVD, due to inheriting rich spatial-temporal knowledge from the backbone pre-trained on large-scale natural videos, whereas the backbones of Endora and Bora lack such knowledge. Moreover, Ophora can generate videos from text, but Endora lacks this capability, as it is an unconditional video generation model. Meanwhile, the synthesized videos of Ophora achieve the highest correspondence with the input text, as they obtain the highest CS, even after these models are fine-tuned on our dataset.

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.

Model	Dataset setting		Metric		
	OphVL [12]	Ophora-160K	FID ↓	FVD ↓	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**

Visualization. We also present the synthesized video frames generated from the input text prompts in Fig. 2. As Endora cannot generate videos from text, we randomly select some synthesized videos for visualization. The synthesized videos from Ophora demonstrate realistic and detailed surgical scenes with proper instruments and coherent surgical actions that follow the input prompts. Although other models are fine-tuned on our dataset, they exhibit limited instruction-following capabilities, as evidenced by the inconsistency between the instruments and actions described in the prompt and those in the synthesized videos.

Ophthalmologist Feedback. We generated 600 videos based on the surgical phase labels of Cataract-1K dataset [6] with 50 videos for each label. These videos were synthesized based on generation instructions that were written by three ophthalmologists according to the phase labels. Then, three ophthalmologists evaluated the authenticity of these synthesized videos based on seven criteria, including: 1) **Phase Matching** that reflects the consistency between the generated video and the corresponding phase label; 2) **Phase Completeness**, which indicates whether the video presents a complete phase; 3) **Construction Realism** representing the authenticity of the anatomical structures or instruments; 4) **Construction Stability** indicating whether the anatomical

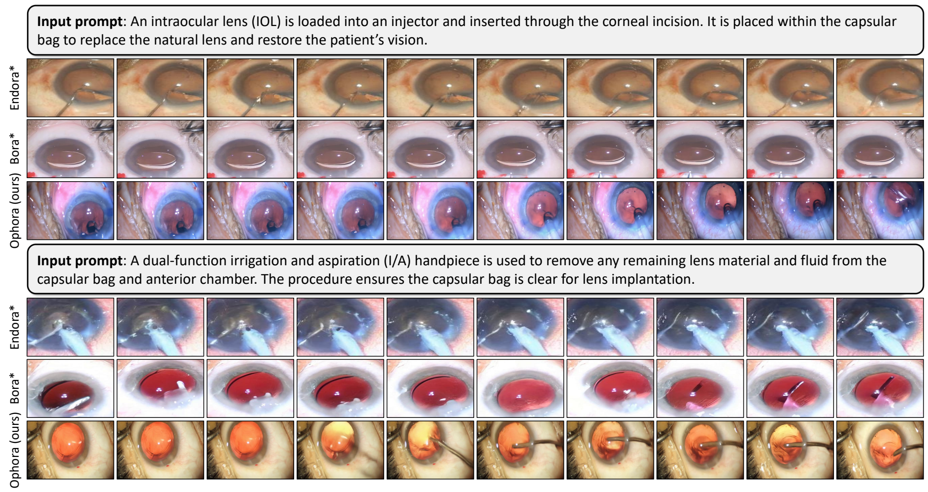

Fig. 2. Synthesized video frames from the input text prompts of different models. '*' denotes that this model was fine-tuned on the proposed Ophora-160K.

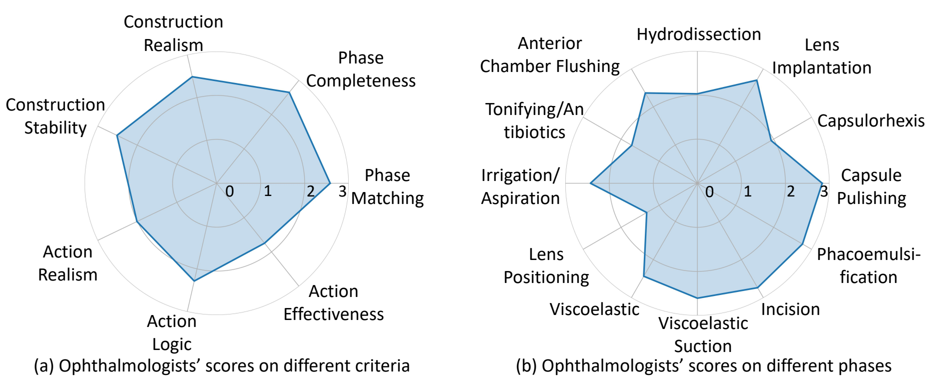

Fig. 3. Ophthalmologists' scores on different criteria (a) and surgical phases (b).

structures or instruments undergo unrealistic deformations; 5) **Action Realism**, which represents the consistency between the surgical actions in videos and actual procedures; 6) **Action Logic** reflecting the consistency of the sequential order of surgical actions with actual orders; 7) **Action Effectiveness,** meaning the realism of the deformation effects on the corresponding tissues caused by surgical actions. Each criterion is scored on a scale of 0 to 3, representing different levels of realism, where a score of 0 indicates 0% $\sim$ 10% realism, 1 corresponds to 10%$\sim$50%, 2 to 50%$\sim$90%, and 3 to 90%$\sim$100% realism, respectively.

The results are shown in Fig. 3. For most phases, Ophora can generate videos being consistent with the phases given phase-related generation instructions, with no erroneous deformation of anatomical structures or instruments and realistic actions, as evidenced by the high scores on different criteria and phases.

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.

Training data	Classifier	Val				Test			
		Phase		Operation		Phase		Operation	
		Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	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**

Data Augmentation on Downstream Task. We evaluate the impact of synthesized videos from Ophora as augmented data on OphNet [11], a large-scale benchmark for ophthalmic surgical workflow understanding, featuring 14,674 instances annotated with 52 surgical phases and 17,508 with 106 operations. We generated 100 videos for each category and compared two clip-level surgical workflow recognition models: SlowFast [5] and MViTv2 [18], under three training data configurations: 1) **source**, using only real videos from the training set of OphNet; 2) **source** + **Bora**, including real videos and synthesized videos from Bora; and 3) **source** + **Ophora**, including real videos and synthesized videos from Ophora. We employed Adam optimizer with learning rate 0.1, epoch number 35, batch size 64 for training SlowFast, and batch size 32 for training MViTv2. We uniformly sampled 8 frames for each video during training.

We report the Top-1 and Top-5 accuracy of two classifiers on the validation and test sets across different training setups for phase-based and operation-based classification tasks in Table 2. Although synthesized videos from Bora yield moderate accuracy gains compared to the source setting, leveraging the videos from Ophora achieves the best performance across all classifiers on both classification tasks. Notably, MViTv2 achieves the highest boost in phase-level Top-1 accuracy on the test set (37.92% $\rightarrow$ 42.24%) due to the adoption of synthesized videos from Ophora. The results indicate that Ophora can generate diverse and effective videos for developing an AI model for ophthalmic workflow understanding.

Ablation Study. We compared the performance of Ophora with: 1) the original CogVideoX-2b; 2) CogVideoX-2b (w/ OphVL) representing directly fine-tuning CogVideoX-2b on OphVL without applying our data curation; and 3) Ophora (TPT-only), where we only conducted transfer pre-training (TPT), as shown in Table 1. Although fine-tuning CogVideoX-2b on OphVL enables ophthalmic video generation, the performance is limited compared to Ophora due to redundant information in narratives and low temporal dynamics quality of videos. In contrast, performing TPT on the same backbone using Ophora-160K further improves performance. Surprisingly, after conducting privacy-preserving fine-tuning, Ophora achieves the best performance, as videos without sensitive information may correspond better with the narrative text.

4 Conclusion

This paper presents Ophora, a pioneering model that can generate ophthalmic surgical videos following input generation instructions. To achieve this, we propose a comprehensive data curation pipeline to convert narrative videos into a large-scale, high-quality video-instruction dataset, Ophora-160K. Then, we propose a progressive video-instruction tuning approach to transfer spatial-temporal knowledge from a T2V model pre-trained on natural video-text datasets for privacy-preserved ophthalmic video generation. Ophora can generate realistic and reliable ophthalmic videos based on user instructions, demonstrating significant potential for developing a general surgical AI system. In the future, we will explore generating videos of more surgery types with longer durations.

Acknowledgments. This work was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (2023ZD0501806), the National Key Research and Development Program (2024YFC2418201), the Interdisciplinary Fund for Medical and Engineering of SJTU (YG2023LC08), the National Key R&D Program of China (2022ZD0160101, 2022ZD0160102), the National Natural Science Foundation of China (Grant No.62272450) and Shanghai Artificial Intelligence Laboratory.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Cepolina, F., Razzoli, R.: Review of robotic surgery platforms and end effectors. Journal of Robotic Surgery **18**(1), 74 (2024)
3. Cheng, Y., Liu, L., Wang, S., Jin, Y., Schönlieb, C.B., Aviles-Rivero, A.I.: Why deep surgical models fail?: Revisiting surgical action triplet recognition through the lens of robustness. In: International Workshop on Trustworthy Machine Learning for Healthcare. pp. 177–189. Springer (2023)
4. Cho, J., Schmidgall, S., Zakka, C., Mathur, M., Kaur, D., Shad, R., Hiesinger, W.: Surgen: Text-guided diffusion model for surgical video generation. arXiv preprint arXiv:2408.14028 (2024)
5. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)

6. Ghamsarian, N., El-Shabrawi, Y., Nasirihaghighi, S., Putzgruber-Adamitsch, D., Zinkernagel, M., Wolf, S., Schoeffmann, K., Sznitman, R.: Cataract-1k dataset for deep-learning-assisted analysis of cataract surgery videos. Scientific Data **11**(1), 373 (2024)
7. He, B., de Smet, M.D., Sodhi, M., Etminan, M., Maberley, D.: A review of robotic surgical training: establishing a curriculum and credentialing process in ophthalmology. Eye **35**(12), 3192–3201 (2021)
8. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
11. 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.: Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 481–500. Springer Nature Switzerland, Cham (2025)
12. Hu, M., Yuan, K., Shen, Y., Tang, F., Xu, X., Zhou, L., Li, W., Chen, Y., Xu, Z., Peng, Z., et al.: Ophclip: Hierarchical retrieval-augmented learning for ophthalmic surgical video-language pretraining. arXiv preprint arXiv:2411.15421 (2024)
13. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21807–21818 (2024)
14. Iliash, I., Allmendinger, S., Meissen, F., Kühl, N., Rückert, D.: Interactive generation of laparoscopic videos with diffusion models. In: Mukhopadhyay, A., Oksuz, I., Engelhardt, S., Mehrof, D., Yuan, Y. (eds.) Deep Generative Models. pp. 109–118. Springer Nature Switzerland, Cham (2025)
15. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.W., Heng, P.A.: Sv-rcnet: Workflow recognition from surgical videos using recurrent convolutional network. IEEE Transactions on Medical Imaging **37**(5), 1114–1126 (2018)
16. Ju, X., Gao, Y., Zhang, Z., Yuan, Z., Wang, X., ZENG, A., Xiong, Y., Xu, Q., Shan, Y.: Miradata: A large-scale video dataset with long durations and structured captions. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 48955–48970. Curran Associates, Inc. (2024)
17. Li, C., Liu, H., Liu, Y., Feng, B.Y., Li, W., Liu, X., Chen, Z., Shao, J., Yuan, Y.: Endora: Video generation models as endoscopy simulators. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 230–240. Springer Nature Switzerland, Cham (2024)
18. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: Mvitv2: Improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4804–4814 (2022)

19. Li, Y., Min, M., Shen, D., Carlson, D., Carin, L.: Video generation from text. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018)
20. Price, W.N., Cohen, I.G.: Privacy in the age of medical big data. Nature medicine **25**(1), 37–43 (2019)
21. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1) (Jan 2020)
22. Sun, W., You, X., Zheng, R., Yuan, Z., Li, X., He, L., Li, Q., Sun, L.: Bora: Biomedical generalist video generation model. arXiv preprint arXiv:2407.08944 (2024)
23. Thia, B.C., Wong, N.J., Sheth, S.J.: Video recording in ophthalmic surgery. Survey of Ophthalmology **64**(4), 570–578 (2019)
24. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
25. Varghese, C., Harrison, E.M., O'Grady, G., Topol, E.J.: Artificial intelligence in surgery. Nature Medicine **30**(5), 1257–1268 (2024)
26. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al.: Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115 (2024)
27. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
28. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In: Linguraru, M.G., Dou, Q., Feragen, A., Giannarou, S., Glocker, B., Lekadir, K., Schnabel, J.A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 306–316. Springer Nature Switzerland, Cham (2024)
29. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE **109**(1), 43–76 (2021)