How far is AI from surgeons for providing an expert analysis of surgical videos? A comprehensive evaluation of current methods.

Authors: Nikhil Sharma, B.Tech.[1,2], Shameema Sikder, M.D.[3,4], Vishal M. Patel, Ph.D.[5], *S. Swaroop Vedula, M.B.B.S., M.P.H., Ph.D.[4], *Daniel Khashabi, Ph.D.[1,2]
* Authors contributed equally to this work.

Affiliations:

[1] Center for Language and Speech Processing, Whiting School of Engineering, Johns Hopkins University

[2] Department of Computer Science, Whiting School of Engineering, Johns Hopkins University

[3] Wilmer Eye Institute, School of Medicine, Johns Hopkins University

[4] Malone Center for Engineering in Healthcare, Whiting School of Engineering, Johns Hopkins University

[5] Department of Electrical and Computer Engineering, Whiting School of Engineering, Johns Hopkins University

Corresponding Author:
S. Swaroop Vedula
340 Malone Hall
3400 N. Charles Street
Baltimore, MD 21218
Email: swaroop@jhu.edu

Word Count: 3196

Acknowledgments: This research was supported by a grant from the National Institutes of Health, USA; R01EY033065. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The dataset used in this research was created with a grant from the Wilmer Eye Institute Pooled Professor's Fund and an unrestricted research grant from Research to Prevent Blindness to the Wilmer Eye Institute.

Abstract

Background

Analysis of surgical videos allows multiple applications such as performance feedback, coaching, video-based assessments, and autonomous intraoperative support for surgeons. Our objective was to evaluate narrative analyses of surgical videos generated by artificial intelligence (AI) systems that use vision language and large language models (VLMs and LLMs).

Methods

We used 19 videos of capsulorhexis, a critical step in cataract surgery. An expert surgeon provided ground truth video analyses using a think-aloud protocol. We developed an eight-level framework for systematic evaluation of model-generated video analyses. We used the Generative Pre-trained Transformer (GPT)-4 model to evaluate whether a given video analysis contained information to answer 102 questions relevant for surgeons. Our primary metric was the proportion of questions that could be answered with a given video analysis. We evaluated narrative video analyses generated by three VLMs and the effect of larger models, contextualized prompts, and bridging the VLM-generated analysis using an LLM.

Results

None of the models we evaluated had generated video analyses comparable to those by the human expert. Increasing model size, contextualized prompting, and bridging with an LLM improved the quality of the video analysis. However, the improvements were limited to low-level abilities such as elementary surgical facts (the lowest level in our evaluation framework).

Conclusion

Our proposed evaluation framework shows that narrative analyses of surgical videos generated by currently available AI models contain only information on elementary facts and have substantial shortcomings compared with analyses by a human expert.

Background

Surgeons continuously acquire technical skills as they practice. While surgeons acquire skills through experience, their patients are at high risk for severe complications[1]. In cataract surgery, each year of independent practice is associated with a 10% (95% confidence interval [CI] 6% to 13%) reduction in the risk of an adverse event[2]. In other words, patients receiving care from surgeons who are early in their careers are not necessarily assured of the best outcomes. These findings from the literature suggest that accelerating skill acquisition among practicing surgeons can improve patient outcomes.

To acquire skill, surgeons rely upon feedback and guidance from experienced teachers. Such feedback and guidance are available to surgeons during training, albeit inconsistently[3-7], but not when they are in independent practice[8]. Furthermore, teachers vary in their effectiveness even though it remains unmeasured[3]. Technologies that provide feedback and guidance to surgeons on a large scale can accelerate surgeons' skill acquisition and assure the best outcomes to all patients[9], and address a priority identified by a multidisciplinary Delphi panel[10]. A critical element of such technologies is to map what is visible in surgical videos to a narrative analysis. The analysis should be sophisticated and interpretable to be useful for feedback.

Artificial intelligence (AI) methods offer an opportunity to map surgical videos to human interpretable narrative analyses. Several deep learning-based foundation models are now available to analyze images and videos of human activities as well as natural language[11]. A foundation model is typically trained on large amounts of data to learn generic representations that can be repurposed for multiple downstream applications[12]. The former, called vision language models (VLMs)[13], take images or videos as input and generate descriptions about them in natural language. The latter, called large language models (LLMs), are trained to generate natural language suitable for a given context[14].

Our objective was two-fold: (1) develop an evaluation framework that provides a granular and reproducible assessment of the AI systems' ability to predict narrative analyses from surgical videos, thereby incentivizing the participation of the broader community; and (2) comprehensively evaluate narrative analyses generated from surgical videos by AI systems that use VLMs and LLMs, which we expect will be useful for expert feedback and guidance to surgeons.

Methods

This study was approved by the Johns Hopkins Institutional Review Board. We used a convenience sample of 19 deidentified videos of capsulorhexis, which is a critical step in cataract surgery[15,16]. We captured the videos from the operating microscope and

processed them to a resolution of 740*480 and 59 frames per second. An experienced faculty surgeon watched each video and provided a narrative analysis using a think-aloud protocol[17]. We instructed the surgeon to provide an explanatory and critical narration about the surgical activity observed in the video. We did not constrain the surgeon from discussing any topic. The narration lasted the duration of the video or five minutes, whichever was longer. We transcribed the narration using YouTube's automatic transcription service and manually corrected it. These transcribed expert analyses are the ground truth (reference standard) against which we evaluated model-generated analyses. One expert surgeon also assessed each video using the International Council of Ophthalmology's Ophthalmology Surgical Competency Assessment Rubric for phacoemulsification (ICO-OSCAR:phaco)[18].

Quantifying the gap in generative surgical video analysis: We aimed to quantify the gap between AI systems and the narrative analysis provided by expert surgeons (Figure 1). We developed a reproducible framework that provides granular evaluation of surgical video analysis. Specifically, our evaluation rates the AI system's abilities for generating narrative analysis in terms of eight levels of increasing complexity (Figure 2). Low-level abilities involve understanding basic concepts about surgery such as the name of surgery, the target organ, or identifying objects in the surgical field. On the other hand, high-level abilities refer to explanations about the rationale for surgical activity, for example, rotating the cystotome when inserting it into the eye to avoid injuring the wound. Details of our evaluation framework are described below in the subsection on model evaluation.

Research questions (RQs): We addressed the following RQs to evaluate narrative analyses generated from surgical videos by AI systems (Figure 1 and Supplement Figure S1):

- **RQ1. Granularity of AI analysis:** Is there a substantial gap in performance of the AI systems for low-level abilities and high-level video analysis?
- **RQ2. Scaling models:** How does scaling the VLMs (more parameters or pre-training data) affect the quality of the generated captions in a low resource domain like cataract surgery?
- **RQ3. Prompting:** Does contextualizing prompts improve the quality of VLM-generated captions?
- **RQ4. Bridging:** Does a collaborative use of LLM and VLM improve the quality and granularity of the predicted narrative analysis to show a higher-level understanding of the surgery?

Models: In our analysis, we used three VLMs: OpenFlamingo[19], BLIP-2[20], and MiniGPT-4[21]. All these architectures are based on transformers [22] and have achieved state-of-the-art results on various existing benchmarks. These models are different (and complementary) in several aspects: (a) The [pre-]training datasets used: OpenFlamingo is trained on LAION-2B[23] and Multimodal C4[24], MiniGPT-4 uses LAION[23], Conceptual Captions[25] and SBU

Captioned Photo Dataset[26], BLIP-2 is trained on various datasets including COCO[27], Visual Genome[28], CC3M[29], CC12M[30], SBU[26] and LAION400M[31]. (b) The vision encoder: While both OpenFlamingo and BLIP-2 employ CLIP[32] embeddings, OpenGPT4 employs EVA-CLIP[33] (cite). (c) The choice of LLMs used to verbalize the language analysis: OpenFlamingo uses MPT-1B for the 3B version, RedPajama-3B for 4B version, MPT-7B for 9B version, BLIP-2 is based on OPT[34], MiniGPT4 is based on Vicuna[35] and LLaMA2[36]. (d) The fusion module for aligning vision and language embeddings: OpenFlamingo employs cross attention modules attached to levels of frozen language model, BLIP-2 uses Querying Transformer (Q-Former), and OpenGPT4 uses only one projection level. (e) Finally, these models are of different sizes: OpenFlamingo model sizes range are 3B, 4B and 9B parameters, BLIP-2 comes in two size 2.7B and 6.7B parameters, MiniGPT-4 comes in 7B or 13B parameters sizes.

Evaluating RQ1: To assess whether there is a substantial performance gap between various low-level abilities and high-level video analysis, we formulate a fine-grained evaluation metric as discussed below.

Evaluating RQ2: We descriptively analyzed the models as a function of their scale (number of parameters).

Evaluating RQ3: To prompt the VLMs, we used three strategies: (1) Prompt 1: “Create a concise narration describing the events in the image.” This prompt is not specifically engineered for cataract surgery. Thus, it allows us to compare acuity of different models to identify details about the surgical procedure. (2) Prompt 2: "Create a concise narration describing the events in the cataract surgery image.” This prompt provides the models that the context is cataract surgery, but not specifically about capsulorhexis. It allows us to evaluate potential “hallucinations”[37] in the LLM-generated language. (3) Prompt 3: "Create a concise narration describing the events in the Rhexis phase of cataract surgery image.” This prompt offers specific context for the VLMs and represents the benchmark for our dataset.

Evaluating RQ4: The goal of bridging with a second LLM is to align the VLM-generated language with that from the expert surgeon. VLMs can generate low-level descriptive captions about images, but they do not provide well-reasoned captions. VLMs are not trained with the overall context about surgery or the specific procedure. On the other hand, LLMs are trained with large corpora of language in multiple domains. Consequently, they have more context than the VLMs. Bridging is useful when the VLMs produce accurate factual details of each image in the video that can be mapped to the ground truth expert analysis.

To bridge VLM-generated analyses closer to the ground truth, we used GPT-4 (32K context window)[38] by prompting it with image-level analyses from the VLM in a one-shot approach.

In the one-shot approach, we provided GPT-4 with one pair of image-level analyses and the corresponding ground truth expert analysis for one video as an example.

Evaluation metric and statistical analysis: Our primary metric to evaluate model-generated analyses is based on their relevance to the downstream task of expert training and guidance, and formative evaluation. We expected that a model-generated analysis should contain information to correctly answer questions relevant to surgeons. Accordingly, we created questions about capsulorhexis in consultation with an expert surgeon (Table S1). The questions correspond to feedback commonly provided to surgeons learning to master capsulorhexis. A video analysis that contains information to answer the questions will provide surgeons with expert training and guidance. We categorized the questions created by the expert surgeon into eight levels of increasing sophisticated reasoning necessary to answer them (Figure 2). Our primary metric was the proportion of questions in our framework that can be answered using model-generated analysis (averaged over different question levels and surgery videos) and its 95% confidence interval (CI). We qualitatively compared the primary metric for the best model in each family.

Verifying accuracy of GPT-4 in computing our primary metric: We used GPT-4 to evaluate video analyses provided by the human expert and those generated by the models. Specifically, we relied on GPT-4 to ascertain whether a given video analysis contained information to answer the questions pertaining to our framework. This allowed us to compute the proportion of questions in our framework that could be answered using a given video analysis. To verify accuracy of GPT-4 in computing our primary metric, we generated answers to 33 questions in our framework using ground truth expert analyses for all 19 videos. We prompted GPT-4 to specify when it cannot answer a question using a given video analysis. An expert surgeon evaluated the responses generated by GPT-4 on a 4-point scale (very incorrect, somewhat incorrect, somewhat correct, very correct). We collapsed the 4-point scale into a binary variable (incorrect or correct) and estimated accuracy, sensitivity, and specificity as measures of GPT-4s utility for evaluating model-generated analyses.

We also evaluated the model-generated analyses using conventional lexical and semantic metrics. Lexical metrics evaluate the token overlap between model-generated and ground truth analyses. Semantic metrics evaluate the semantic similarity between the model-generated and ground truth analyses. We computed one lexical metric (BLEU score [39]), and five semantic metrics (ROUGE-1[40], ROUGE-2[40], ROUGE-L[40], BLEURT[41], and the BERTScore[42]). However, these conventional metrics of text similarity are not enough when the generated text is long and contains nuanced analysis.

We used Python 3.9.18 for all our analyses. We used OpenAI's API to implement the GPT-4 (8K) model with a temperature of 0.9 and top-p of 1 for computing our primary metric, and GPT-4 (32K-0613) model with a temperature of 1 and top-p of 1 for bridging.

Results

The mean (standard deviation) of the duration of videos in our sample was 171.2 seconds (142.2 seconds). 11 of 19 videos in our sample were assessed to have a score of 5 on at least one of the two items for capsulorhexis and a score of at least 4 on the other item (Supplement Table S2). The 102 questions across the levels in our evaluation framework are shown in Supplement Table S1.

Results of verifying accuracy of GPT-4 in computing our primary metric: In our evaluation of GPT-4's utility for evaluating model-generated analyses, GPT-4 was 90.5% accurate (95% CI = 88.2% to 92.8%), with a sensitivity of 84.98% (95% CI = 81.4% to 88.6%) and specificity of 96% (95% CI = 93.5% to 98.4%).

Our findings below are based on the primary metric (Supplement Table S3). The estimates for the conventional metrics are shown in Supplement Table S4.

RQ1. Granularity of AI-generated video analysis: Overall, none of the models surpassed the expert surgeon. The metric decreased with increasing sophistication of questions in higher levels of our framework (Supplement Figure S2).

RQ2. Scaling models: Figure 3 and Table S5 show improvement in estimates of our primary metric with increasing model size. The average improvement was up to 100% for OpenFlamingo models (19.94 vs. 40.17), 11.1% for BLIP2 models (3.24 vs. 3.64), and 14.8% for MiniGPT4 models (38.23 vs. 44.8). When comparing models in the MiniGPT4 family with different architectures, the average improvement was 28.95% (1.21 vs.1.71).

RQ3. Prompting: Figure 4 and Table S6 show estimates of the primary metric for three strategies of prompting the VLMs. For the MiniGPT-4 Vicuna (13B) model, which was the best overall model in the MiniGPT-4 family (Supplement Figure S2), the primary metric improved when the prompt included specific context for questions in Level 0, but not in the other levels in our framework. Models in the OpenFlamingo and BLIP2 families returned no text when we used Prompt 3, which included the specific word "Rhexis".

RQ4. Bridging: Figure 5 and Table S7 show estimates of the primary metric for models with and without bridging to enrich the VLM-generated analysis. Bridging improved the metric in all cases except one (OpenFlamingo, Level 1). In fact, bridging improved the metric for Levels 4 to 7 beyond what we observe with increasing model size or prompting. We observed an improvement of 0.00 vs. 8.33 for OpenFlamingo (9B), 0.00 vs. 14.35 for BLIP2 (6.7B), and 0.00 vs. 12.50 for MiniGPT-4 Vicuna (13B) in Level 4 to Level 7.

Discussion

Our evaluation of narrative analyses generated by AI systems from videos of capsulorhexis shows a substantial gap compared with analyses by an expert surgeon. To systematically measure the gap, we devised a new framework that describes the ability of an AI system to generate narrative analyses of surgical videos. Using this framework, our evaluation shows that the model-generated analyses contained relevant elementary surgical facts but not information needed for sophisticated reasoning. The quality of the model-generated analyses improved with increasing scale of the VLMs, contextualization of the prompts to the VLMs, and collaborative use of LLMs to bridge the VLM-generated analysis with the ground truth.

Our findings show that conventional metrics used to assess text similarity are not useful to evaluate the surgical video analyses generated by AI systems (Supplement Table S4). The metrics were not discriminative of the different models or approaches we evaluated. On the other hand, the metric we proposed shows the gap between the model-generated analyses and the ground truth in terms of abilities relevant for downstream applications. While it is possible to manually review and evaluate the model-generated analyses, we used GPT-4 to predict whether the analyses contain information to answer pre-specified questions. Even though we do not evaluate the correctness of the answer, GPT-4 had 96% specificity and it can be used on a large scale. For example, our findings suggest that GPT-4 may be used to evaluate model-generated analyses of capsulorhexis videos in large numbers even without ground truth responses to the questions.

Our observation that the ground truth video analyses by the expert surgeon did not have a perfect score supports validity of the eight-level evaluation framework. We asked for a think-aloud narrative from the surgeon with no specifications about its contents. The resulting narratives left out low-level information such as individual surgical actions. Instead, they emphasized high-level information that appeared to be relevant for feedback and assessment.

The clinical scope of our work is limited to capsulorhexis. While the limited scope constrains our ability to comprehensively evaluate AI models, it allows us to understand their abilities and limitations with specificity. Our evaluation approach only verifies whether a question can be answered by GPT-4 using a given narrative analysis, but not the validity of the answer. We expect evaluating the validity of the answers would scale down the overall performance of the models and show the need for more progress with models. Furthermore, evaluating the validity will require manual annotations and obtaining them for large video datasets is costly. We did not evaluate all AI models available now for generating language from images instead of videos because the latter are still rudimentary and computationally expensive. Instead, we chose model families that allows us to evaluate the effects of model size. Because we used image to language models, we did not

include any questions for Level 3 (Localization). Localization requires temporally aligned video and language data, which was not a feature of our dataset.

We also did not use the GPT-4 vision model because of privacy concerns at the time this research was conducted (preventing use of our data for other purposes). We used a small sample of videos and ground truth narrative analyses by one expert surgeon. The collection of data is onerous and the protocols for how to capture ground truth video analyses from human experts is still not informed by relevant research. So, we conducted a pilot study with a convenience sample. It may be possible to improve the output from the VLMs through more extensive prompting. Our approach represents a baseline.

Despite the limitations, our work to evaluate generation of sophisticated and interpretable analyses addresses a novel question and enables multiple downstream applications of surgical videos such as providing performance feedback for surgeons[1,8-10]. While the idea of using videos to provide feedback to surgeons is at least several decades old[43], enabling it with AI is the innovation we addressed in this work. Besides performance feedback, video analyses in different formats can be useful to explain video-based assessments[10], support surgical coaching[8], automate documentation[44], and to create surgical copilot technologies that provide live intraoperative support[44]. The application determines the content necessary in a video analysis, which in turn influences what ground truth data are required to train the generative AI models.

The eight-level framework we proposed to evaluate narrative analyses of surgical procedures can facilitate their use for downstream applications. The requisite detail in information in the AI-generated video analyses is determined by the downstream application. In the surgical context, there is insufficient understanding of how video analyses are used as performance feedback for surgeons, to explain assessments, coach surgeons, or document intraoperative care. Further user-centered research is necessary to understand how granular the information in the video analyses should be to enable the applications. Our framework provides a structured and reproducible approach to evaluate granularity of information in the narrative analyses, whether it is obtained from humans watching surgical procedures live in the operating room, in videos, or generated using AI.

It Is clear from our study that a substantial research gap needs to be addressed for VLMs and LLMs to be useful for applications that rely on video analysis. The VLMs should be trained to learn features of multiple granularities from the videos. VLMs such as the one reported in Yuan, et.al., involve pretraining with narrations accompanying videos available on YouTube[45]. However, the narrations do not necessarily have the sophistication of expert video analyses that the model is expected to generate. It is unrealistic to create colossal datasets in the surgical context unlike those available for general language models. Instead, datasets that are well curated through protocols that draw upon qualitative research methods to specify systematic data collection methods are necessary to develop AI models that enable next-generation applications for surgeons.

Conclusion

AI models that generate language from vision (images and videos) hold substantial promise to enable applications such as surgical performance feedback, coaching, explanations for video-based assessments, and co-pilot technologies. Models available now fall short of the abilities required for surgical applications. Our work provides a framework and metric to quantify the gap in the quality of model-generated video analyses compared with human expert analyses, and to indicate directions for improvements in AI models to fill the gap.

References

1. Balvardi S, Kammili A, Hanson M, et.al. The association between video-based assessment of intraoperative technical performance and patient outcomes: a systematic review. Surg Endosc. 2022;36:7938–48.
2. Campbell RJ, El-Defrawy SR, Gill SS, Whitehead M, de LP Campbell E, Hooper PL, Bell CM, ten Hove M. New surgeon outcomes and the effectiveness of surgical training: a population-based cohort study. Ophthalmology. 2017;124(4):532-8.
3. Bell RH. Why Johnny cannot operate. Surgery. 2009 Oct 1;146(4):533-42.
4. Bohnen JD, Chang DC, George BC. Changing trends in operating room times between teaching and non-teaching cases: less time for learning? *J Am Coll Surg*. 2015;221(4):S49.
5. Snyder RA, Tarpley MJ, Tarpley JL, Davidson M, Brophy C, Dattilo JB. Teaching in the operating room: results of a national survey. *J Surg Educ*. 2012;69(5):643-49.
6. Butvidas LD, Anderson CI, Balogh D, Basson MD. Disparities between resident and attending surgeon perceptions of intraoperative teaching. *Am J Surg*. 2011;201(3):385-89.
7. Rose JS, Waibel BH, Schenarts PJ. Disparity between resident and faculty surgeons' perceptions of preoperative preparation, intraoperative teaching, and postoperative feedback. *J Surg Educ*. 2011;68(6):459-464.
8. Stefanidis D, Anderson-Montoya B, Higgins RV, Pimentel ME, Rowland P, Scarborough MO, Higgins D. Developing a coaching mechanism for practicing surgeons. Surgery. 2016;160(3):536-45.
9. Trehan A, Barnett-Vanes A, Carty MJ, McCulloch P, Maruthappu M. The impact of feedback of intraoperative technical performance in surgery: a systematic review. BMJ open. 2015;5(6):e006759.
10. Vedula SS, Ghazi A, Collins JW, Pugh C, Stefanidis D, Meireles O, Hung AJ, Schwaitzberg S, Levy JS, Sachdeva AK. Artificial intelligence methods and artificial intelligence-enabled metrics for surgical education: a multidisciplinary consensus. Journal of the American College of Surgeons. 2022 Jun 1;234(6):1181-92.
11. Awais M, Naseer M, Khan S, Anwer RM, Cholakkal H, Shah M, Yang MH, Khan FS. Foundational models defining a new era in vision: A survey and outlook. arXiv preprint arXiv:2307.13721. 2023 Jul 25. Accessed on April 14, 2024.
12. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. 2021 Aug 16. Accessed on April 14, 2024.

13. Zhang J, Huang J, Jin S, Lu S. Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024 Feb 26. Doi: 10.1109/TPAMI.2024.3369699. Accessed on April 14, 2024.
14. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nature medicine. 2023 Aug;29(8):1930-40.
15. Kim TS, O'Brien M, Zafar S, Hager GD, Sikder S, Vedula SS. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. International journal of computer assisted radiology and surgery. 2019 Jun;14(6):1097-105.
16. Hira S, Singh D, Kim TS, Gupta S, Hager G, Sikder S, Vedula SS. Video-based assessment of intraoperative surgical skill. International journal of computer assisted radiology and surgery. 2022 Oct;17(10):1801-11.
17. Van Someren M, Barnard YF, Sandberg J. The think aloud method: a practical approach to modelling cognitive processes. London: Academic Press. 1994.
18. Golnik KC, Beaver H, Gauba V, Lee AG, Mayorga E, Palis G, Saleh GM. Cataract surgical skill assessment. Ophthalmology. 2011;118(2):427-427.E5.
19. Awadalla A, Gao I, Gardner J, et al. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. arXiv:2308.01390 [cs.CV]. DOI: 10.48550/arXiv.2308.01390
20. Li J, Li D, Savarese S, Hoi S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs.CV]. https://doi.org/10.48550/arXiv.2301.12597.
21. Zhu D, Chen J, Shen X, et al. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592 [cs.CV]. DOI: 10.48550/arXiv.2304.10592. Accessed on April 14, 2024.
22. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv:1706.03762 [cs.CL]. https://doi.org/10.48550/arXiv.1706.03762.
23. Schuhmann C, Beaumont R, Vencu R, et al. LAION-5B: An open large-scale dataset for training next generation image-text models2. In: Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), Track on Datasets and Benchmarks. 2022. DOI: 10.48550/arXiv.2210.08402
24. Zhu W, Hessel J, Awadalla A, et al. Multimodal C4: An Open, Billion-scale Corpus of Images Interleaved with Text. arXiv preprint arXiv:2304.06939v3. DOI: 10.48550/arXiv.2304.06939. Accessed on April 14, 2024.
25. Sharma P, Ding N, Goodman S, et al. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long

Papers); 2018:2556–2565, Melbourne, Australia. Association for Computational Linguistics.
26. Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2Text: describing images using 1 million captioned photographs. In Proceedings of the 24th International Conference on Neural Information Processing Systems (pp. 1143–1151). Red Hook, NY, USA: Curran Associates Inc.
27. Lin TY, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs.CV]. Accessed on April 14, 2024.
28. Krishna R, Zhu Y, Groth O, et al. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. arXiv:1602.07332 [cs.CV]. 2016. DOI: 10.48550/arXiv.1602.07332. Accessed on April 14, 2024.
29. Sharma P, Ding N, Goodman S, et al. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2018:2556–2565. Melbourne, Australia: Association for Computational Linguistics. DOI: 10.18653/v1/P18-1238
30. Changpinyo S, Sharma P, Ding N, et al. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2021); 2021. DOI: 10.48550/arXiv.2102.08981. Accessed on April 14, 2024.
31. Schuhmann C, Vencu R, Beaumont R, et al. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. arXiv:2111.02114 [cs.CV]. https://doi.org/10.48550/arXiv.2111.02114. Accessed on April 14, 2024.
32. Radford A, Kim JW, Hallacy C, et al. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]. DOI: 10.48550/arXiv.2103.00020. Accessed on April 14, 2024.
33. Quan S, Yuxin F, Ledell W, Xinlong W, Yue C. EVA-CLIP: Improved Training Techniques for CLIP at Scale. arXiv:2303.15389 [cs.CV] https://doi.org/10.48550/arXiv.2303.15389. Accessed on April 14, 2024.
34. Zhang S, Roller S, Goyal N, et al. OPT: Open Pre-trained Transformer Language Models. arXiv:2205.01068 [cs.CL]. [Submitted on 2 May 2022 (v1), last revised 21 Jun 2022 (this version, v4)]. DOI: 10.48550/arXiv.2205.01068.
35. The Vicuna Team. Vicuna: an open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. https://lmsys.org/blog/2023-03-30-vicuna/. Accessed on April 14, 2024.
36. Touvron, H., Martin, L., Stone, K., et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]. DOI: 10.48550/arXiv.2307.09288.

37. Huang L, Yu W, Ma W, et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv preprint arXiv:2311.05232.[Submitted on 9 Nov 2023]. Available from: https://doi.org/10.48550/arXiv.2311.05232. Accessed on April 14, 2024.
38. OpenAI, Josh A, Steven A et al. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]. Submitted on 15 Mar 2023, last revised 4 Mar 2024. DOI: 10.48550/arXiv.2303.08774. Accessed on April 14, 2024.
39. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (pp. 311–318). Philadelphia, Pennsylvania: Association for Computational Linguistics. DOI: 10.3115/1073083.1073135.
40. Lin C-Y. ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. Association for Computational Linguistics. 2004:74-81.
41. Sellam T, Das D, Parikh AP. BLEURT: Learning Robust Metrics for Text Generation. arXiv preprint arXiv:2004.04696, 2020.
42. Zhang T, Kishore V, Wu F, et al. BERTScore: Evaluating Text Generation with BERT. In: Proceedings of the International Conference on Learning Representations. 2020. Available from: https://openreview.net/forum?id=SkeHuCVFDr. Accessed on April 24, 2024.
43. Goldman LI, Maier WP, Rosemond GP, Saltzman SW, Cramer LM. Teaching surgical technique by the critical review of videotaped performance--the surgical instant replay. Surgery. 1969;66(1):237-41.
44. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, Eisenmann M, Feussner H, Forestier G, Giannarou S, Hashizume M. Surgical data science for next-generation interventions. Nature Biomedical Engineering. 2017;1(9):691-6.
45. Yuan K, Srivastav V, Yu T, Lavanchy J, Mascagni P, Navab N, Padoy N. Learning multi-modal representations by watching hundreds of surgical video lectures. arXiv preprint arXiv:2307.15220. 2023 Jul 27. Accessed on April 25, 2024.

Figure Captions

Figure 1.

Title: Overview of our research questions in this study.

Caption: Our objective was to quantify the gap between a video analysis generated by artificial intelligence (AI) methods and an expert surgeon. The figure shows examples of an AI-generated and expert surgeon analysis for one clip in a video of capsulorhexis. The figure also shows examples of other ways to describe information in a surgical video clip, including naming the objects and activity, and triplets consisting of names for the instrument, the action it performs, and the target anatomy.

Figure 2.

Title: Framework to evaluate narrative analyses of surgical procedures using videos.

Caption: An eight-level framework explaining the type of information and its granularity that can be observed in narrative analyses of surgical videos produced by humans or artificial intelligence-driven models.

Figure 3.

Title: Effect of scaling complexity of models for generating analyses from videos of capsulorhexis.

Caption: Analyses generated by more complex models scored better. The differences in scores between models were apparent for low level information (Levels 0 and 1). For high level, the analyses did not have sufficient information to evaluate them. Our evaluation did not include Level 3 because we used image to language models that may not be relevant to localize analyses to specific timepoints in the video.

Figure 4.

Title: Effect of increasing context in the prompts on quality of captions generated using vision language models.

Caption: Contextual prompts improved the analysis generated by MiniGPT-4 Vicuna (13B), which was the best model overall in the MiniGPT-4 family, for Level 0 (low-level information).

Figure 5.

Title: Effect of collaborative use of a large language model to enrich analyses generated by vision-language models.

Caption: For the best vision language model in all three model families we evaluated, enriching the model-generated analysis with a large language model improved the evaluation metric in nearly all cases. The improvement appeared to be significant for information at low (Levels 0 and 1)- as well as high-levels (Levels 2 to 7). Our evaluation did not include Level 3.