npj | digital medicine

Perspective

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01649-4

Multimodal generative AI for interpreting 3D medical images and videos

Jung-Oh Lee[1], Hong-Yu Zhou[2], Tyler M. Berzin[3], Daniel K. Sodickson[4] & Pranav Rajpurkar[2] ✉

This perspective proposes adapting video-text generative AI to 3D medical imaging (CT/MRI) and medical videos (endoscopy/laparoscopy) by treating 3D images as videos. The approach leverages modern video models to analyze multiple sequences simultaneously and provide real-time AI assistance during procedures. The paper examines medical imaging's unique characteristics (synergistic information, metadata, and world model), outlines applications in automated reporting, case retrieval, and education, and addresses challenges of limited datasets, benchmarks, and specialized training.

Current unimodal AI models that interpret either text or images/videos already benefit physicians by summarizing electronic health records[1], identifying high-risk patients for cancers[2], and detecting lesions in medical images/videos[3,4]. However, vision-language generative AI, which integrates both linguistic and visual information, has immense potential to surpass these capabilities across the entire healthcare system, transcending specific medical fields. Techniques such as clinical report generation from images, visual question answering, and synthetic data generation are anticipated to significantly transform clinical workflows and educational practices. Nonetheless, these advancements are as of yet predominantly limited to 2D medical images. The application of multimodal generative models to high-diagnostic-value 3D medical imaging examinations, like CT and MRI, and video data, such as endoscopy or surgical video, remains in its early stages.

This perspective proposes approaches to revolutionize the interpretation and analysis of complex medical imaging data by leveraging the latest advancements in vision-language generative models, particularly video-text generative models. We argue that these powerful AI systems, originally designed for natural video understanding, possess untapped potential to transform how we process and interpret 3D medical images and medical videos. In this paper, "3D medical images" specifically refer to tomographic images from CT and MRI, while "medical videos" focus on gastrointestinal (GI) endoscopy and laparoscopy, excluding other types like 3D ultrasound or patient behavior recordings.

Video-Text Generative/Foundation Models

Video-text generative AI is capable of creating text and videos from inputs in either modality. Foundation models[5] within this category are trained on large multimodal datasets of videos, images, and text, making them highly versatile across various tasks. These models can generate video descriptions[6], answer questions about video content[7], and retrieve specific content within videos using textual queries[8]. Additionally, they are capable of tasks such as video generation[9] and object detection/tracking[10].

The successful integration of 2D images and text through Contrastive Language-Image Pre-training (CLIP)[11] and its application to videos[12,13] opened a new era for video-text foundation models. Contemporary video-text models can handle four or more modalities simultaneously[14], including audio and subtitles in addition to videos and text inputs[15], and can process thousands of frames at once[16]. As large language models (LLMs) gain multimodal capabilities, some video-text generative models have emerged from LLMs[17,18]. Multimodal LLMs such as OpenAI's GPT-4o[19], Google DeepMind's Gemini 1.5[20], and their successors can now perform video-related tasks, placing them within the video-text generative model category.

Labeling video data is challenging and costly, prompting the widespread adoption of self-supervised learning for training vision-text foundation models[21,22]. These self-supervised models, primarily trained with masked modeling[23] and contrastive learning[24] on extensive datasets, have outperformed supervised models trained on specific datasets[25–27]. The same self-supervised approach can be applied to 3D images[28] and videos[29] in medicine, reducing the need for expert labeling and addressing the scarcity of labeled data in the medical field.

Given structural similarities between video and 3D medical images, video-text models can be adapted to both medical videos and tomograms. However, it's important to recognize that CT and MRI differ significantly from standard videos, and even medical videos from endoscopy and laparoscopy have distinct characteristics. These unique properties must be carefully considered when applying video-text generative models to the medical context.

[1]Department of Radiology, Seoul National University Hospital, Seoul, Republic of Korea. [2]Department of Biomedical Informatics, Harvard Medical School, Boston, USA. [3]Center for Advanced Endoscopy, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, USA. [4]Center for Advanced Imaging Innovation and Research, Department of Radiology, New York University Grossman School of Medicine, New York, USA.
✉e-mail: pranav_rajpurkar@hms.harvard.edu

Unique Characteristics of 3D Medical Images and Medical Videos

Differences between the medical imaging data and standard video mainly arise from the specific technology and physics of medical imaging devices, as well as from the specific anatomical targets being imaged. These distinctions make it challenging to apply video-text generative models to tomograms and medical videos.

Data Format and Video Profile

Standard video frames are typically 8-bit, 3-channel RGB images, whereas 3D medical images are usually grayscale in DICOM format, with 12-bit or 16-bit pixel ranges[30]. The broader pixel range in medical images necessitates proper windowing to enable radiologists to accurately interpret different tissue types. For instance, CT scans require different value ranges to visualize bone and lung tissues, while MRI requires specific windowing to highlight lesions, depending on the vendor and protocol used. Although deep learning models can use raw 12/16-bit DICOM images, video models pre-trained on 8-bit color images/videos need proper preprocessing, such as windowing or normalization[31].

GI endoscopy and laparoscopy videos also differ from natural videos in their color space and magnification. They often use advanced imaging techniques like narrow-band imaging (NBI)[32,33], which enhances visibility of specific anatomical features by emphasizing certain wavelengths of light, such as those absorbed by hemoglobin, making blood vessels and tissue abnormalities more visible. In red dichromatic imaging (RDI)[34], which uses green, amber, and red light, blood appears as dark yellowish-green, improving visibility of bleeding lesions for accurate hemostasis. The resolution and magnification of endoscopy videos may vary depending on the endoscope used and its distance from the tissue surface. Certain specialized endoscopes, when placed directly against GI mucosa, can provide up to 520x magnification[35], capturing extremely fine details of mucosal and vascular patterns. This allows gastroenterologists to differentiate between normal tissue and dysplasia, potentially obviating the need for biopsies. Such extreme zooming is not typical in standard videos.

Self-multimodality and Synergistic Information

Compared to videos, 3D medical images often include additional dimensions beyond the extra spatial axis, such as pulse sequence[36] and contrast phase[37]. For example, MRI uses diverse pulse sequences like T1-weighted, T2-weighted, and diffusion-weighted images to demonstrate different tissue characteristics within the same anatomic structure. In multi-phase CT or MR examinations, multiple image stacks of the same organ are captured at different times, as the injection of exogenous contrast agents creates variations in image contrast as a function of time. Dual-source CT scans generate multiple stacks with different tube voltages. Furthermore, raw MRI or CT data can be post-processed to produce parametric maps with distinctly different information, such as signal relaxation rates or flow velocities. Consequently, 3D medical images exhibit self-multimodal properties.

Medical videos also display self-multimodality. Narrow-band imaging and red dichromatic imaging, among other techniques, can be toggled during procedures, generating different images of the same anatomy. During endoscopic or surgical interventions, clinicians may apply dyes or stains to better visualize and assess lesions, analogous to contrast phases in tomograms. Certain specialized endoscopic procedures combine multiple modalities such as ultrasound[38], and/or fluoroscopy[39], further demonstrating the self-multimodal properties of medical videos.

The sequences or phases in 3D medical images contain synergistic information, making interpretation more complex than simple video analysis. Image interpretation experts can derive significant synergistic insights by analyzing all sequences or phases together, such as identifying hemodynamic features of a lesion or making differential diagnoses based on lesion composition. This synergistic information is pivotal for accurate interpretation, as illustrated in Fig. 1.

Similarly, in medical videos, integrating data from different modalities provides critical synergistic information for diagnosis. For instance, in diagnosing subepithelial tumors in the stomach, light-sourced endoscopic videos provide details on the lesion's location and mucosal features, while endoscopic ultrasound reveals internal echo characteristics essential for distinguishing between conditions like ectopic pancreas and other subepithelial tumors.

Metadata in Interpretation

Metadata is crucial for interpreting 3D medical images and medical videos, more so than for standard video formats. In MRI, for example, the absence of metadata detailing pulse sequences or reconstructed parametric maps—such as transit time, cerebral blood flow, or cerebral blood volume—would complicate the evaluation of MR perfusion maps[40] in stroke patients. In colonoscopy, clinicians typically begin their detailed inspection during scope withdrawal, only after reaching the ileocecal valve[41]. This procedural metadata helps describe the approximate anatomic location of a colon polyp or the extent of mucosal involvement in ulcerative colitis. Metadata is also required in surgical videos, where experts may struggle to identify key structures or surgery phases from short clips. Moreover, since surgeries often involve procedures not fully captured on video, like vaginal procedures in laparoscopic hysterectomy[42], metadata on surgical techniques is indispensable for accurate video interpretation.

Metadata on patient demographics also significantly impacts the interpretation of tomograms and medical videos, especially for differential diagnosis. For instance, an anterior mediastinal mass on a chest CT may suggest thymic neoplasm or goiter in an elderly patient, whereas lymphoma or germ cell tumor is more likely in a younger patient[43]. Similarly, recommendations on the same gastroscopy findings can differ based on the prevalence of gastric cancer in different populations[44], implying the synergistic value of metadata. These examples underscore the importance of metadata in providing accurate diagnoses and interpretations for medical examinations and procedures.

World Models

A world model[45] refers to an intrinsic system learned by AI to understand and predict the dynamics of an environment. While video models offer a promising starting point, notable differences exist between world models required for interpreting standard videos and those for 3D medical imaging. These differences primarily arise from the additional dimensions: videos add a temporal axis to 2D images, while 3D medical images add a spatial axis, possibly along with other parametric axes. Figure 2 illustrates some of the key attributes characteristic of 3D medical image models: object connectivity across frames, causality between objects, and uniqueness of similar structures.

In medical videos, a key difference lies in directionality rather than an additional spatial axis. Although the endoscope's path might seem linear, it involves curvature and rotation, making orientation within the tubular GI lumen challenging. Even clinicians sometimes inject water into the lumen to pool in gravity-dependent areas, aiding orientation. In laparoscopy videos, navigating anatomical structures can be counterintuitive. For example, moving the laparoscope to the right in the video corresponds to the patient's left side when imaging the upper body, but to the right side when imaging the lower body, due to the laparoscope's insertion through the belly button. This necessitates a distinct approach to understanding orientation compared to traditional videos.

Adapting Video-Text Models to 3D Medical Imaging

Despite the fundamental differences between standard videos and tomograms, video pre-training has been applied to 3D medical imaging tasks[46–48]. However, using video-text models for complex multimodal tasks in 3D imaging, like report generation, remains underexplored. We propose leveraging video-text generative AI to improve the interpretative capability of deep learning models in 3D medical imaging.

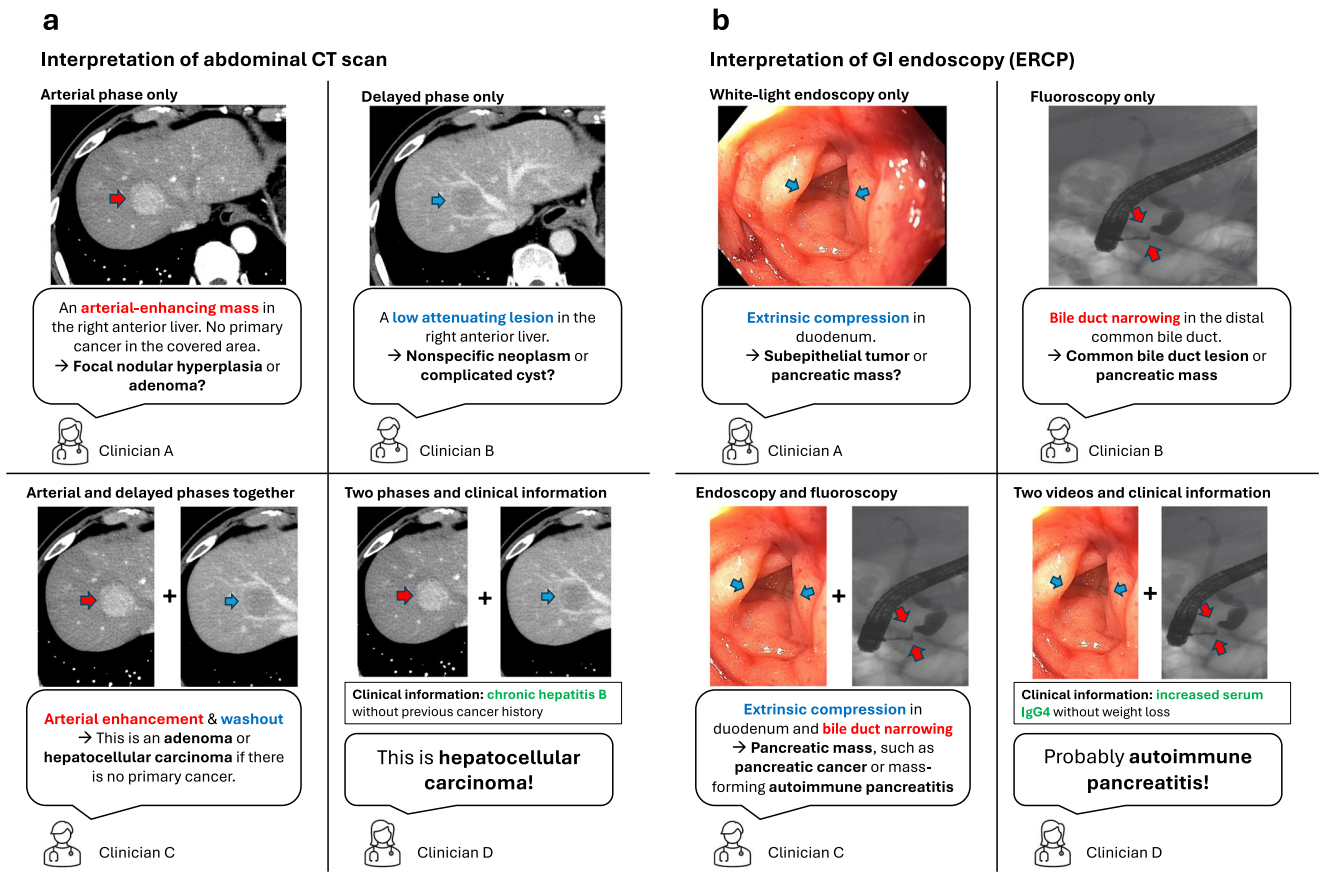

Fig. 1 | An illustrative example of the role of synergistic information in the accurate interpretation of 3D medical images and medical videos. Each panel shows how clinicians might interpret (**a**) an abdominal CT scan and (**b**) an endoscopy video (ERCP, endoscopic retrograde cholangiopancreatography) in the context of particular pieces of given information. Only when synergistic information from multiphase/multimodal images and clinical information is available (as seen with clinician D) can a clinician be confident of an accurate diagnosis.

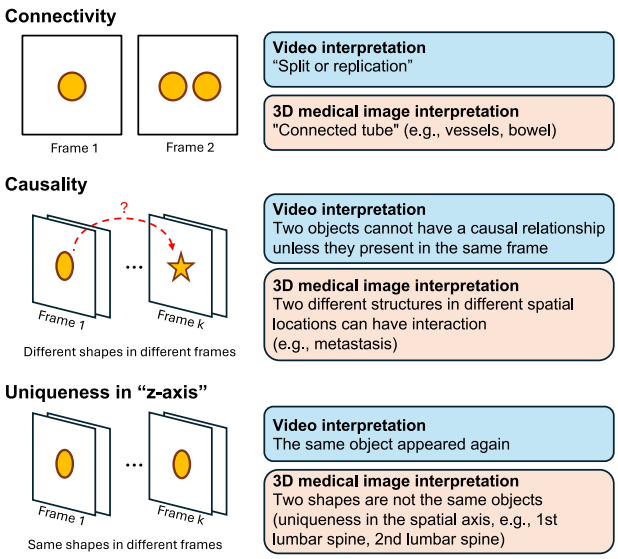

Fig. 2 | Differences in world models between videos and 3D medical imaging. (1) Connectivity: If one object becomes two identical objects in the next frame, videos may show split or replicated objects, while 3D scans capture connectivity of structures. (2) Causality: In videos, causal relationships often require co-occurrence within the same frame, but in 3D imaging, spatially separated structures can influence each other across slices. (3) Uniqueness along the z-axis: Unlike video frames, 3D volumes do not exhibit repetition of objects along the depth dimension.

Transforming 3D Medical Images into a Continuous Video

A straightforward way to adapt video-text models for 3D medical images is by converting grayscale DICOM slices to RGB and concatenating them along the "time axis" to create a long video[49]. If multiple scan windows are required due to the broader pixel intensity range of DICOM images compared to standard RGB videos, separate stacks with each window can be produced and concatenated along the time axis. Stacks of multiple sequences, contrast phases, parametric maps, or cross-sectional planes can be concatenated similarly. This method can also be applied to other types of 3D medical imaging, like PET-CT, and even multiple exams taken on different dates. By treating these concatenated slices as video frames, the dynamic capabilities of video-text models can be fully leveraged.

The key motivation for this simple concatenation method is the recent enhancement in video-text models' capacity to process large numbers of frames simultaneously. Earlier models typically handle only 8 to 16 frames at once, but advancements in training methods and cloud computing have now enabled modern video-text models to process thousands of frames concurrently[16]. These advanced models can analyze relevant findings from text input across multiple videos, even those lasting over an hour[20]. Considering that 3D medical imaging studies often comprise several hundred slices, this capability is sufficient to accommodate not only a complete individual scan but also multiple related studies simultaneously. Consequently, modern high-capacity models can compare multiple sequential studies by processing them as a continuous video or comprehensively interpret MRI alongside additional multimodal inputs such as CT and X-ray images within a single video.

Creating a long video, rather than concatenating 3D medical images channel-wise, addresses issues like scan range variations and positional

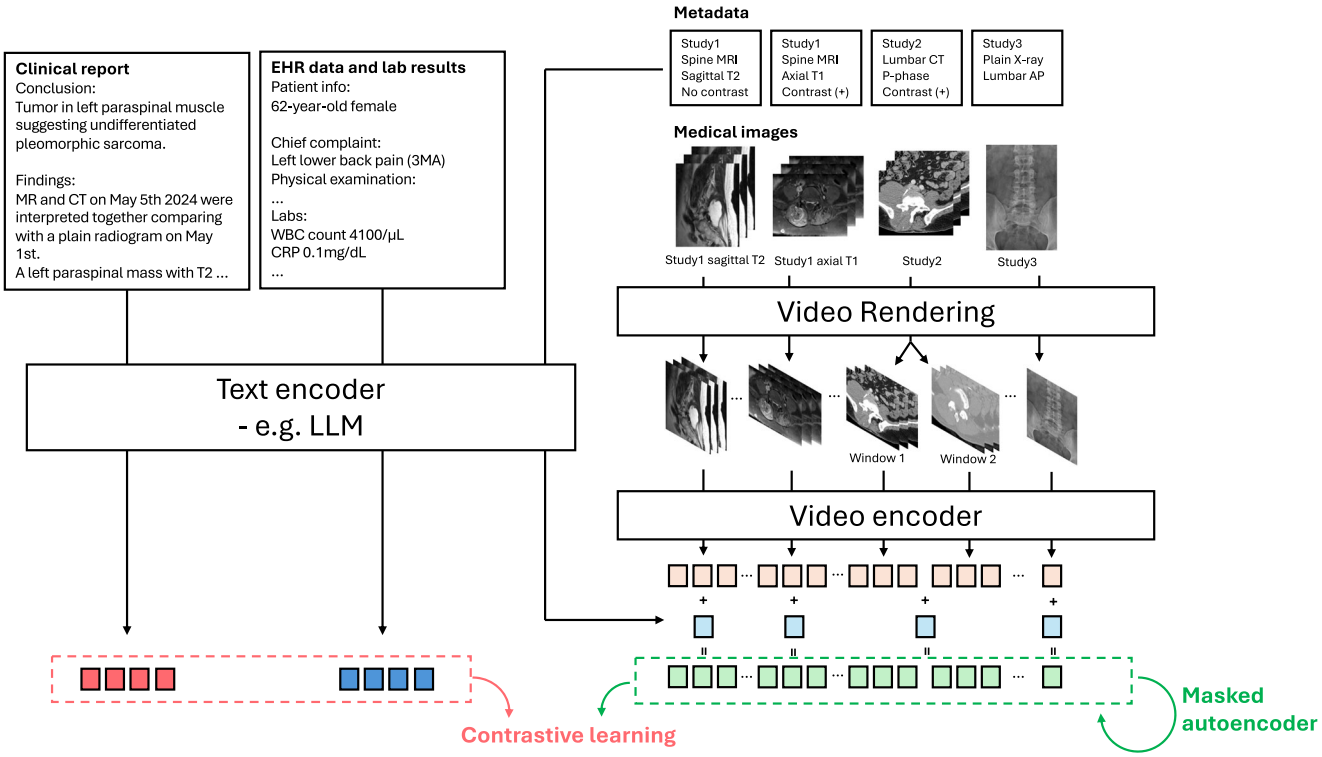

Fig. 3 | A suggested training method to adapt video-text models for 3D medical imaging. Clinical reports and electronic health records (EHR) data are input alongside multiple imaging modalities, including MRI, CT, and X-ray scans. These images can be reconfigured as a video by converting DICOM slices into RGB frames and concatenating them along the time axis, forming a long video. The video encoder processes these frames, allowing for the interpretation of complex medical imaging data. The contrastive learning approach aligns video features with text features, while the masked autoencoder aids in the training of video-text models to capture synergistic information from multimodal inputs. This method enables the comprehensive analysis of sequential studies and multimodal images within a unified framework, leveraging the enhanced capabilities of modern video-text models.

differences due to patient respiration between sequences/phases, thus enhancing the generalizability of the video-text models. For example, in liver CT scans, only the portal phase typically covers the entire abdomen to minimize radiation, while the arterial phase does not. These variations, coupled with diaphragm displacement between phases, complicate the registration of multiple sequences/phases, making channel-wise concatenation impractical.

Beyond Visual Data: Integrating Multimodal Information

To extract synergistic information and accurately interpret 3D medical images, it's crucial to simultaneously process multiple phases and sequences while also comprehending the medical context and metadata. Video-text models can integrate textual inputs from electronic health records, including clinical histories and lab findings, as well as metadata such as date, phase, and acquisition parameters about the 3D images[50]. In leveraging this extensive contextual information, video-text generative models derived from LLMs[51] can significantly benefit from a knowledge base built through massive text pretraining.

Figure 3 outlines a potential training method for video-text models integrating comprehensive information beyond traditional 3D imaging data. Unlike existing 3D medical imaging foundation models[52–54], which typically rely only on single-phase images during training, this approach incorporates multiple image sequences alongside reports, electronic health records, and relevant metadata. Specialized training strategies, such as sequence- or phase-specific masking, can be implemented to improve the model to better understand distinct imaging phases or sequences. Furthermore, applying organ-specific masking guided by segmentation models[55] across different imaging sequences can strengthen the model's capability to integrate synergistic information on anatomical structures. This refined training approach may help models effectively address the unique challenges associated with interpreting tomographic images. For optimal performance, precise alignment of multimodal and longitudinal studies with their corresponding reports is essential, ensuring the models fully leverage all available data during training.

A practical pre-training approach for video-text models in 3D medical imaging is utilizing clinical reports[54,56], which are typically paired with 3D images. However, since clinical findings in the reports often correspond to specific small volumes within the entire 3D image, using these findings as labels for the entire image can hinder accurate representation learning for video-text models. Inspired by dense captioning tasks[57], we can decompose clinical reports into dense captions targeting specific body regions or organs. Segmentation models can then localize video frames for each organ[55], providing detailed and localized training targets for video-text models. Moreover, video-text models capable of dense captioning can generate frame-by-frame annotations and provide visual grounding, which is essential for verifying model reliability.

Adapting Video-Text Models to Medical Videos

Video-text models with long video inputs are particularly well-suited for analyzing medical videos, which often contain rich temporal information. For instance, during endoscopy, observing peristaltic motion in the gastrointestinal tract can provide valuable insights into digestive function[58]. Such temporal analysis is beyond the capability of non-video models. The dynamic nature of endoscopy/laparoscopy, characterized by variable lighting and the presence of fluids, bubbles, and debris, often leads to errors in traditional AI models[59]. However, video-text models with extended inputs can potentially mitigate these issues by utilizing comprehensive information across the entire videos. Additionally, combining previous and

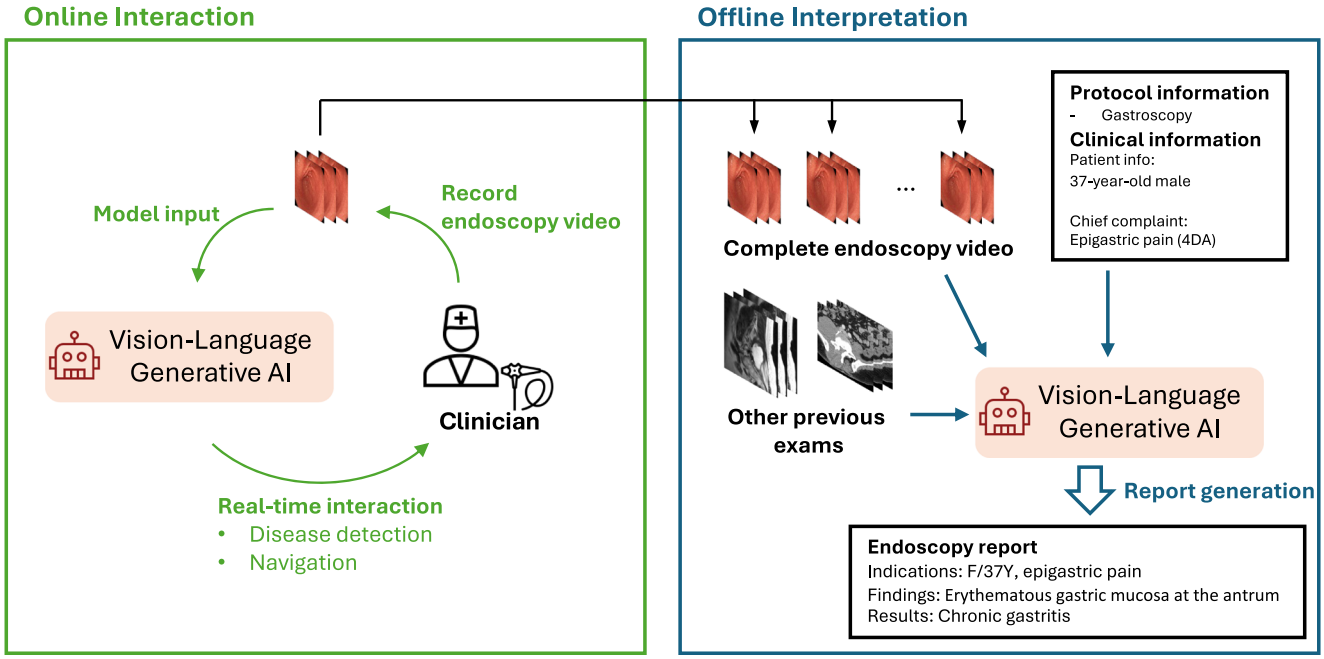

Fig. 4 | A clinical application of vision-language generative AI in endoscopy. Through real-time interaction with physicians, the vision-language AI assists in ensuring no lesions are missed. After a physician records a whole endoscopy video, the AI generates a comprehensive report that incorporates the endoscopy protocol, the patient's clinical history, and findings from previous exams such as CT and MRI scans.

current videos as long inputs may allow tracking lesion progression or assessing therapeutic interventions over time.

Treating medical videos similarly to standard videos is a natural approach, but specific preprocessing steps may be required to handle the unique characteristics of endoscopy and laparoscopy data. These steps may include color space conversion[60], image enhancement[61], and removal of specular reflections caused by the endoscope light source[62]. For labeling, a dense captioning approach[63] can be applied to process clinical reports to create captions for medical videos, focusing on specific anatomical landmarks, lesions, or abnormalities observed during the endoscopic or surgical procedures.

A critical factor to consider when implementing video-text generative models in medical videos is real-time interaction with a physician. Unlike 3D medical images, which are generally captured in a predetermined protocol and interpreted after patient encounters, medical videos can be influenced by the physician's decisions and AI assistance during the procedure. For example, the gastroenterologist or surgeon may extend inspection time in critical areas identified by AI. Real-time AI support can aid live clinical decisions, such as whether to biopsy or resect lesions, necessitating swift interaction between the AI model and a physician. Given the live, interactive nature of endoscopy and surgery videos, inference speed is crucial. Recent advancements in vision-language generative AI, such as GPT-4o, have shown the potential for real-time interaction.

As discussed earlier, metadata is essential for interpreting endoscopic and laparoscopic videos. Protocol data and clinical information should be provided to enhance the accurate interpretation of these videos. The potential application of a video-text generative model considering these factors is illustrated in Fig. 4.

Opportunities and Applications

The impact of video-text generative models on clinicians' workflow could be extensive and profound, enhancing both efficiency and patient outcomes. These models can automatically draft preliminary reports for both 3D medical images[56] and medical videos[64], significantly reducing the time clinicians spend on documentation. Such initial drafts can facilitate emergency triage, enabling timely intervention through prompt alerts to clinicians. Moreover, workflow phase recognition for surgical and endoscopic video may enable post-hoc analytics focused on measuring efficiency and quality of procedures while providing constructive feedback to clinicians[65]. Generative models may also offer real-time guidance to augment procedural performance and reduce cognitive load during challenging endoscopic or surgical tasks, for instance by highlighting safe planes of tissue dissection[66], or by automatically halting cautery current if an electrosurgical knife is inadvisably close to a critical structure.

Video-text retrieval techniques powered by multimodal AIs can assist clinicians to quickly search for similar cases in a database of tomograms or medical videos, facilitating comparative research and aiding in the diagnosis of rare or challenging cases. This technology can also improve communication among clinicians. A common challenge in medical data is correlating written findings with the images without proper annotation. Traditionally, this requires consultation with specialists. Video-text generative models can streamline this process by displaying relevant image slices or video frames corresponding to the written report and offering question-answering functionalities for further clarification.

These video-text generative models can also create educational content based on specific diagnoses or textual descriptions. Modern text-to-video generation models have demonstrated a remarkable ability to simulate realistic and temporally coherent videos from text inputs[67,68]. These synthetic surgical or endoscopy videos and simulated 3D medical images can serve as valuable training tools for a wide range of clinicians while preserving patient privacy.

The learned representations from training video-text generative models on 3D medical images or medical videos can have broader applications beyond their primary uses. These representations, which capture 3D anatomical relationships and multimodal correlations, could enhance diagnostic accuracy even with limited or low-quality imaging data, such as from abbreviated exams or point-of-care devices.

Finally, recent advances in scaling test-time compute[69] have substantially boosted the performance of vision-language models across various benchmarks[70,71]. These improvements in reasoning capabilities provide promising solutions for tackling the inherent complexity of interpreting 3D medical images and medical videos. Since 3D medical images and videos are typically voluminous, with clinically relevant details often distributed in complex patterns across the data, models with advanced reasoning abilities

can more effectively identify and integrate these scattered findings. This capability substantially increases the practical utility of video-text models in clinical settings, particularly for diagnostic tasks requiring sophisticated integration of spatial and temporal information. Thus, reasoning models hold significant potential for bringing the clinical applications discussed above closer to real-world implementation.

Challenges and Future Directions

The primary challenge in adapting video-text models to 3D medical imaging and medical videos is data scarcity. While self-supervised methods for video-text models have advanced, there are only a few small open-source datasets[72,73] suitable for 3D medical images and medical videos. For example, a recent relatively large video-text dataset contains seven million videos with 234 million video clips[74], whereas CLIP on 2D images was trained on 400 million paired image-text pairs[6]. This highlights the lack of equivalent datasets for 3D medical images. A practical approach, therefore, involves pretraining video-text models on existing video-text datasets and then fine-tuning them on 3D medical data or medical videos[51,75]. Concurrently, global efforts to create and share more open-source datasets are crucial.

Various vision-language models have emerged in the medical field, including BioMedGPT[76], MedPaLM[77], LLaVa-Med[78], and MedVersa[79]. However, most of these models cannot adequately handle multiphase 3D images or long medical videos, and none utilize multiple phases or sequences simultaneously during training. Our proposed training methods could help these models develop robust interpretive capabilities, but implementation requires carefully curated datasets containing complete image sequences and longitudinal studies. However, such comprehensive datasets raise legitimate privacy concerns, as 3D medical images can be reconstructed into recognizable shapes, and the inclusion of multiple time points increases the risk of patient re-identification. Possible solutions include institution-level deidentification, as used in MIMIC dataset[80], and mild distortion of surface areas to protect patient privacy.

The field also lacks suitable datasets for downstream tasks and benchmarks to evaluate model performance in interpreting 3D medical images and videos, especially regarding synergistic information and unique world models. Without these resources, we cannot properly assess how video-text models fail in medical image and video interpretation. Suggesting differential diagnosis and lesion localization represent promising downstream tasks to verify proper interpretation of synergistic information. For example, distinguishing between liver hemangioma, adenoma, and focal nodular hyperplasia, or differentiating between brain abscesses and cysts on MRI, requires models to integrate information from multiple phases and sequences. Similarly, accurate lesion description in medical videos demands a sophisticated world model from video-text models.

While reasoning models show promise in this domain, appropriate reasoning datasets to train them do not yet exist. Fortunately, radiologic and endoscopic/laparoscopic reports already contain substantial reasoning information, potentially simplifying the construction of such datasets. This approach may offer the most direct path to enhancing the interpretive performance of video-text models.

Beyond the unique characteristics previously discussed, further complexities arise when adapting video-text models to 3D medical imaging and videos. Clinical reports paired with the medical data often contain varied information such as comparisons with prior studies or different types of exams, complicating the training process. Moreover, real-world tomograms can also include reconstructed volume images and dynamic videos. Techniques like cerebral CT angiography and MR cholangiography generate 3D visualizations of anatomical structures from various angles, while cardiac MRI captures heart motion in video sequences. These added complexities necessitate continuous research to enhance the training methodologies for video-text generative AI models in the medical field.

Conclusion

The successful application of video-text generative models in medicine can revolutionize clinical workflows, enhance diagnostic accuracy, facilitate clinician-clinician communication, and provide valuable tools for education and training. Despite this promising potential, several challenges must be addressed. These include the limited availability of large-scale, open-source datasets suitable for self-supervised learning, the complexity of training models on data containing synergistic information, and the engineering hurdles associated with the unique structures of 3D medical images and medical videos. To overcome these barriers, future research should focus on creating comprehensive datasets that preserve patient privacy, developing benchmarks to evaluate models' capabilities to integrate multi-sequence information, and advancing training methodologies tailored specifically to the complex multimodal characteristics of 3D medical imaging and medical videos.

Data Availability

No datasets were generated or analyzed during the current study.

Received: 1 November 2024; Accepted: 18 April 2025;
Published online: 13 May 2025

References

1. Chi, E. A. et al. Development and validation of an artificial intelligence system to optimize clinician review of patient records. *JAMA Netw. Open* **4**, e2117391 (2021).
2. Yala, A. et al. Toward robust mammography-based models for breast cancer risk. *Sci. Transl. Med.* **13**, eaba4373 (2021).
3. Homayounieh, F. et al. An artificial intelligence–based chest X-ray model on human nodule detection accuracy from a multicenter study. *JAMA Netw. Open* **4**, e2141096 (2021).
4. Wang, P. et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* **68**, 1813–1819 (2019).
5. Bommasani, R. et al. On the Opportunities and Risks of Foundation Models. Preprint at http://arxiv.org/abs/2108.07258 (2022).
6. Rafiq, G., Rafiq, M. & Choi, G. S. Video description: A comprehensive survey of deep learning approaches. *Artif. Intell. Rev.* **56**, 13293–13372 (2023).
7. Antol, S. et al. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)* 2425–2433 (IEEE, Santiago, Chile, 2015). https://doi.org/10.1109/ICCV.2015.279.
8. Bain, M., Nagrani, A., Varol, G. & Zisserman, A. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 1708–1718 (IEEE, Montreal, QC, Canada, 2021). https://doi.org/10.1109/ICCV48922.2021.00175.
9. Singer, U. et al. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *The Eleventh International Conference on Learning Representations* (2023).
10. Nguyen, P., Quach, K. G., Kitani, K. M. & Luu, K. Type-to-Track: Retrieve Any Object via Prompt-based Tracking. In *Thirty-seventh Conference on Neural Information Processing Systems* (2023).
11. Radford, A. et al. Learning Transferable Visual Models From Natural Language Supervision. In *Proc. 38th International Conference on Machine Learning* (eds. Meila, M. & Zhang, T.) vol. 139 8748–8763 (PMLR, 2021).
12. Luo, H. et al. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing* **508**, 293–304 (2022).
13. Xu, H. et al. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* 6787–6800 (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021). https://doi.org/10.18653/v1/2021.emnlp-main.544.
14. Zhu, B. et al. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. In *The Twelfth International Conference on Learning Representations* (2024).

15. Chen, S. et al. VAST: A Vision-Audio-Subtitle-Text Omni-Modality Foundation Model and Dataset.
16. Liu, H., Yan, W., Zaharia, M. & Abbeel, P. World Model on Million-Length Video And Language With Blockwise RingAttention. In *The Thirteenth International Conference on Learning Representations* (2025).
17. Zhang, H., Li, X. & Bing, L. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 543–553 (Association for Computational Linguistics, Singapore, 2023). https://doi.org/10.18653/v1/2023.emnlp-demo.49.
18. Lin, B. et al. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In *Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing* 5971–5984 (Association for Computational Linguistics, Miami, Florida, USA, 2024). https://doi.org/10.18653/v1/2024.emnlp-main.342.
19. OpenAI. *GPT-4o* https://openai.com/index/hello-gpt-4o/ (2024).
20. Gemini Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Preprint at http://arxiv.org/abs/2403.05530 (2024).
21. Wang, W. et al. Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 19175–19186 (IEEE, Vancouver, BC, Canada, 2023). https://doi.org/10.1109/CVPR52729.2023.01838.
22. Li, J., Li, D., Savarese, S. & Hoi, S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. 40th International Conference on Machine Learning* vol. 202 19730–19742 (JMLR.org, Honolulu, Hawaii, USA, 2023).
23. He, K. et al. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 15979–15988 (IEEE, New Orleans, LA, USA, 2022). https://doi.org/10.1109/CVPR52688.2022.01553.
24. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In *Proc. of the 37th International Conference on Machine Learning* vol. 119 1597–1607 (JMLR.org, 2020).
25. Wang, Y. et al. InternVideo: General Video Foundation Models via Generative and Discriminative Learning. Preprint at http://arxiv.org/abs/2212.03191 (2022).
26. Zhao, Y., Misra, I., Krähenbühl, P. & Girdhar, R. Learning Video Representations from Large Language Models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 6586–6597 (IEEE, Vancouver, BC, Canada, 2023). https://doi.org/10.1109/CVPR52729.2023.00637.
27. Yu, E. et al. Merlin: Empowering Multimodal LLMs with Foresight Minds. In *Computer Vision – ECCV 2024* (eds. Leonardis, A. et al.) vol. 15062 425–443 (Springer Nature Switzerland, Cham, 2025).
28. Liu, C. et al. T3D: Towards 3D Medical Image Understanding through Vision-Language Pre-training. Preprint at http://arxiv.org/abs/2312.01529 (2025).
29. Yuan, K. et al. Learning Multi-modal Representations by Watching Hundreds of Surgical Video Lectures. Preprint at http://arxiv.org/abs/2307.15220 (2024).
30. National Electrical Manufacturers Association. Digital Imaging and Communications in Medicine (DICOM) Part 5: Data Structures and Encoding. (2024).
31. Masoudi, S. et al. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. *J. Med. Imaging* **8**, (2021).
32. Gono, K. et al. Appearance of enhanced tissue features in narrow-band endoscopic imaging. *J. Biomed. Opt.* **9**, 568 (2004).
33. Takei, Y. et al. New diagnostic strategy using narrow-band imaging (NBI) during laparoscopic surgery for patients with colorectal cancer. *Surg. Endosc.* **36**, 8843–8855 (2022).
34. Oka, K. et al. Red dichromatic imaging improves visibility of bleeding during gastric endoscopic submucosal dissection. *Sci. Rep.* **13**, 8560 (2023).
35. Barua, I. et al. Real-time artificial intelligence–based optical diagnosis of neoplastic polyps during colonoscopy. *NEJM Evid*. **1**, (2022).
36. Bitar, R. et al. MR pulse sequences: what every radiologist wants to know but is afraid to ask. *RadioGraphics* **26**, 513–537 (2006).
37. Fleischmann, D. Use of high-concentration contrast media in multiple-detector-row CT: principles and rationale. *Eur. Radiol.* **13**, 14–20 (2003).
38. Van Der Merwe, S. W. et al. Therapeutic endoscopic ultrasound: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. *Endoscopy* **54**, 185–205 (2022).
39. Chathadi, K. V. et al. The role of ERCP in benign diseases of the biliary tract. *Gastrointest. Endosc.* **81**, 795–803 (2015).
40. Petrella, J. R. & Provenzale, J. M. MR Perfusion Imaging of the Brain: Techniques and Applications. *Am. J. Roentgenol.* **175**, 207–219 (2000).
41. Lee, R. H. et al. Quality of colonoscopy withdrawal technique and variability in adenoma detection rates (with videos). *Gastrointest. Endosc.* **74**, 128–134 (2011).
42. Aarts, J. W. M. et al. Surgical approach to hysterectomy for benign gynaecological disease. *Cochrane Database Syst. Rev.* **2015**, CD003677 (2015).
43. Juanpere, S. et al. A diagnostic approach to the mediastinal masses. *Insights Imaging* **4**, 29–52 (2013).
44. Huang, R. J., Choi, A. Y., Truong, C. D., Yeh, M. M. & Hwang, J. H. Diagnosis and Management of Gastric Intestinal Metaplasia: Current Status and Future Directions. *Gut Liver* **13**, 596–603 (2019).
45. Ha, D. & Schmidhuber, J. World Models. https://doi.org/10.5281/zenodo.1207631 (2018) .
46. Rajpurkar, P. et al. AppendiXNet: Deep Learning for Diagnosis of Appendicitis from A Small Dataset of CT Exams Using Video Pretraining. *Sci. Rep.* **10**, 3958 (2020).
47. Ke, A. et al. Video pretraining advances 3D deep learning on chest CT tasks. In *Medical Imaging with Deep Learning* (2023).
48. Zunair, H., Rahman, A. & Mohammed, N. ViPTT-Net: Video pretraining of spatio-temporal model for tuberculosis type classification from chest CT scans. In *Proc. Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021* (eds. Faggioli, G., Ferro, N., Joly, A., Maistro, M. & Piroi, F.) vol. 2936 1412–1421 (CEUR-WS.org, 2021).
49. Dai, C. et al. Deep Learning Assessment of Small Renal Masses at Contrast-enhanced Multiphase CT. *Radiology* **311**, e232178 (2024).
50. Xu, H. et al. mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video. In *Proc. 40th International Conference on Machine Learning* vol. 202 38728–38748 (JMLR.org, Honolulu, Hawaii, USA, 2023).
51. Saab, K. et al. Capabilities of Gemini Models in Medicine. Preprint at http://arxiv.org/abs/2404.18416 (2024).
52. Pai, S. et al. Foundation model for cancer imaging biomarkers. *Nat. Mach. Intell.* **6**, 354–367 (2024).
53. Zhu, W. et al. 3D Foundation AI Model for Generalizable Disease Detection in Head Computed Tomography. Preprint at https://doi.org/10.48550/arXiv.2502.02779 (2025).
54. Blankemeier, L. et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. Preprint at http://arxiv.org/abs/2406.06512 (2024).
55. Wasserthal, J. et al. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiol. Artif. Intell.* **5**, e230024 (2023).
56. Hamamci, I. E., Er, S. & Menze, B. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024* vol. LNCS 15012 (Springer Nature Switzerland, 2024).

57. Krishna, R., Hata, K., Ren, F., Fei-Fei, L. & Niebles, J. C. Dense-Captioning Events in Videos. In *2017 IEEE International Conference on Computer Vision (ICCV)* 706–715 (IEEE, Venice, 2017). https://doi.org/10.1109/ICCV.2017.83.
58. Matsubara, M. et al. Clinical significance of esophagogastroduodenoscopy in patients with esophageal motility disorders. *Dig. Endosc. J. Jpn. Gastroenterol. Endosc. Soc.* **33**, 753–760 (2021).
59. Hsieh, Y.-H., Tang, C.-P., Tseng, C.-W., Lin, T.-L. & Leung, F. W. Computer-Aided Detection False Positives in Colonoscopy. *Diagnostics* **11**, 1113 (2021).
60. Goel, N., Kaur, S., Gunjan, D. & Mahapatra, S. J. Investigating the significance of color space for abnormality detection in wireless capsule endoscopy images. *Biomed. Signal Process. Control* **75**, 103624 (2022).
61. Sato, T. TXI: texture and color enhancement imaging for endoscopic image enhancement. *J. Healthc. Eng.* **2021**, 1–11 (2021).
62. Nie, C., Xu, C., Li, Z., Chu, L. & Hu, Y. Specular Reflections Detection and Removal for Endoscopic Images Based on Brightness Classification. *Sensors* **23**, 974 (2023).
63. Yang, A. et al. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10714–10726 (IEEE, Vancouver, BC, Canada, 2023). https://doi.org/10.1109/CVPR52729.2023.01032.
64. Zhang, L. et al. Effect of a deep learning–based automatic upper GI endoscopic reporting system: a randomized crossover study (with video). *Gastrointest. Endosc.* **98**, 181–190.e10 (2023).
65. Yanik, E., Schwaitzberg, S. & De, S. Deep learning for video-based assessment in surgery. *JAMA Surg.* **159**, 957 (2024).
66. Kumazu, Y. et al. Automated segmentation by deep learning of loose connective tissue fibers to define safe dissection planes in robot-assisted gastrectomy. *Sci. Rep.* **11**, 21198 (2021).
67. Peebles, B. et al. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators (2024).
68. Google Deepmind. Veo. https://deepmind.google/technologies/veo/ (2024).
69. Snell, C. V., Lee, J., Xu, K. & Kumar, A. Scaling LLM Test-Time Compute Optimally Can be More Effective than Scaling Parameters for Reasoning. In *The Thirteenth International Conference on Learning Representations* (2025).
70. Xu, G. et al. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. Preprint at https://doi.org/10.48550/arXiv.2411.10440 (2025).
71. Sun, G. et al. video-SALMONN-o1: Reasoning-enhanced Audio-visual Large Language Model. Preprint at https://doi.org/10.48550/arXiv.2502.11775 (2025).
72. Bai, F., Du, Y., Huang, T., Meng, M. Q.-H. & Zhao, B. M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models. Preprint at http://arxiv.org/abs/2404.00578 (2024).
73. Misawa, M. et al. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointest. Endosc.* **93**, 960–967.e3 (2021).
74. Wang, Y. et al. InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation. In *The Twelfth International Conference on Learning Representations* (2024).
75. Sun, W. et al. Bora: Biomedical Generalist Video Generation Model. Preprint at https://doi.org/10.48550/arXiv.2407.08944 (2024).
76. Zhang, K. et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nat. Med.* **30**, 3129–3141 (2024).
77. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
78. Li, C. et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2023).
79. Zhou, H.-Y., Adithan, S., Acosta, J. N., Topol, E. J. & Rajpurkar, P. A Generalist Learner for Multifaceted Medical Image Interpretation.
80. Johnson, A. et al. MIMIC-IV. PhysioNet https://doi.org/10.13026/KPB9-MT58.

Acknowledgements

We thank Michael Moritz for his valuable comments and feedback on the manuscript. This study received no funding.

Author contributions

J.L. conceived the concept and led the project. P.R. further developed the concept and supervised the project. H.Z. conducted the literature review and provided critical feedback on the manuscript. T.M.B. contributed expertise in medical videos, particularly endoscopy. D.K.S. provided expertise in 3D medical imaging. All authors discussed the concepts and examples and reviewed and commented on the manuscript at all stages.

Competing interests

T.M.B. is a consultant for Boston Scientific, Medtronic, and Magentiq Eye. All the other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Pranav Rajpurkar.