nature reviews electrical engineering

https://doi.org/10.1038/s44287-025-00166-6

Perspective

Check for updates

Innovating robot-assisted surgery through large vision models

Zhe Min[1,2,4], Jiewen Lai[3,4] & Hongliang Ren[3] ✉

Abstract

The rapid development of generative artificial intelligence and large models, including large vision models (LVMs), has accelerated their wide applications in medicine. Robot-assisted surgery (RAS) or surgical robotics, in which vision has a vital role, typically combines medical images for diagnostic or navigation abilities with robots with precise operative capabilities. In this context, LVMs could serve as a revolutionary paradigm towards surgical autonomy, accomplishing surgical representations with high fidelity and physical intelligence and enabling high-quality data use and long-term learning. In this Perspective, vision-related tasks in RAS are divided into fundamental upstream tasks and advanced downstream counterparts, elucidating their shared technical foundations with state-of-the-art research that could catalyse a paradigm shift in surgical robotics research for the next decade. LVMs have already been extensively explored to tackle upstream tasks in RAS, exhibiting promising performances. Developing vision foundation models for downstream RAS tasks, which is based on upstream counterparts but necessitates further investigations, will directly enhance surgical autonomy. Here, we outline research trends that could accelerate this paradigm shift and highlight major challenges that could impede progress in the way to the ultimate transformation from 'surgical robots' to 'robotic surgeons'.

Sections

Introduction

Large vision models for robot-assisted surgery upstream tasks

Large vision models for downstream tasks for RAS

Looking ahead

[1]School of Control Science and Engineering, Shandong University, Jinan, China. [2]Department of Medical Physics and Biomedical Engineering, University College London, London, UK. [3]Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China. [4]These authors contributed equally: Zhe Min, Jiewen Lai. ✉e-mail: hlren@ee.cuhk.edu.hk

Key points

- Large vision model-driven surgical robots could become a new paradigm for achieving higher level of surgical autonomy.
- In robot-assisted surgery (RAS), vision-related tasks are broadly classified into upstream tasks including classification, detection, segmentation and registration and downstream counterparts encompassing cognition, simulation, diagnosis and robot control.
- Large vision models have been extensively explored to tackle upstream tasks in RAS, demonstrating great effectiveness and promising performances.
- Incorporating vision foundation models into downstream RAS tasks that typically involve multidimensional and multimodal data directly enhances surgical autonomy and intelligence, which to some extent can be achieved through leveraging upstream image processing advancements but necessitates further investigations.
- Future research trends towards developing large models for downstream tasks (and beyond) and increasing autonomy level in RAS encompass enhancing data collection, achieving physics-aware artificial intelligence models, developing surgical large multimodal models, boosting models' explainability and strengthening cross-disciplinary collaborations.
- Looking forward, with technical, application, ethical and regulatory challenges to be tackled along the road, a pathway to develop multimodal, downstream-task-oriented, high-dimensional and physics-aware large models for achieving higher RAS autonomy level is on the horizon.

Introduction

The advent of large vision models (LVMs)[1] represents a fundamental shift in foundation model development, challenging the dominance of large language models (LLMs). Evolving from conventional deep-learning-based models that rely on manual labels, LVMs are intended to autonomously learn from massive data sets with hundreds of millions of parameters — generally in a self-supervised learning manner — powering their capability to dig out intricate patterns and inexplicable connections within images. LVMs ingest massive data sets to recognize and conceive visual concepts, benchmarked by the popular Imagen by Google DeepMind and Stable Diffusion by Stability AI. Transfer learning, which involves fine-tuning a model trained for one task to perform a different one (usually from a general domain to a specific domain[2]), is often used to adapt pretrained LVMs to specific new tasks. LVMs have exhibited either great performances or promising potentials for a wide range of vision-related tasks, from the upstream object recognition problem to downstream healthcare[3] and robotics applications[4]. Robot-assisted surgery (RAS) is one of the fastest-growing areas among the ones that can be equipped with the latest breakthroughs in computer vision, image analysis and precise robot control algorithms[5]. Consequently, it is necessary to fully understand how LVMs can contribute to the development of more intelligent surgical robots with the ambitious goal to achieve copilot robotic surgery, or even autopilot RAS procedures[6].

RAS leverages the high accuracy and robustness associated with robots and the intelligence of computerized algorithms to simultaneously enhance the interventional performance and levels of automation, as well as the ultimate outcome of surgery with less labour input from healthcare professionals[7,8]. As a major way of perceiving, vision intertwines with RAS practices in various degrees (such as from vision-based detection of lesions to enhanced precision for the closed-loop robot motion control). Thus, it is reasonable to hypothesize that LVMs could benefit both upstream and downstream tasks in RAS (Box 1 and Fig. 1). Developing LVM-driven surgical robots would bring three advantages. First, a new paradigm towards surgical autonomy across various data modalities and different tasks can be established, possibly with risk awareness and control mechanisms. Second, realistic surgical representations with high fidelity and physical intelligence can be realized, possibly using generative spatial artificial intelligence (AI), which will retain or produce more refined details of surgical scenes or instruments. Third, high-quality data use and long-term algorithmic development can be achieved, possibly with self-supervision and in-context learning scheme, decreasing the burden of modifying incorrect data annotations and continuously enhancing the intelligence, versatility and capability of the surgical robotic system.

However, directly applying 'unseasoned' models in RAS tasks is impractical because of the substantial domain gap between natural and surgical scenes. More specifically, although current LVMs are adept at identifying pets, tourist attractions and commonly used items — owing to the abundance of internet natural images in their training data sets — they falter when it comes to specific image categories in professional fields. One such category includes the specialized visions within surgical robotic systems, ranging from the preoperative diagnostic images to the intraoperative cameras[9], which exhibit considerable discrepancies with natural images (for example, light conditioning, texture and patterns). Furthermore, the distinct nature of RAS introduces various challenges in adapting LVMs to surgical scenarios, such as the limited availability of specialized medical/surgical image data for training or fine-tuning, the high complexity of personalized medical profiles and the unique requirement of RAS (for example, high accuracy and safety guarantee).

LVMs for universal segmentation of a wide range of anatomical structures and lesions, such as MedSAM[10], have attracted attention in the field of medical image computing. Nevertheless, segmentation represents only one of the numerous upstream tasks in RAS. We are thus particularly motivated to showcase ways of applying or adapting generalist LVMs to various upstream and downstream tasks (and beyond) in RAS, to realize LVMs' benefits for RAS under the above-outlined challenges. In contrast to previous reviews of applications of AI in the field of medical robotics[6] or healthcare[2,3], our study has a distinct focus on how prospering LVMs could particularly benefit key vision-related tasks in RAS and further enhance surgical autonomy, along with identified future research trends and major challenges.

In this Perspective, we discuss how LVM-equipped approaches could have game-changer roles in RAS given its distinct characteristics. We first introduce LVMs specifically developed for upstream tasks in RAS, showcasing superior performances of LVMs over existing task-specific or modality-specific supervised approaches. Furthermore, we describe LVMs relevant with downstream tasks and elaborate their great promising impact for RAS. The dependence of downstream tasks on upstream counterparts is also discussed, for example, the downstream cognition task during surgery necessitates all upstream

categories (that is, surgical phase classification, surgical instrument detection and segmentation and multimodal medical image registration). Finally, we identify the main future research trends of developing surgical foundation models and enhancing autonomy levels of RAS (including data enhancement, robophysics, surgical large multimodal models (LMMs), explainability and cross-disciplinary collaboration) and core challenges that need to be overcome for clinical deployment of large models, including technical, application, ethical and regulatory aspects.

Large vision models for robot-assisted surgery upstream tasks

The main upstream tasks in RAS include classification, detection, segmentation and registration (Fig. 1). Domain-specific LVMs have been explored to tackle upstream tasks (indicated by the rapidly growing publications shown in Fig. 2) and have demonstrated their effectiveness and advantages over existing algorithms.

Classification

Medical image classification is the ability of assigning specific categories to input images — which can be binary or multiclass — such as distinguishing malignant from benign lesions or tumour grading[11]. Foundation models could benefit this upstream task by reducing training time and the use of labelled data. For example, RETFound[12], a foundation model specific for retinal images, was successively trained on large-scale unlabelled retina images through self-supervised learning and then fine-tuned on task-specific labelled data, resulting in superior performance in downstream tasks including diabetic retinopathy classification and label efficiency compared with state-of-the-art competing models[12]. Furthermore, the zero-shot contrastive learning-based classification model CheXzero[13] was trained on pairs of unannotated X-rays and radiology reports, achieving comparable performances in the pathology classification task[14] with those obtained by expert radiologists or using supervised methods. A multimodal vision-text contrastive learning method used to decouple images and texts, such as CLIP[15], can be adapted to a medical domain, for example, MedCLIP[16], achieving a 10% improvement over baselines under both zero-shot and supervised classification scenarios.

Classifying surgical instruments correctly in endoscopic images is essential for downstream cognition tasks in RAS. Given the fact that instrument classification is indeed strongly coupled with detection and segmentation tasks, there is no standalone surgical instrument classification foundation model. Nevertheless, we anticipate that the contrast learning technique could benefit the classification of surgical instruments[17]. For the endoscopic polyp classification, a foundation model for gastrointestinal endoscopic image analysis (GastroNet-5M)[18] has been established by first constructing a large data set consisting of 5 million gastrointestinal endoscopic images. This model exploits self-supervised learning techniques such as DINO[19] to pretrain the model[18], whose superior performance demonstrates the marked benefits of in-domain pretraining. By pretraining GastroNet-5M with DINO, the model achieved an area under the receiver-operating curve of 93.36% in the polyp classification task on a downstream data set[20], outperforming the 87.76% achieved with a supervised pretrained model on ImageNet-1K (both are fine-tuned). For the prostate cancer grading (classification) task, a histopathology image evaluation foundation model called CHIEF[21] has been developed using more than 60,530 whole-slide images, including 10,616 whole-slide images from prostate biopsies in the PANDA challenge[22].

Box 1 | Upstream and downstream tasks in robot-assisted surgery

Here, we classify vision-related tasks involved in robot-assisted surgery as upstream and downstream categories (Fig. 1). Upstream tasks are defined as fundamental or independent problems including the classification of medical images, the single-organ/multi-organ segmentation in the diagnostic images[33,34], detection and segmentation of lesions in diagnostic images[10,98], segmentation and pose estimation of surgical tools during surgery, depth estimation and 3D reconstruction of surrounding tissues[45,137,138], registration (rigid and deformable) of multiple image modalities[52,53,55,56,98] and so forth. Downstream tasks refer to cognitive tasks based on and beyond upstream tasks, including surgical image understanding such as surgical action triplet recognition (that are instrument, verb and target)[139], surgical captioning[140], visual question answering in surgery[67,69,71,136], surgical phase recognition and surgical skill assessment[74], computer-assisted diagnosis that helps diagnose and grade the disease automatically[141], surgical navigation that provides accurate interventional guidance[142], robot control[6] and decision-making assistance mechanism that provides informed decisions during surgery[117]. On the basis of our categorization, we conducted an advanced search in Scopus to identify the number of publications from 2018 to 2024 (as of November), revealing a noticeable disparity in the development of upstream and downstream tasks, with upstream tasks receiving more attention (Fig. 2).

Detection

Medical object detection refers to the localization and categorization of interested objects (such as organs, lesions and surgical instruments) in medical images. For surgical instrument detection, Surgical-DeSAM[23] leveraged and fine-tuned the transformer-based detector called DETR[24] to automatically obtain the instrument bounding box prompts for segment anything model (SAM), being a popular promptable image segmentation system[25]. For poly detection in endoscopic images[26], by leveraging the pretrained vision-language models GLIP[27] in which object detection is reformulated as a phrase grounding task and manually text prompts, the fine-tuned MIU-VL[28] achieved an improvement of 6.5% in average precisions over the best-supervised model. Besides, 'one-for-all' generalist models are also desired in biomedical detection. For example, BiomedParse[29], a biomedical vision-language foundation modal, can jointly conduct segmentation, detection and recognition for 82 object types across 9 distinct imaging modalities. Developing 3D foundation models for medical object detection is desirable but non-trivial as more pretrained models are trained on 2D[24]. In this regard, the 3D medical foundation localization model MedLAM[30] was developed with two self-supervised tasks and several template images (that is, few-shot learning), demonstrating comparable or better performances in detecting anatomical structures compared with fully supervised models. For example, 5-shot MedLAM[30] achieved a mean Intersection-of-Unions of 76.7% in detecting the left femur's head, significantly outperforming ($P < 0.05$) nnDetection[31] (70.5%) and Mask RNN[32] (20.0%). Furthermore, MedLAM[30] can be used to automatically produce box prompts that are further adopted by MedSAM[10], resulting in the fully automated foundation segmentation model MedLSAM[30].

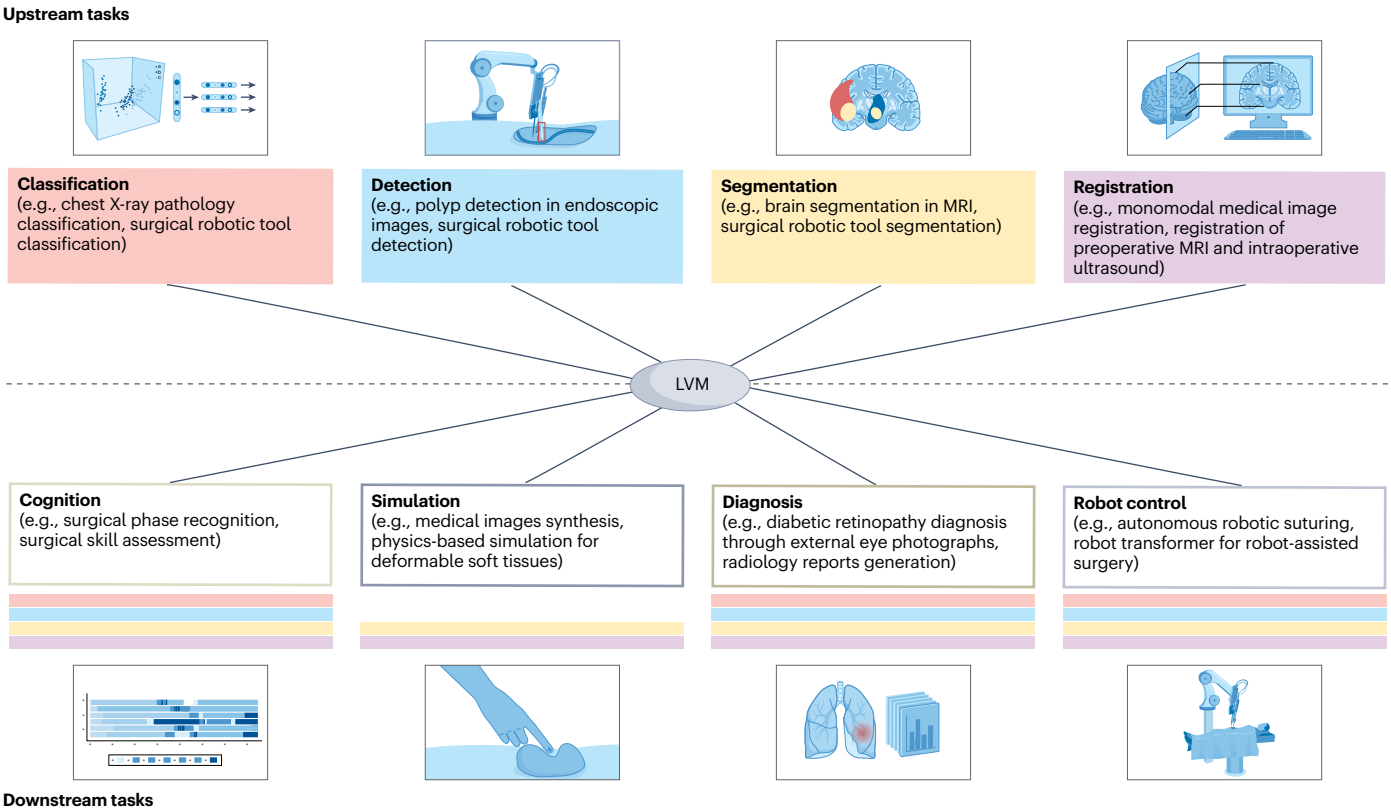

Fig. 1 | Applications of large vision models to upstream and downstream robot-assisted surgery tasks. Upstream tasks: a growing number of pilot studies are exploring the use of large vision models (LVMs), whether pre-trained with general images or domain-specific medical images, on the upstream robot-assisted surgery tasks, including technical implementations in classification, detection, segmentation and registration. Downstream tasks: in downstream robot-involved applications, a solid upstream implementation paves the way for realizing 'tasks beyond medical image processing', including high-level cognition, simulation, medical diagnosis beyond expert capability and complicated semi/full robotic control. Traditional approaches for downstream robot-assisted surgery tasks usually accept data sets in limited domains, resulting in high dependency on specific sources that may not always be accessible. Meanwhile, with LVM-based exploration, robots might operate with versatile visual perception and eventually act more like well-trained healthcare professionals.

Segmentation

Medical image segmentation refers to the delineation of interested regions through pixel-wise or voxel-wise classification in medical images. Between 2023 and 2024, three heuristic works[10,33,34] explored the feasibility and efficacy of adapting SAM[25], with strong zero-shot generalization, for medical image segmentation. Because of the substantial difference between natural and medical images, especially those images segmenting medical targets with weak boundaries or low contrast, one feasible solution is to fine-tune the SAM on medical images[10,33]. For example, by fine-tuning the SAM with more than 1 million medical images spanning 10 image modalities and over 30 cancer types, MedSAM[10] demonstrated better accuracy and robustness than modality-wise specialist models. Specifically, MedSAM achieved the median dice similarity coefficient of 97.4%, for segmenting prostate gland in MRI on their external validation data set, significantly outperforming ($P < 0.05$) SAM (90.0%), modality-wise specialist nnUNet[35] (95.0%) and DeepLabV3+ (ref. 36) (90.4%).

SAMs have also been exercised in dealing with typical RAS tasks. For surgical robotic instrument segmentation, the performance of SurgicalSAM[37] implied the importance of box prompts and domain adaptation but also raised concerns about the adverse effect of the domain gap between natural and surgical scenes. Those effects include failures to identify instruments in complex surgical scenarios (influenced, for example, by the presence of blood and reflection) and performance degradation under data corruption (caused by, for example, noise or blur). It is however possible to mitigate this gap through fine-tuning using low-rank adaption (LoRA)[38]. To eliminate the requirement of point or box prompts using SAM, another SurgicalSAM[17] integrated a prototype-based class prompt encoder with SAM whereas a discriminative class prototype was obtained with contrastive prototype learning. However, the reliance on the class prompt necessitated knowing the instrument class in advance. In addition, SAMs can also be leveraged in segmenting anatomical structures (for example, gall bladder, liver and gastrointestinal tract) in surgical images[39–41], in which the name of a class of interest is provided as the text prompt. Furthermore, MedSAM achieved a median dice similarity coefficient of 97.7% in segmenting polyps in endoscopic images on their external validation data set, outperforming SAM (93.9%), modality-wise specialist nnUNet (95.8%) and DeepLabV3+ (95.2%), respectively ($P < 0.05$).

It is inevitable that the development of LVM for medical image segmentation will expand from 2D to 3D[42]. For instance, MedSAM2 (ref. 40) treated 3D medical images as videos and unlocked the one-prompt

segmentation capability. It achieved a segmentation dice score of 88.57%, outperforming zero-shot SAM-2 (ref. 41) and fully supervised MedSegDiff[43], whose dice scores are 51.6% and 87.9%, respectively, for the multi-organ segmentation task on the Beyond the Cranial Vault data set[44], which is a computerized axial tomography data set. Meanwhile, segmenting 3D minimally invasive surgical scenes remains unaddressed because of the scarcity in 3D data availability and spatial complexity. Depth estimation is often discussed with semantic segmentation, as they contribute to the scene understanding. It refers to estimating the depth value, that is, the distance relative to the camera, of each pixel given a single monocular image or stereo image pairs. In this regard, Surgical-DINO[45] made attempts to adapt foundation models for 3D reconstruction of RAS scenes. To explicitly consider the domain gap between natural and surgical scenes, LoRA[38] was added to the image encoder of DINOv2 (ref. 46) to fine-tune the model. The generalist Depth-Anything Model (DAM)[47] has been tested on laparoscopic images, showing impressive zero-shot capability[48]. Surgical-Depth-Anything[49] further fine-tunes the DAM using surgical data sets with ground-truth depth maps and has reduced the absolute error by 15%, 44% and 50% for the stomach, small intestine and colon, respectively. These advances in exploring the anatomical landscape offer valuable tools for subsequent cognitive and diagnostic applications.

Registration

Compared with the other three upstream tasks, registration tasks, that connect prevalent image processing technologies to physical robots, have been less explored using LVMs (Fig. 2). Medical image registration refers to the estimation of the optimal spatial transformation, which can be a parametric (for example, rigid, affine or thin-plate splines) or non-parametric (such as voxel-wise dense displacement vectors) representation, that aligns a pair of source and fixed images[50,51]. In 2024, two medical image registration foundation models achieved good performance across multiple anatomical structures (lung, knee, brain and abdomen) with different modalities (such as CT, cone beam CT and MRI)[52,53]. This result is otherwise infeasible using conventional learning-based registration methods. Nevertheless, uniGradICON[53] focused on monomodal medical image registration, whereas DINOReg[52] was validated only to register preoperative fan-beam CT and intraoperative cone beam CT.

However, multimodal medical image registration models are essential for typical RAS procedures to enable downstream interventional navigation[54] where distinct imaging modalities usually exist in preoperative and intraoperative spaces and have to be registered accurately. For example, the preoperative MRIs should be fused with the intraoperative ultrasound data to enable MRI-targeted transperineal prostate biopsy[55–57]. In this regard, MultiGradICON[58] can be considered the first foundation model for both monomodal and multimodal medical image registration. The model was directly trained on an image corpus containing 5 anatomical regions and 1 whole-body MRI data set and 12 different image modalities, achieving an 11% dice score improvement on the MRI–CT registration task compared with the uniGradICON[53]. Segmentation information can benefit the registration process. For example, SAMReg[59] leverages the off-the-shelf SAM model to segment multiple regions-of-interest in two images to be registered and adopts their similarities in the feature space to estimate regions-of-interest correspondences. The estimated regions-of-interest correspondences can be optionally converted to voxel-wise dense displacement fields. This training-free registration approach demonstrates superior and comparative performances to conventional, unsupervised and weakly supervised registration approaches, achieving mean dice of 75.67% (SAMReg) against 7.68% (NiftyReg[60]), 56.84% (VoxelMorph[61]) and 77.32% (LabelReg[62]), respectively, for the intersubject 3D prostate registration task.

As of the beginning of 2025, the routine of developing foundation models for upstream tasks in RAS is to fine-tune the natural image-based pretrained models with medical images with a few exceptions. Existing biomedical LVMs generally exhibit superior performance over the organ, modality or task-specific supervised approaches. They are applicable to 2D but are rapidly evolving to the 3D domain. The only exception is that most existing registration foundation models can handle 3D medical data, whereas some adopt 2D large models to operate on each slice. Thus far, LVMs have attracted much attention to tackle upstream tasks, with marked benefits such as reducing the need for costly labels and incorporating medical knowledge.

There are some other upstream tasks whose solutions have yet to be developed through leveraging large models. Examples include surgical tracking, localization and mapping[63]. Tracking refers to the motion estimation of the camera, instruments or tissue, which has potential applications in RAS, including autonomous scanning and image guidance. Tissue tracking is achieved either through dense optical flow or sparse feature matching. Mapping refers to creating consistent underlying representations (such as 3D point cloud or mesh) of the surgical scene, which is crucial for the surgeon or robot to accomplish downstream tasks, including decision-making, planning and navigation. Although in 2024, few works have developed foundation models for object tracking in the general computer vision field[64], both classical and neural-network-based methods cover the majority of work for these tasks[63]. As one essential technical building block, feature detection and descriptors have a key role in tracking and mapping tasks, such as being leveraged to establish data correspondences between images. For example, tissue tracking is achieved either through dense optical flow or sparse feature matching. Furthermore,

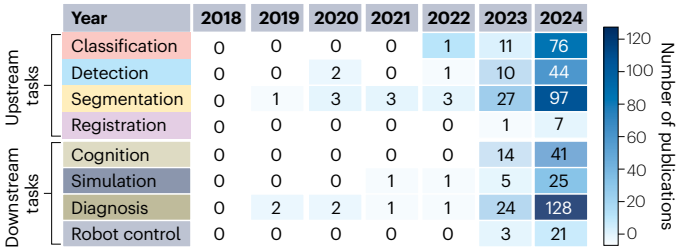

Fig. 2 | Number of publications in upstream and downstream robot-assisted surgery tasks in the years 2018–2024. The statistics is based on advanced searching in Scopus concerning titles, keywords and abstracts in journals and conference proceedings, with conference reviews and non-relevant articles excluded. For example, the 'detection' articles are searched via the query string TITLE-ABS-KEY (detection AND 'IMAGE' AND ('medical' OR ('surgery' OR 'surgical')) AND ('large vision model' OR 'large language model' OR 'LVM' OR 'LLM' OR 'foundation model')) AND PUBYEAR > 2017 AND PUBYEAR < 2025 AND NOT DOCTYPE ('cr'). Synonymous words are also included in some tasks, such as 'simulated', 'synthesis' and 'synthetic'. Key references are: Classification: RETFound[12], CheXzero[13], MedCLIP[16], GastroNet[18], CHIEF[21]; Detection: Surgical-DeSAM[23], MIU-VL[28], BiomedParse[29], MedLSAM[30]; Segmentation: MedSAM[10], SurgicalSAM (Prototype)[17], SurgicalSAM (LoRA)[37], AdaptiveSAM[39], MedSAM2[40], MA-SAM[42]; Registration: DINOReg[52], UniGradICON[53], MultiGradICON[58], SAMSeg[59]; Cognition: Surgical-GPT[136], Surgical-VQA[69], Surgical-VQLA[71], SAIS[74], GSViT[72], GP-VLS[73]; Simulation: Endora[80], Bora[82], ConStructS[85], Surgical-CD[86], PBD-RM-ON[90]; Diagnosis: BiomedGPT[97], GPT-4V[101]; Robot control: SRT[107], RAS-RT[111].

the data association terms using feature vectors can be adopted as terms in the cost function to be optimized in Simultaneous Localization and Mapping, where the map of the environment is created and the sensor (such as the surgical camera) is localized at the same time[63]. In this regard, the features extracted using LVMs can help these tasks in a way like the depth estimation problem.

Large vision models for downstream tasks for RAS

Downstream tasks in RAS include the resultant cognition, generative simulation, medical diagnosis and vision-guided robot control (Fig. 1). Incorporating visual foundation models into downstream surgical robotic tasks will enhance surgical autonomy and degrees of intelligentization. These tasks are established on top of image processing outcomes from the upstream counterparts. As with upstream tasks, the development of LVMs for downstream tasks has been on the rise since 2023 (Fig. 2).

Cognition

The watershed for upstream and downstream RAS tasks is those that require cognitive vision. Cognitive vision is the ability to process visual content beyond computational processing and manipulating and deliver comprehensive interpretations based on multiple sources and knowledge in different domains. Surgical scene understanding is a key example of this ability. For example, surgical phase recognition is a fundamental task for RAS systems. Most current learning-based methods rely on costly and labour-intensive hand annotations, which are detailed with precise timeframes for each surgical phase. Generating pseudo labels to limited timestamps for the training to reduce manual annotation proves the feasibility of effective surgical phase recognition[65], and recording finite temporal information with diverse importance demonstrates improvements in phase recognition outcomes[66]. However, these temporal-dependent methods are likely unable to recognize unplanned changes in surgical phases when ambiguous image frames are registered in the internal phase. By contrast, surgeons can always tell the intricate surgical phases not just by relying on operation time and surgical view but also knowledge from the past, standard operating procedure, preoperative assessment, intraoperative evaluation and so on, all of which can be 'visually' perceived. In this regard, LVM-based methods, leveraging more general, multimodal, non-domain-specific visual data sets, are more likely to establish an understanding of surgical scenarios and to recognize discontinuous surgical phases such as human surgeons[67]. These methods can be considered more anthropomorphic, as humans comprehend visual content through categorization within just 160 ms of seeing it[68]. Early works in exploring visual question answering for surgical scenes addressed the challenge of limited annotated medical data. For example, by introducing a residual multilayer perceptron-based VisualBert encoder to enhance visual–text token interaction, Surgical-VQA[69] outperformed the general visual question answering baseline in answering questions about surgical tools, their interactions and surgical procedures[70]. A similar study called Surgical-VQLA[71] also supported the necessity of developing domain-specific LVM specifically for surgery, either from the perspective of critical module or surgical data.

A vision model (GSViT)[72] based on the transformer recognized the surgical phase in a cholecystectomy with just a 'blink' (that is, a single frame) by leveraging a data set of 70 million frames from general surgical videos on YouTube. When leveraging domain-specific knowledge in surgery, models such as GP-VLS[73] achieved improved performances compared with the SOTA in four surgical visual question answering tasks, including automatic recognition of surgical phases (+8.2%), actions (+20.7%), tools (+14.5%) and triplet actions (+16.9%). Another example is SAIS[74], which is a unified framework that decodes subphase recognition, gesture classification and skill assessment, from surgical videos in RAS, by leveraging vision transformer and supervised contrastive learning. These examples indicate that integrating upstream techniques can initiate more advanced downstream tasks. For instance, the downstream cognition task containing subtasks such as surgical phase classification, surgical instrument detection and segmentation and interventional navigation (that usually necessitates multimodal image registration) could benefit from the development of large models with all four upstream tasks (Fig. 1).

Simulation

Traditional robotic simulations often rely on well-established models in mathematical or physical forms, indicating that the simulation is more of a graphic 'visualization' of some known settings. However, simulating RAS is challenging, as it involves interaction with soft tissues as well as dynamic, wet, responsive anatomy and diverse medical images. This approach seemed impossible for a reliable model-based reproduction in the virtual environment until the rise of LVMs. Synthesizing medical images using generative adversarial networks[75] and diffusion models[76] has shown great advantages in data augmentation, despite the fact that most existing methods focus on 2D images. The technology showcases great potential in data set problems that have long struggled with insufficiency and privacy concerns[77]. However, it has also been found that training on recursively generated data leads to the collapse of the AI model[78]. The success of text-to-image generating models, such as DALL-E, extends the capability of traditional generative adversarial networks with larger parameters to produce more realistic images in shorter times. Although roboticists are exploring how static images could contribute to robot learning[79], text-to-video models, as OpenAI Sora released in early 2024, spark more imagination in not just data augmentation but also physics-free clinical simulation. A preliminary attempt can be found in Endora[80] that simulates clinical endoscopy scenes with explicit modelling of spatial–temporal dynamics, showing more than 39% reduction of Fréchet Video Distance (a noise-based metric for generative models of video[81]; the lower, the better) on Kvasir-Capsule and CholecTriplet data set compared with major video generative adversarial network models. Bora[82] is a diffusion probabilistic model for text-guided (with LLM) biomedical video generation, covering endoscopy, ultrasound, MRI and cellular visualization. Generative models were also used to create realistic endoscopic images[83–86]. Among others, ConStructS[85] and Surgical-CD[86] are, respectively, generative adversarial network-based and multistage latent diffusion methods for generating surgical images through simulated-image to real-image translations. In a cholecystectomy, for example, the involved regions include the abdominal wall, liver, gall bladder, liver ligaments, fat and surgical tools. Almost, all existing work simulates 2D images with the depth unmodelled. This fashion awaits further investigations into other medical imaging modalities.

Meanwhile, with the continuing development of imitation learning and LVMs, it can be expected that a surgical robot could learn about robotic motion from demonstrative video sequences with marked domain shifts (for example, from hand-held surgical tool's motion to robot motion) enabled by LVM-based generative models. At that time, simulators will no longer simply be visualizers but bits of intelligence who can imagine their eye-of-sight. Nevertheless, critiques remain,

concerning the physical realism of the generative contents and their ability to deal with the spatial and temporal complexities of our physical world by only exploiting 2D vision[87]. In 2024, a couple of studies demonstrated how synthetic 3D models can be used as useful visual sources for action learning[88,89]. However, at the current stage, exact kinematic parameters are still paired as part of the learning, making the models inferior to advanced human-liked visual learning. Physics-based simulation techniques can power deformable objects (such as soft tissues) manipulation in RAS. To tackle the real-to-simulation gap caused by inaccurate calibration and unmodelled physical effects, the residual deformation and stiffness parameters can be estimated by optimizing residuals between the simulated particles (obtained from physics-based simulators) and the observed point cloud, acquired by combining the depth estimation and tissue segmentation using SAM from images acquired by endoscopic cameras[90]. The downstream simulation task could directly benefit from rapidly developing key AI methods used by the upstream tasks such as segmentation and registration (Fig. 1).

Diagnosis

Incisive medical diagnostic methods can be expected with the adoption of LVMs. Emerging as one of the fastest proving grounds for real-world applications, diagnosis has garnered the most research interest and the corresponding large models development is the fastest among the four typical downstream tasks (Fig. 2). Diagnosis is defined as 'the identification of the diseases that are most likely to be causing the patient's symptoms, given their medical history'[91]. Doctors figure out the disease (diseases) based on observation, queries or tests before giving an accurate prescription or suitable treatment. Current medical imaging techniques, such as CT and MRI, provide direct means of see-through observations of the anatomy, which substantially improves the diagnosis precision. Despite some methodological failures and translational discrepancy[92,93], AI methods in analysing medical images for better diagnosis have grown exponentially since 2012, for their ability to perform prognosis and detection on par or beyond human physicians[94]. For example, deep-learning approaches can detect diabetic retinopathy, diabetic macular oedema and poor blood glucose control through only external eye photographs[95], a task impossible for human doctors. Additionally, Alzheimer disease, traditionally diagnosed using various tests, including the history of illness, cognitive testing, medical imaging and so on, can now be confirmed with more than 60% confidence by only feeding brain MRIs to convolutional neural networks, previously trained with large brain-imaging cohorts of ageing individuals at risk of developing the disease[96]. Without covering all the examples, we can still conclude that these diagnostic AI models are specifically trained for selected abnormalities, lesions or diseases based on a large amount of image–label pairs of the same kind. As a result, these AI methods often still require a human pre-diagnosis to at least choose which model to use. Using large-scale multi-institutional medical images (such as MIDRC Data) along with standardized and high-quality annotations for training LVMs leads the way to the development of more general image-based diagnostic paradigms. For example, in 2024, the introduction of BiomedGPT[97] provided a generalist vision-language foundation model for diverse biomedical tasks, demonstrating predictive capabilities with a low error rate of 3.8% in question answering and satisfactory performance in generating complex radiology reports, with an error rate of 8.3%. Additionally, its summarization ability earned a preference score nearly equivalent to those of human experts. The downstream diagnostic task could gain from all four upstream counterparts (Fig. 1). As an example, prostate cancer diagnosis in MRI may involve grading or classification, detection and segmentation of prostatic lesions and the registration of multi-parametric MRI[98].

Considering that the quality of X-rays, CT scans and MRIs is more standardized, a wealth of these AI-ready data sets is more readily available compared with the variety found in RAS-related images. For instance, in BiomedGPT, the endoscopy modality constitutes only 1.41% of the pre-trained data set, whereas standardized modalities make up more than 70%. In general, the RAS systems are designed to perceive the surgical site in situ across varying scales, ranging from macroscopic views (for example, using an endoscope) to microscopic perspectives (for example, through optical coherence tomography or intravascular ultrasound images). These informative visual contents, presented with minimal image processing and, crucially, in a dynamic manner that could reflect the physiological status, are invaluable for diagnosis. For example, with the assistance of convolutional neural networks and vision transformers, it is feasible to perform a ternary classification of gastrointestinal tract endoscopic images, efficiently categorizing them into specific diseases such as ulcerative colitis, polyps and esophagitis with high throughput[99]. Despite substantial research efforts and commercial interest, most diagnostic algorithms still fall short of matching human doctors' accuracy in differential diagnosis, where identifying the exact cause among multiple possible causes of a patient's symptoms presents a challenge[100]. Metaphorically speaking, previous AI-based diagnostic methods resemble those of a medical school junior student who can identify conditions based on strong indicators learned from textbooks and lectures. By contrast, an LVM-based approach can possibly mirror the expertise of a senior doctor on specific tasks, offering in-depth explanations that consider a multitude of factors, such as the dynamic biological behaviour of motile tissues and organs that have yet to be explored. This comprehensive understanding is akin to drawing from years of experience analysing an extensive collection of medical data sets from an end-to-end perspective. Furthermore, diagnosis typically results from an interplay of multiple factors, with medical images constituting just one piece of the puzzle. The factors include textual data such as medical histories, biological characteristics and patients' verbal descriptions of symptoms. Therefore, integrating LVMs with language models (that are vision-language models such as CLIP[15]) that could use different types of information rather than solely relying on images is essential to leverage the full spectrum of diagnostic information. However, existing language-vision models, such as GPT-4V(vision), exhibit caution in direct diagnoses[101], likely because of the stringent safeguard mechanisms implemented by the developers. Nevertheless, this result does not necessarily negate the potential of vision-language models in this area.

Robot control

In the realm of RAS, from a technological standpoint, we contend that robotic control stands to be the ultimate beneficiary of recent advancements in LVMs, with a twofold evolution to be addressed, namely, autonomy and adaptability. Physical robots execute precise operations, with motion commands generated from planning policies and real-time environmental perception. Autonomy can occur when navigating the robot to the target sites, undertaking manipulation, or assisting with these procedures. For instance, robot manipulators can be used to perform semi-automatic wound suturing on ex vivo phantoms by encoding real-time visual feedback to a series of predefined action primitives without human intervention[102]. With accurate registration and visual feedback, robots can execute high-precision

surgical procedures, provided that these procedures, such as tooth implant placement[103] and keratoplasty suturing[104], can be standardized into a sequence of quasi-static steps. Although robots surpass human surgeons in terms of motion precision and stability, they often lack the adaptability required to effectively respond to dynamic environments encountered in real-world scenarios. For example, autonomous robotic suturing can be performed on ex vivo soft phantoms using marker-less tracking techniques. However, a series of digital labels are still required for the training data set. In a specific case, the suture plan could be updated in real time when tissue deformation exceeded a threshold of 5 mm, granting the robot system some adaptability to dynamic environments[105]. However, missed attempts were reported, and the system is apparently not yet comparable to a real surgeon.

Human surgeons demonstrate remarkable adaptability across a diverse array of cases, effortlessly handling minor variations from patient to patient as well as significant differences that necessitate distinct surgical decisions. This adaptability stems from surgeons' extensive medical training and years-long hands-on experience in operating rooms. In essence, roboticists, such as surgeons, must navigate a complex landscape, making informed decisions and precise actions to ensure the successful operation of autonomous robots. In robotics, learning from demonstration is a paradigm in which robots acquire new skills by imitating demonstrations without extensive scripting over traditional robot programming. This technique can be meaningful for controlling redundant robotic systems, such as the prevalent surgical continuum robots, to work in irregular intraluminal environments. For example, in 2024, the feasibility of eye-in-hand vision-based transoral navigation learning from simulated demonstrations was reported using a high-DOF endoscopic robot[106]. This result substantially simplifies what would have been a cumbersome process if undertaken through conventional programming on robotic motion control.

Nevertheless, the majority of vision-based learning from demonstration methods continues to necessitate tailored image processing and specific control policies to effectively encode demonstrations for robot learning, although exceptions can be found in imitation learning[107]. The rise of LVMs alters the previous practices, as not only can demonstrations be enriched by synthetic data but also the image processing (upstream tasks) can be conducted with less manual work[108]. Since 2023, a novel category of multimodal models termed vision-language-action (VLA) models has emerged[109], garnering rapidly increasing interests within the research community[110]. This technique is ideal for surgical robots to learn about complicated motions and intensive decision-making in RAS. Generative vision models, such as Sora, have the potential to simulate multiscale surgical procedures in numerous ways. A robot could assimilate a breadth of operational knowledge through visual information far surpassing that of human surgeons, enabling increasing autonomy and intelligence on RAS systems, as depicted in transformer-based RAS-RT[111], through the development of multimodal, multitask, VLA models.

Looking ahead

Despite the success of incorporating LVMs in the upstream tasks of RAS, there is potential for expanding their use in downstream tasks in the attempt of covering the last step in the evolution from 'surgical robots' to 'robotic surgeons' (Fig. 3a). The magnitude of this step is evidently enormous, yet it can be rapidly advanced by a trigger point. Just as, at the beginning of the 2010s, little was known about how AI would impact the surgical robotics ecosystem, in the middle of the 2020s, whether LVM could be the next trigger point is highly discussed in the community. The evident trend of developing larger models with more parameters can be clearly observed by recapping the development history of AI models from 2012 to 2024 (Fig. 3b). Researchers are making a lot of efforts to solidify the performance of LVMs and other foundation models in the upstream RAS tasks. However, within the context of current RAS practice, upstream tasks predominantly aim at providing human surgeons with essential computational information. In the end, it still falls to the surgeons to carry out the remaining downstream tasks.

The development ecology can be advanced by integrating existing techniques for upstream tasks, associating the four major downstream ones and their upstream counterparts (Fig. 1). Each downstream task combines two or more upstream tasks in various possible configurations, involving multidimensional and multimodality data, which notably increases development complexity. For model deployment, it is crucial that models for downstream tasks are robust, accurate and safe, as they directly impact surgical outcomes.

Trends

Specifically, five representative research trends can be identified for the development of surgical vision (and beyond) large models: enhancing data collection, particularly with high-dimensional 3D spatial data; achieving physics-aware surgical AI models (defined as Robophysics); developing surgical LMMs with incorporated medical domain knowledge; proposing explainability techniques to ensure reliability and safety of surgical foundation models; and strengthening the collaborations among multidisciplinary specialists. Large models will bring AI–human collaboration and surgical autonomy to a new level by integrating multimodalities and combining multi-downstream tasks (possibly through unifying task-specific models). Potential examples are production of alerts under risk or uncertainty; automatic surgical data analysis (for example, surgical image understanding, surgical phase detection); and generation of information (in visual or audio formats) to surgeon's questions to assist the decision-making procedure. Although existing 'preliminary' autonomy has been achieved for certain surgical tasks, such as robotic endoscopy[112], viscoelastic tissue debridement[113] and suturing[114–116], most of these procedures have been demonstrated only on phantom tissues or animal models. Large models, with their superior accuracy in upstream tasks, enhanced cognitive abilities crucial for intraoperative decision-making, and generalization capabilities can accelerate the development process towards task autonomy (level of autonomy 2) and conditional autonomy (level of autonomy 3) for more complex surgical procedures and environments[117]. This progress will also facilitate their real clinical translation. Once downstream tasks are effectively managed by LVMs, particularly those with certified diagnostic and interventional capabilities, the technical development of surgical robots with high autonomy (level of autonomy 4) for specific surgeries could possibly become a reality (Fig. 3a).

Data enhancement. Although the used data might not require costly expert labels given the success of self-supervision, the development of large or foundation models still necessitates collecting unprecedented amounts of medical data. Medical imaging data are both scarce and highly valuable, prompting stakeholders to closely guard their access and distribution. Different medical imaging modalities (such as CT, X-ray and ultrasound) in RAS exhibit significant variations in scale and quantity (Fig. 3c). Moreover, unwanted issues, such as risks for patients' privacy or social bias caused by the fact that models might be trained using data only from specific populations,

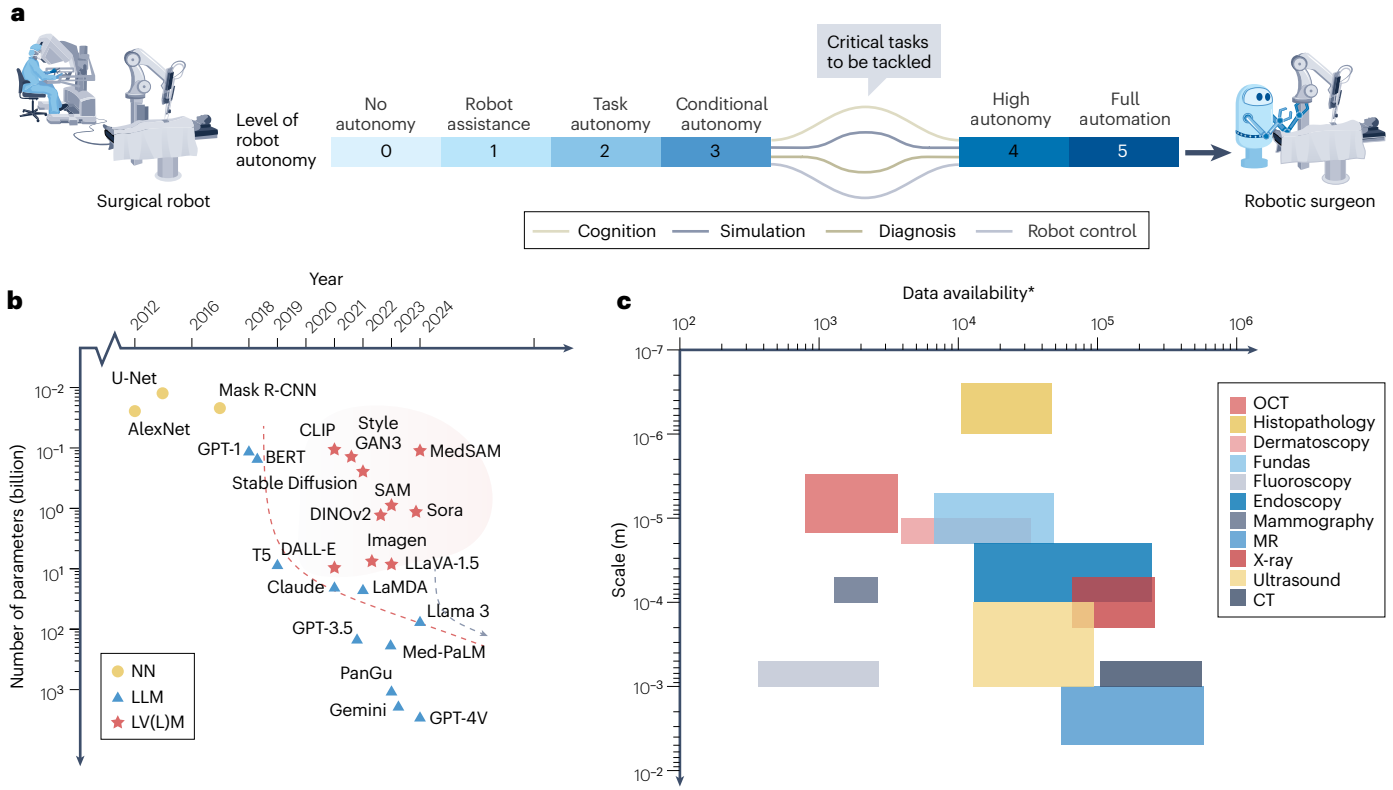

Fig. 3 | The role of large vision models in enhancing surgical robot autonomy. **a**, Innovation from 'surgical robot' to 'robotic surgeon' involves addressing four major downstream tasks: cognition, simulation, diagnosis and robot control. These tasks must be tackled to achieve a high level (level 4) of robot autonomy. **b**, Representative deep-learning-based and large vision models (LVMs) developed from 2012 to 2024. The parameters used by LVMs currently lag those used by large language models (LLMs). The trend indicated that LVMs will increasingly use more parameters, becoming large vision-language models (LV(L)Ms). **c**, Medical image modalities applicable to robot-assisted surgery, with data availability assessed based on the data sets used by state-of-the-art LV(L)Ms[10,97]. MR, magnetic resonance; NN, neural network; OCT, optical coherence tomography.

further complicate real-world medical data collection[118]. Consequently, data synthesis is emerging as a potential way out[119]. Nevertheless, collecting more high-quality medical data (particularly 3D data such as MedShapeNet[120]) is vital for enhancing performance. Collected data should be modality-balanced and class-balanced to develop a foundation model of medical image classification. In addition, more work will and has to be done for 3D RAS tasks, which are fundamental for downstream high-level tasks such as surgical navigation (as in augmented reality-assisted surgery). Several foundation models for 3D data processing have already been developed (such as point cloud segmentation[121] and 3D medical image segmentation[30,122]). Applications vary from typical upstream tasks including object classification, part segmentation and object detection[122–125] to downstream tasks such as 3D question answering, text/image/audio/point-to-mesh generation and multimodal reasoning[126], which however are restricted to the general computer vision field but are evidently useful for planning, visualization and navigation once adapted to surgical scenarios.

Robophysics. Generative AI, such as Sora, might generate physically implausible movements[127]. In this regard, the combination of physics and data-driven techniques is promising, with their respective advantages of interpretability and capability of viable predictions, to prune or regularize clearly unreasonable solutions. For example, the biomechanical constraints in the form of partial differential equations, describing complex plausible deformations in computer-assisted interventions, have been explicitly incorporated into the optimization objective of the upstream multimodal medical image registration task[128]. The three tiers of incorporating physics into surgical AI pipelines contain: (1) the synthesis of feasible data (such as organ deformations) constrained by physical laws (such as elasticity) possibly through the use of physically realistic generative models through residual learning that predicts the residual of measurements away from the physical model's predictions; (2) the physical fusion in which physics is treated as multimodal inputs; and (3) the incorporation of the physics into the overall loss function as a regularization, which ideally should be differentiable[129]. In any case, it is very beneficial in most cases and even a must in specific scenarios for surgical robots to be aware of the physical characteristics of the surrounding environments or tissues to complete certain tasks. For example, soft tissues' stiffness parameters can be estimated online with a fast simulation framework[90], bridging the sim-to-real gap, ensuring deformation to adhere to physical constraints and enabling accurate and automated manipulation of deformable tissue by the robotic surgeon.

Surgical large multimodal models. Multimodal foundation models (such as vision, virtual reality, language, force, haptic and kinematics) to

Glossary

Area under the receiver-operating curve
A single scalar that quantitatively summarizes the performance of one classification model across all classification thresholds.

Contrastive learning
A self-supervised learning technique through maximizing and minimizing agreements between similar and dissimilar pairs in the latent space, respectively.

Dice similarity coefficient
The ratio of intersection of predicted and ground-truth segmented regions in the context of segmentation.

Foundation models
Large-scale artificial intelligence models pretrained on massive amounts of diverse data sets and can be adapted to various downstream tasks.

Knowledge graphs
A graph-based representation of knowledge that describes entities and their relationships.

Large language models
(LLMs). Large artificial intelligence models pretrained on massive amounts of language data (for example, text and radiology reports) and can be applied to numerous language downstream tasks such as question answering and dialogue.

Large vision models
(LVMs). Large artificial intelligence models pretrained on massive amounts of vision data (for example, medical images and surgical videos) and can be applied to numerous vision downstream tasks such as medical image classification.

Segment anything model
(SAM). A segmentation foundation model that exhibits impressive zero-shot performance.

Self-supervised learning
Models that derive supervisory signals directly from unlabelled data.

Vision-language-action (VLA) models
Large multimodal models that process vision and language information and generate robot actions.

Vision-language models
Large multimodal models pretrained with image–text pairs and can be applied to various downstream vision-language tasks (for example, image captioning and visual question answering).

be used for high-level downstream tasks — such as interventional navigation, automatic low-risk decision-making or assistance mechanism — are potentially achievable with the development of LMMs, which might also serve as an important bridge between upstream and downstream tasks. For example, one viable way of automating surgical interventions is to use the LLMs as high-level task planner, which interprets human intentions and decomposes complex surgical tasks into simpler subtasks, according to the perceived environment by LVMs, executable by low-level control policies which themselves could be VLA models[110]. There is no doubt that medical knowledge and expertise in literature, publications, clinical notes, case studies and knowledge graphs, combined with language models, will further enhance disease diagnostic performances and increase surgical autonomy by generating algorithms that can correctly reason step-by-step surgical tasks or subtasks in an automatic or cooperative way. The benefits include enhanced medical reasoning abilities to deal with long-tail problems that contain unseen or usual conditions during surgery. At the same time, the large model itself could do medical reasoning to some extent. Both inputs and outputs of surgical LMMs could be complex and multimodal.

Explainability. Good explainability of the large models and interpretability of the predicted results would help the deployment of AI algorithms on robots with confidence — although not essential — as the aim is to meet the requirement of applications instead of knowing everything. Yet, the fact is that knowing a bit more about what we are doing and why we are doing it helps to develop better models. In addition, much relevant to the model's explainability, surgical foundation models should express their uncertainties with confidence, especially in worse-case scenarios (such as unexpected complications and unseen phenomena during surgery). In other words, surgical AI models need to know when they are unsure or even fail. This is especially important for those adopting multimodal models and when unusual situations are encountered to help the operators determine how much they can trust the algorithm in making decisions and decide when human surgeons should take over the control. In this respect, uncertainty estimation[130–132] for models associated with these upstream tasks is desired. Moreover, the level of confidence in tackling downstream tasks could be further estimated through uncertainty propagation theories in a linear or, most probably, a nonlinear way[133].

Cross-disciplinary collaboration. The continuing development and deployment of surgical LMMs, across the preoperative, intraoperative and postoperative stages, require close collaborations among multidisciplinary specialists, including but not limited to clinicians, radiologists, pathologists, oncologists, AI researchers, roboticists, statisticians and policymakers. For example, it is the responsibility of medical doctors to verify the accuracy of surgical AI models' outputs for downstream tasks. In this regard, a multidisciplinary panel of specialists might be needed to verify complex tasks involving different downstream applications (such as intraoperative decision-making). Additionally, collaborations among clinicians, AI researchers, roboticists, policymakers and regulatory bodies are required to tackle the inevitable challenges concerning the AI models' safety, reliability, ethics and regulatory issues. The opportunities for multidisciplinary collaborations will further increase along with the development of LVMs for RAS. One evident example would be the exponential growth of multimodal research involving clinician notes/radiologist reports (LMMs) and medical images (LVMs), which were rather difficult or even impossible using conventional deep-learning methods. Along the pathway, open-source platforms, such as the Medical Open Network for Artificial Intelligence (MONAI) and the Grand Challenge projects, mitigate geographical and deployment barriers for collaboration.

Challenges

Surgery poses several challenges covering technical, application, ethical and regulatory aspects that need to be tackled along the roadmap of integrating large AI models and therapeutic interventions to achieve the goal of robotic surgeons.

Technical challenges. First, the mistakes, biases, hallucinations and incorrectness of the surgical AI models should be refrained as much as possible. One of the research directions for tackling this challenge is to augment the large models with a knowledge graph[134] or medical expertise[16] and explicit physical laws[129]. In addition, it is still challenging to conduct precise soft-tissue modelling and simulation and reproduce the interaction between the surgical instrument and surrounding organs. In this regard, the specific challenges that have not been fully resolved include but are not limited to the accurate estimation of material properties with soft tissues[90], the biomechanics-constrained

deformable medical image registration[135] and the sim-to-real gap that needs to be compensated within a reasonable computational cost frame. Developing physics-aware AI models could be a solution.

Moreover, although the large-scale medical data set might or will be sizable and of high quality, another issue that should carefully be considered is the inevitable distribution shift problem, in which data distributions in training and deployment are often different from each other. This harmful problem could be alleviated with in-context learning techniques, predominantly observed in LLMs as of the beginning of 2025 (ref. 44), which should be further developed for surgical LVMs and LMMs.

Application challenges. Although we believe that vision will remain the central role of information source in surgical AI, yet as we predict and other researchers have identified[9,111,118], the development of multimodal foundation models for surgery is unstoppable. Despite the benefits of flexible interactions between large models and users, surgical LMMs will require different (even additional specialized and high-end hardware) interfaces to take in and process distinct modalities. To be able to deploy the models properly, changes and shifts in traditional surgical routines and paradigms might thus be necessary.

The heavy costs of unprecedented massive data collection and large model training, practical issues posed by the markedly increasing model size, could be alleviated by large-scale data-sharing efforts and the incorporation of large and small models. For example, the large-scale foundation models for the upstream tasks in RAS, whose outputs or intermediate numeric representations are being leveraged, can serve as the basis for the downstream counterparts that could be smaller specialist models tailored for specific cognitive tasks or organs.

Ethical challenges. The heterogeneous data in surgery include physiological measurements, electronic health records, clinical notes, medical images and so on. This large amount of multimodal data will add more difficulties in overcoming the long-standing challenges in developing and deploying data-centred surgical AI models, including data privacy, security, social bias and regulatory foresight. As a sector of evidence-based medicine, surgery necessitates extensive validations and verification, such as sizable and reliable randomized trials, to enable clinical translation and enhance the general public's acceptance level of surgical innovations. The multimodal and multidimensional inputs and outputs, the ground-truth deficiency, particularly for the advanced downstream RAS tasks, together with the versatility of surgical foundation models, will pose multi-faced in-negligibly pragmatic challenges for relevant regulatory approval.

Regulatory challenges. Large models are thriving in many other fields and evolving on almost weekly basis. However, their widespread use in medicine and surgery remains conservative, as some could involve matters of life and death and some are considered 'not helpful enough' or 'not worth it' in clinical practice. An across-the-board rule for all medical AI methods will discourage the development from tip to toe. Therefore, well-designed regulations that balance restrictions with encouragement are essential. For example, the European Union commenced the world-first Artificial Intelligence Act (EU AI Act) in August 2024. The EU AI Act classifies non-exempt AI applications by their risk of causing harm to categories such as unacceptable, high, limited, minimal risks and general-purpose AI. Notably, Article 13 on 'transparency and provision of information to deployers' mandates a certain level of interpretability for AI systems, which could restrict their use in the European market. Therefore, substantial challenges persist in the development and deployment of LVM-based tools for RAS that are both explainable and trustworthy.

Conclusion

The introduction of LVMs in robotic surgery is promising to warrant ongoing investment and research. This frontier technique has great potential to benefit RAS in many crucial aspects, facilitating the surgical automation process towards conditional autonomy for more procedures and high autonomy for some specific interventions, producing more generalizable, versatile and accurate intelligence algorithms and enhancing the degree of human–robot interaction. Looking forward, we can foresee a vital and inevitable pathway to develop multimodal, downstream-tasks-oriented, high-dimensional and physics-aware large models for RAS.

Published online: 12 May 2025

References

1. Khan, S. et al. Transformers in vision: a survey. *ACM Comput. Surv.* **54**, 1–41 (2022).
2. Krishnan, R., Rajpurkar, P. & Topol, E. J. Self-supervised learning in medicine and healthcare. *Nat. Biomed. Eng.* **6**, 1346–1352 (2022).
3. Qiu, J. et al. Large AI models in health informatics: applications, challenges, and the future. *IEEE J. Biomed. Health Inform.* **27**, 6074–6087 (2023).
4. Firoozi, R. et al. Foundation models in robotics: applications, challenges, and the future. *Int. J. Robot. Res.* https://doi.org/10.1177/02783649241281508 (2024).
5. Dupont, P. E. et al. A decade retrospective of medical robotics research from 2010 to 2020. *Sci. Robot.* **6**, eabi8017 (2021).
6. Yip, M. et al. Artificial intelligence meets medical robotics. *Science* **381**, 141–146 (2023). **This work reviews artificial intelligence techniques for medical robotics.**
7. Fiorini, P., Goldberg, K. Y., Liu, Y. & Taylor, R. H. Concepts and trends in autonomy for robot-assisted surgery. *Proc. IEEE* **110**, 993–1011 (2022).
8. Marcus, H. J. et al. The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring. *Nat. Med.* **30**, 61–75 (2024).
9. Varghese, C., Harrison, E. M., O'Grady, G. & Topol, E. J. Artificial intelligence in surgery. *Nat. Med.* **30**, 1257–1268 (2024).
10. Ma, J. et al. Segment anything in medical images. *Nat. Commun.* **15**, 654 (2024). **This study introduces a foundation model for universal segmentation of a wide spectrum of anatomical structures and lesions across different medical imaging modalities.**
11. Wang, D. et al. A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Sci. Data* **10**, 574 (2023).
12. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
13. Tiu, E. et al. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat. Biomed. Eng.* **6**, 1399–1406 (2022). **This study presents a self-supervised model that performs pathology classification by leveraging image–text pairs of unannotated X-rays and accompanying radiology reports.**
14. Irvin, J. et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proc. AAAI'19: AAAI Conference on Artificial Intelligence* 590–597 (AAAI, 2019).
15. 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.) 8748–8763 (PMLR, 2021). **This study presents contrastive language-image pretraining (CLIP), a multimodal approach that jointly trains an image encoder and a text encoder to predict correct image–text pairs, to learn transferable image representations and to support zero-shot prediction.**
16. Wang, Z., Wu, Z., Agarwal, D. & Sun, J. MedCLIP: contrastive learning from unpaired medical images and text. In *Proc. 2022 Conference on Empirical Methods in Natural Language Processing* (eds Goldberg, Y. et al.) 3876–3887 (Association for Computational Linguistics, 2022).
17. Yue, W. et al. SurgicalSAM: efficient class promptable surgical instrument segmentation. In *Proc. AAAI'2024: AAAI Conference on Artificial Intelligence* (eds Wooldridge, M. et al.) 6890–6898 (AAAI, 2024).
18. Boers, T. G. W. et al. Foundation models in gastrointestinal endoscopic AI: impact of architecture, pre-training approach and data efficiency. *Med. Image Anal.* **98**, 103298 (2024).
19. Caron, M. et al. Emerging properties in self-supervised vision transformers. In *Proc. 2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 9630–9640 (IEEE, 2021).
20. van der Zander, Q. E. et al. Real-time classification of colorectal polyps using artificial intelligence — a prospective pilot study comparing two computer-aided diagnosis systems and one expert endoscopist. *Gastrointest. Endosc.* **95**, AB250–AB251 (2022).

21. Wang, X. et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970–978 (2024).
22. Bulten, W. et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat. Med.* **28**, 154–163 (2022).
23. Sheng, Y., Bano, S., Clarkson, M. J. & Islam, M. Surgical-DeSAM: decoupling SAM for instrument segmentation in robotic surgery. *Int. J. Comput. Assist. Radiol. Surg.* **19**, 1267–1271 (2024).
24. Carion, N. et al. End-to-end object detection with transformers. In *Proc. Computer Vision – ECCV 2020: 16th European Conference* (eds Vedaldi, A. et al.) 213–229 (Springer, 2022).
25. Kirillov, A. et al. Segment Anything. In *Proc. 2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 4015–4026 (IEEE, 2023).
26. Fan, D.-P. et al. Pranet: parallel reverse attention network for polyp segmentation. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference* (eds Martel, A. L. et al.) 263–273 (Springer, 2020).
27. Li, L. H. et al. Grounded language-image pre-training. In *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10955–10965 (IEEE, 2022).
28. Qin, Z., Yi, H. H., Lao, Q. & Li, K. Medical image understanding with pretrained vision language models: a comprehensive study. In *Proc. 11th International Conference on Learning Representations* (ICLR, 2023).
29. Zhao, T. et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nat. Methods* **22**, 166–176 (2025).
This work proposes a foundation model for joint segmentation, detection and recognition tasks of biomedical objects along with a large biomedical data set across nine modalities.
30. Lei, W., Xu, W., Li, K., Zhang, X. & Zhang, S. MedLSAM: localize and segment anything model for 3D CT images. *Med. Image Anal.* **99**, 103370 (2025).
31. Baumgartner, M., Jäger, P. F., Isensee, F. & Maier-Hein, K. H. nnDetection: a self-configuring method for medical object detection. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference* (eds de Bruijne, M. et al.) 530–539 (Springer, 2021).
32. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. In *Proc. 2017 IEEE International Conference on Computer Vision (ICCV)* 2980–2988 (IEEE, 2017).
33. Mazurowski, M. A. et al. Segment anything model for medical image analysis: an experimental study. *Med. Image Anal.* **89**, 102918 (2023).
34. Huang, Y. et al. Segment anything model for medical images? *Med. Image Anal.* **92**, 103061 (2024).
This work extensively validated the performance and limitations of segment anything models in medical scenarios.
35. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211 (2021).
36. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. Computer Vision – ECCV 2018: 15th European Conference* (eds Ferrari, V. et al.) 833–851 (Springer, 2018).
37. Wang, A., Islam, M., Xu, M., Zhang, Y. & Ren, H. SAM meets robotic surgery: an empirical study on generalization, robustness and adaptation. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops: ISIC 2023, Care-AI 2023, MedAGI 2023, DeCaF 2023, Held in Conjunction with MICCAI 2023* (eds Celebi, M. E. et al.) 234–244 (Springer, 2023).
This is the very first work to examine and analyse the robustness and zero-shot generalizability of the segment anything model in the field of robotic surgery.
38. Hu, E. J. et al. LoRA: low-rank adaptation of large language models. In *Proc. 10th International Conference on Learning Representations* (ICLR, 2022).
39. Paranjape, J. N., Nair, N. G., Sikder, S., Vedula, S. S. & Patel, V. M. AdaptiveSAM: towards efficient tuning of SAM for surgical scene segmentation. In *Proc. Medical Image Understanding and Analysis: 28th Annual Conference, MIUA 2024* (eds Yap, M. H. et al.) 187–201 (Springer, 2024).
40. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., & Wu, J. Medical SAM 2: segment medical images as video via Segment Anything Model 2. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2408.00874 (2024).
41. Ravi, N. et al. SAM 2: segment anything in images and videos. Preprint at https://arxiv.org/abs/2408.00714 (2024).
42. Chen, C. et al. MA-SAM: modality-agnostic SAM adaptation for 3D medical image segmentation. *Med. Image Anal.* **98**, 103310 (2024).
43. Wu, J. et al. MedSegDiff: medical image segmentation with diffusion probabilistic model. In *Proc. Medical Imaging with Deep Learning (MIDL 2023)* (eds Oguz, I. et al.) 1623–1639 (PMLR, 2024).
44. Landman, B. et al. MICCAI 2015: multi-atlas labeling beyond the cranial vault - workshop and challenge. *Synapse* https://doi.org/10.7303/syn3193805 (2015).
45. Cui, B., Islam, M., Bai, L. & Ren, H. Surgical-DINO: adapter learning of foundation model for depth estimation in endoscopic surgery. *Int. J. Comput. Assist. Radiol. Surg.* **16**, 1013–1020 (2024).
46. Oquab, M. et al. DINOv2: learning robust visual features without supervision. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=a68SUt6zFt (2024).
47. Yang, L. et al. Depth Anything: unleashing the power of large-scale unlabeled data. In *Proc. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10371–10381 (IEEE, 2024).
48. Han, J. J., Acar, A., Henry, C. & Wu, J. Y. Depth anything in medical images: a comparative study. Preprint at https://arxiv.org/abs/2401.16600 (2024).
49. Lou, A., Li, Y., Zhang, Y. & Noble, J. Surgical depth anything: depth estimation for surgical scenes using foundation models. Preprint at https://arxiv.org/abs/2410.07434 (2024).
50. Fu, Y. et al. DeepReg: a deep learning toolkit for medical image registration. *J. Open Source Softw.* **5**, 2705 (2020).
51. Chen, J. et al. A survey on deep learning in medical image registration: new technologies, uncertainty, evaluation metrics, and beyond. *Med. Image Anal.* **100**, 103385 (2025).
52. Song, X., Xu, X. & Yan, P. DINO-Reg: general purpose image encoder for training-free multi-modal deformable medical image registration. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference* (eds Linguraru, M. G. et al.) 608–617 (Springer, 2024).
53. Tian, L. et al. uniGradICON: a foundation model for medical image registration. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference* (eds Linguraru, M. G. et al.) 749–760 (Springer, 2024).
This paper presents the first foundation model uniGradICON for medical image registration, demonstrating great performances across multiple data sets and zero-shot capabilities for new registration tasks.
54. Wang, S. et al. The use of three-dimensional visualization techniques for prostate procedures: a systematic review. *Eur. Urol. Focus.* **7**, 1274–1286 (2021).
55. Min, Z. et al. Non-rigid medical image registration using physics-informed neural networks. In *Proc. Information Processing in Medical Imaging: 28th International Conference, IPMI 2023* (eds Frangi, A. et al.) 601–613 (Springer, 2023).
This study presents a biomechanically constrained medical image registration approach using physics-informed neural networks.
56. Min, Z. et al. Biomechanics-informed non-rigid medical image registration and its inverse material property estimation with linear and nonlinear elasticity. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference* (eds Linguraru, M. G. et al.) 564–574 (Springer, 2024).
57. Ahdoot, M. et al. MRI-targeted, systematic, and combined biopsy for prostate cancer diagnosis. *N. Engl. J. Med.* **382**, 917–928 (2020).
58. Demir, B. et al. MultiGradICON: a foundation model for multimodal medical image registration. In *Proc. Biomedical Image Registration: 11th International Workshop, WBIR 2024, Held in Conjunction with MICCAI 2024* (eds Modat, M. et al.) 3–18 (Springer, 2024).
59. Huang, S. et al. SAMReg: SAM-enabled image registration with ROI-based correspondence. Preprint at https://arxiv.org/abs/2410.14083 (2024).
60. Modat, M. et al. Fast free-form deformation using graphics processing units. *Comput. Methods Prog. Biomed.* **98**, 278–284 (2010).
61. Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. Voxelmorph: a learning framework for deformable medical image registration. *IEEE Trans. Med. Imaging* **38**, 1788–1800 (2019).
62. Hu, Y. et al. Weakly-supervised convolutional neural networks for multimodal image registration. *Med. Image Anal.* **49**, 1–13 (2018).
63. Schmidt, A., Mohareri, O., DiMaio, S., Yip, M. C. & Salcudean, S. E. Tracking and mapping in medical computer vision: a review. *Med. Image Anal.* **94**, 103131 (2024).
64. Hong, L. et al. OneTracker: unifying visual object tracking with foundation models and efficient tuning. In *Proc. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 19079–19091 (IEEE, 2024).
65. Ding, X. et al. Less is more: surgical phase recognition from timestamp supervision. *IEEE Trans. Med. Imaging* **42**, 1897–1910 (2023).
66. Liu, Y. et al. SkiT: a fast key information video transformer for online surgical phase recognition. In *Proc. 2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 21017–21027 (IEEE, 2023).
67. Bai, L., Islam, M. & Ren, H. CAT-ViL: co-attention gated vision-language embedding for visual question localized-answering in robotic surgery. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference* (eds Greenspan, H. et al.) 397–407 (Springer, 2023).
68. Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and time. *Nat. Neurosci.* **17**, 455–462 (2014).
69. Seenivasan, L., Islam, M., Krishna, A. K. & Ren, H. Surgical-VQA: visual question answering in surgical scenes using transformer. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference* (eds Wang, L. et al.) 33–43 (Springer, 2022).
70. Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J. & Chang, K.-W. Visualbert: a simple and performant baseline for vision and language. Preprint at https://arxiv.org/abs/1908.03557 (2019).
71. Bai, L., Islam, M., Seenivasan, L. & Ren, H. Surgical-VQLA: transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. In *Proc. 2023 IEEE International Conference on Robotics and Automation (ICRA)* 6859–6865 (IEEE, 2023).
72. Schmidgall, S., Kim, J. W., Jopling, J. & Krieger, A. General surgery vision transformer: a video pre-trained foundation model for general surgery. Preprint at https://arxiv.org/abs/2403.05949 (2024).
73. Schmidgall, S., Cho, J., Zakka, C. & Hiesinger, W. GP-VLS: a general-purpose vision language model for surgery. Preprint at https://arxiv.org/abs/2407.19305 (2024).
This paper presents GP-VLS, a general-purpose vision language model for surgery, that understands both medical and surgical knowledge and tackles surgical visual question answering problems such as phase and triplet action recognition.

74. Kiyasseh, D. et al. A vision transformer for decoding surgeon activity from surgical videos. *Nat. Biomed. Eng.* **7**, 780–796 (2023).
This study introduces the unified surgical AI system (SAIS) leveraging a vision transformer and supervised contrastive learning, to decode subphase recognition, gesture classification and skill assessment from videos collected during robotic surgeries.
75. Yi, X., Walia, E. & Babyn, P. Generative adversarial network in medical imaging: a review. *Med. Image Anal.* **58**, 101552 (2019).
76. Kazerouni, A. et al. Diffusion models in medical imaging: a comprehensive survey. *Med. Image Anal.* **88**, 102846 (2023).
77. Xu, M., Islam, M., Bai, L. & Ren, H. Privacy-preserving synthetic continual semantic segmentation for robotic surgery. *IEEE Trans. Med. Imaging* **43**, 2291–2302 (2024).
78. Shumailov, I. et al. AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).
79. Kapelyukh, I., Vosylius, V. & Johns, E. Dall-E-Bot: introducing web-scale diffusion models to robotics. *IEEE Robot. Autom. Lett.* **8**, 3956–3963 (2023).
80. Li, C. et al. Endora: video generation models as endoscopy simulators. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference* (eds Linguraru, M. G. et al.) 230–240 (Springer, 2024).
This study introduces the medical video generation framework Endora that achieves high-fidelity endoscopy simulations, by utilizing video transformer for spatial–temporal modelling and 2D vision foundation models for feature extraction.
81. Unterthiner, T. et al. Towards accurate generative models of video: a new metric & challenges. Preprint at https://arxiv.org/abs/1812.01717 (2018).
82. Sun, W. et al. Bora: biomedical generalist video generation model. Preprint at https://arxiv.org/abs/2407.08944 (2024).
83. Kaleta, J., Dall'Alba, D., Płotka, S. & Korzeniowski, P. Minimal data requirement for realistic endoscopic image generation with stable diffusion. *Int. J. Comput. Assist. Radiol. Surg.* **19**, 531–539 (2024).
84. Venkatesh, D. K., Rivoir, D., Pfeiffer, M., Kolbinger, F. & Speidel, S. Synthesizing multi-class surgical datasets with anatomy-aware diffusion models. Preprint at https://arxiv.org/abs/2410.07753 (2024).
85. Venkatesh, D. K. et al. Exploring semantic consistency in unpaired image translation to generate data for surgical applications. *Int. J. Comput. Assist. Radiol. Surg.* **19**, 985–993 (2024).
86. Venkatesh, D. K., Rivoir, D., Pfeiffer, M. & Speidel, S. SurgicaL-CD: generating surgical images via unpaired image translation with latent consistency diffusion models. Preprint at https://arxiv.org/abs/2408.09822 (2024).
87. Liu, Y. et al. Sora: a review on background, technology, limitations, and opportunities of large vision models. Preprint at https://arxiv.org/abs/2402.17177 (2024).
88. Ng, C., Gao, H., Ren, T.-A., Lai, J. & Ren, H. Navigation of tendon-driven flexible robotic endoscope through deep reinforcement learning. In *Proc. 2024 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)* 134–139 (IEEE, 2024).
89. Moghani, M. et al. SuFIA: language-guided augmented dexterity for robotic surgical assistants. In *Proc. 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 6969–6976 (IEEE, 2024).
90. Liang, X. et al. Real-to-sim deformable object manipulation: optimizing physics models with residual mappings for robotic surgery. In *Proc. 2024 IEEE International Conference on Robotics and Automation (ICRA)* 15471–15477 (IEEE, 2024).
91. Richens, J. G., Lee, C. M. & Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat. Commun.* **11**, 3923 (2020).
92. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digit. Med.* **5**, 48 (2022).
93. Markowetz, F. All models are wrong and yours are useless: making clinical prediction models impactful for patients. *npj Precis. Oncol.* **8**, 54 (2024).
This article points out problems with existing medical models that are not applicable to practical medical applications.
94. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
95. Babenko, B. et al. Detection of signs of disease in external photographs of the eyes via deep learning. *Nat. Biomed. Eng.* **6**, 1370–1383 (2022).
96. Wen, J. et al. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. *Med. Image Anal.* **63**, 101694 (2020).
97. Zhang, K. et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nat. Med.* **30**, 3129–3141 (2024).
This work presents a generalist lightweight vision-language foundation model that can perform versatile biomedical tasks, such as disease diagnosis, report generation and summarization.
98. Min, Z. et al. Segmentation versus detection: development and evaluation of deep learning models for prostate imaging reporting and data system lesions localisation on bi-parametric prostate magnetic resonance imaging. *CAAI Trans. Intell. Technol.* https://doi.org/10.1049/cit2.12318 (2024).
99. Wu, S. et al. High-speed and accurate diagnosis of gastrointestinal disease: learning on endoscopy images using lightweight transformer with local feature attention. *Bioengineering* **10**, 1416 (2023).
100. Semigran, H. L., Levine, D. M., Nundy, S. & Mehrotra, A. Comparison of physician and computer diagnostic accuracy. *JAMA Intern. Med.* **176**, 1860–1861 (2016).
101. Wu, C. et al. Can GPT-4v(ision) serve medical applications? Case studies on GPT-4v for multimodal medical diagnosis. Preprint at https://arxiv.org/abs/2310.09909 (2023).
102. Lu, B., Chu, H. K., Huang, K. & Cheng, L. Vision-based surgical suture looping through trajectory planning for wound suturing. *IEEE Trans. Autom. Sci. Eng.* **16**, 542–556 (2018).
This work introduces a dynamic motion planning approach for coordinated motions of laparoscopic robotic arms to enable a higher level of dexterity and optimal workspace towards the automation of surgical knot tying.
103. Yang, S. et al. Accuracy of autonomous robotic surgery for single-tooth implant placement: a case series. *J. Dent.* **132**, 104451 (2023).
104. Feng, X., Zhang, X., Shi, X. & Li, L. AIRS: autonomous intraoperative robotic suturing based on surgeon-like operation and path quantification in keratoplasty. *IEEE Trans. Ind. Electron.* **71**, 11115–11124 (2024).
105. Kam, M. et al. Autonomous system for vaginal cuff closure via model-based planning and markerless tracking techniques. *IEEE Robot. Autom. Lett.* **8**, 3916–3923 (2023).
106. Lai, J. et al. Sim-to-real transfer of soft robotic navigation strategies that learns from the virtual eye-in-hand vision. *IEEE Trans. Ind. Inform.* **20**, 2365–2377 (2024).
107. Kim, J. W. et al. Surgical robot transformer (SRT): imitation learning for surgical tasks. In *Proc. 8th Conference on Robot Learning* (eds Agrawal, P. et al.) 130–144 (PMLR, 2025).
108. Zhu, X. et al. Diff-LfD: contact-aware model-based learning from visual demonstration for robotic manipulation via differentiable physics-based simulation and rendering. In *Proc. 7th Conference on Robot Learning* (eds Tan, J. et al.) 499–512 (PMLR, 2023).
109. Zitkovich, B. et al. RT-2: vision-language-action models transfer web knowledge to robotic control. In *Proc. 7th Conference on Robot Learning* (eds Tan, J. et al.) 2165–2183 (PMLR, 2023).
110. Ma, Y., Song, Z., Zhuang, Y., Hao, J. & King, I. A survey on vision-language-action models for embodied AI. Preprint at https://arxiv.org/abs/2405.14093 (2024).
111. Schmidgall, S., Kim, J. W., Kuntz, A., Ghazi, A. E. & Krieger, A. General-purpose foundation models for increased autonomy in robot-assisted surgery. *Nat. Mach. Intell.* **6**, 1275–1283 (2024).
This work introduces a conceptual path for enhancing surgical robot autonomy by developing a multimodal, multitask, vision-language-action model.
112. Wijsman, P. J. M. et al. First experience with THE AUTOLAP™ SYSTEM: an image-based robotic camera steering device. *Surg. Endosc.* **32**, 2560–2566 (2018).
113. Murali, A. et al. Learning by observation for surgical subtasks: multilateral cutting of 3D viscoelastic and 2D orthotropic tissue phantoms. In *Proc. 2015 IEEE International Conference on Robotics and Automation* (*ICRA*) 1202–1209 (IEEE, 2015).
114. Chiu, Z.-Y., Liao, A. Z., Richter, F., Johnson, B. & Yip, M. C. Markerless suture needle 6D pose tracking with robust uncertainty estimation for autonomous minimally invasive robotic surgery. In *Proc. 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*) 5286–5292 (IEEE, 2022).
115. Sen, S. et al. Automating multi-throw multilateral surgical suturing with a mechanical needle guide and sequential convex optimization. In *Proc. 2016 IEEE International Conference on Robotics and Automation* (*ICRA*) 4178–4185 (IEEE, 2016).
116. Saeidi, H. et al. Autonomous robotic laparoscopic surgery for intestinal anastomosis. *Sci. Robot.* **7**, eabj2908 (2022).
117. Yang, G.-Z. et al. Medical robotics — regulatory, ethical, and legal considerations for increasing levels of autonomy. *Sci. Robot.* **2**, eaam8638 (2017).
This work presents an analysis on the regulatory, ethical and legal barriers imposed on medical robots.
118. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
119. Frangi, A. F., Tsaftaris, S. A. & Prince, J. L. Simulation and synthesis in medical imaging. *IEEE Trans. Med. Imaging* **37**, 673–679 (2018).
120. Li, J. et al. MedShapeNet – a large-scale dataset of 3D medical shapes for computer vision. *Biomed. Eng. Biomed. Tech.* **70**, 71–90 (2025).
121. Liu, Y. et al. Segment any point cloud sequences by distilling vision foundation models. In *Proc. 37th Conference on Neural Information Processing Systems* (eds Oh, A. et al.) 37193–37229 (NeurIPS, 2023).
122. Wang, H. et al. SAM-Med3D: towards general-purpose segmentation models for volumetric medical images. Preprint at https://arxiv.org/abs/2310.15161 (2023).
123. Pang, Y. et al. Masked autoencoders for point cloud self-supervised learning. In *Proc. Computer Vision – ECCV 2022: 17th European Conference* (eds Avidan, S. et al.) 604–621 (Springer, 2022).
124. Yu, X. et al. Point-BERT: Pre-training 3D point cloud transformers with masked point modeling. In *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 19291–19300 (IEEE, 2022).
125. Zhang, R. et al. Point-M2AE: multi-scale masked autoencoders for hierarchical point cloud pre-training. In *Proc. 36th Conference on Neural Information Processing Systems* (eds Koyejo, S. et al.) 27061–27074 (NeurIPS, 2022).
126. Guo, Z. et al. Point-Bind & Point-LLM: aligning point cloud with multi-modality for 3D understanding, generation, and instruction following. Preprint at https://arxiv.org/abs/2309.00615 (2023).
127. Waisberg, E., Ong, J., Masalkhi, M. & Lee, A. G. Concerns with OpenAI's Sora in medicine. *Ann. Biomed. Eng.* **52**, 1932–1934 (2024).
128. López, P. A., Mella, H., Uribe, S., Hurtado, D. E. & Costabal, F. S. WarpPINN: cine-MR image registration with physics-informed neural networks. *Med. Image Anal.* **89**, 102925 (2023).
129. Kadambi, A., de Melo, C., Hsieh, C.-J., Srivastava, M. & Soatto, S. Incorporating physics into data-driven computer vision. *Nat. Mach. Intell.* **5**, 572–580 (2023).
130. Chen, A. et al. Modeling and understanding uncertainty in medical image classification. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference* (eds Linguraru, M. G. et al.) 557–567 (Springer, 2024).

131. Deng, G. et al. SAM-U: multi-box prompts triggered uncertainty estimation for reliable SAM in medical image. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops: MTSAIL 2023, LEAF 2023, AI4Treat 2023, MMMI 2023, REMIA 2023, Held in Conjunction with MICCAI 2023* (eds Woo, J. et al.) 368–377 (Springer, 2023).
132. Zhang, X. et al. Heteroscedastic uncertainty estimation framework for unsupervised registration. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference* (eds Linguraru, M. G. et al.) 651–661 (Springer, 2024).
133. Gawlikowski, J. et al. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **56**, 1513–1589 (2023).
134. Gilbert, S., Kather, J. N. & Hogan, A. Augmented non-hallucinating large language models as medical information curators. *npj Digit. Med.* **7**, 100 (2024).
135. Lin, S. et al. SuPerPM: a surgical perception framework based on deep point matching learned from physical constrained simulation data. In *Proc. 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 12780–12786 (IEEE, 2024).
136. Seenivasan, L., Islam, M., Kannan, G. & Ren, H. SurgicalGPT: end-to-end language-vision GPT for visual question answering in surgery. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference* (eds Greenspan, H. et al.) 281–290 (Springer, 2023).
This work proposes a language-vision GPT model for visual question answering tasks in surgical scenarios.
137. Cui, B., Islam, M., Bai, L., Wang, A. & Ren, H. EndoDAC: efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference* (eds Linguraru, M. G. et al.) 208–218 (Springer, 2024).
This work designs dynamic vector-based low-rank adaptation and intrinsic parameter estimator head to adapt depth foundation models to surgical scenes with only surgical videos.
138. Rau, A. et al. SimCol3D — 3D reconstruction during colonoscopy challenge. *Med. Image Anal.* **96**, 103195 (2024).
139. Nwoye, C. I. et al. Cholectriplet2021: a benchmark challenge for surgical action triplet recognition. *Med. Image Anal.* **86**, 102803 (2023).
140. Xu, M., Islam, M. & Ren, H. Rethinking surgical captioning: end-to-end window-based MLP transformer using patches. In *Proc. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference* (eds Wang, L. et al.) 376–386 (Springer, 2022).
141. Wang, S. et al. Interactive computer-aided diagnosis on medical image using large language models. *Commun. Eng.* **3**, 133 (2024).
142. Auloge, P. et al. Augmented reality and artificial intelligence-based navigation during percutaneous vertebroplasty: a pilot randomised clinical trial. *Eur. Spine J.* **29**, 1580–1589 (2020).

Acknowledgements
This work was supported in part by the National Natural Science Foundation of China (NSFC) under grants 62303275, 62403402 and 22IAA01849; in part by Jinan Municipal Bureau of Science and Technology under Grant 202333011; in part by the Hong Kong Research Grants Council (RGC) under grants CRF C4026-21G, RIF R4020-22, GRF 14203323 and 14216022; in part by the NSFC/RGC Joint Research Scheme under grant N_CUHK420/22; and in part by the CUHK Direct Grant for Research under grant 4055213.

Author contributions
H.R. conceived and initiated the project. Z.M. and J.L. researched data for the article and wrote the manuscript. All the authors contributed to the discussion of the content and revised/edited the manuscript.

Competing interests
The authors declare that they have no competing interests.

Additional information
Peer review information *Nature Reviews Electrical Engineering* thanks Adam Schmidt and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Related links
EU AI Act: https://artificialintelligenceact.eu/
Grand Challenge: https://grand-challenge.org/
Imagen: https://deepmind.google/technologies/imagen-2/
Medical Open Network for Artificial Intelligence (MONAI): https://monai.io/
MIDRC Data: https://www.midrc.org/midrc-data
Sora: https://openai.com/index/sora/
Stable Diffusion: https://stability.ai/