Towards Next-Generation Healthcare: A Survey of Medical Embodied AI for Perception, Decision-Making, and Action

First Author[1,2*], Second Author[2,3†] and Third Author[1,2†]

[1*]Department, Organization, Street, City, 100190, State, Country.
[2]Department, Organization, Street, City, 10587, State, Country.
[3]Department, Organization, Street, City, 610101, State, Country.

*Corresponding author(s). E-mail(s): iauthor@gmail.com;
Contributing authors: iiauthor@gmail.com; iiiauthor@gmail.com;
†These authors contributed equally to this work.

Abstract

Foundation models have demonstrated impressive performance in enhancing healthcare efficiency. However, their limited ability to perceive and interact with the physical world significantly constrains their utility in real-world clinical workflows. Recently, embodied artificial intelligence (AI) provides a promising physical-interactive paradigm for intelligent healthcare by integrating perception, decision-making, and action within a closed-loop system. Nevertheless, the exploration of embodied AI for healthcare is still in its infancy. To support these advances, this review systematically surveys the key components of embodied AI, focusing on the integration of perception, decision-making, and action. Additionally, we present a comprehensive overview of representative medical applications, relevant datasets, major challenges in clinical practice, and further discuss the key directions for future research in this emerging field. The associated project can be found at XXXX.

Keywords: Embodied Artificial Intelligence, Healthcare, Embodied Perception, Embodied Decision-Making, Embodied Action

1 Introduction

The widespread adoption of artificial intelligence (AI) technologies in the medical field has significantly improved the efficiency and accuracy of clinical diagnosis [1]. For example, Convolutional Neural Networks (CNNs) have achieved outstanding performance in tasks such as disease classification [2] and lesion segmentation [3]. In recent years, Large Language Models (LLMs) and their multimodal extensions have been increasingly applied to medical report generation and clinical decision support, demonstrating remarkable capabilities in language comprehension and generation [4–7]. However, the above approaches remain confined to a "perception and decision" paradigm that relies primarily on static data and lacks the capacity for physical interaction within real-world medical environments, which disconnects them from clinical practice and limits their applicability in realistic scenarios.

In contrast, Embodied Artificial Intelligence (Embodied AI) possesses the ability for perception, decision-making, and action within physical environments, which has attracted increasing research attention in recent years (Fig.1a) and has opened up new avenues for the development of medical artificial intelligence [8, 9]. Currently, embodied AI has been applied in a range of medical tasks, such as surgical robotics [10, 11], surgical navigation [12–17], and rehabilitation assistance [18–20]. For instance, surgical robots equipped with visual and tactile sensing, which enable highly precise operations in minimally invasive procedures, can significantly enhance surgical efficiency and safety [21, 22]. With the integration of intelligent perception and decision-making, rehabilitation systems can dynamically tailoring treatment strategies to individual functional states [23–26]. Mobile robotic platforms help mitigate resource constraints in healthcare systems by enabling the autonomous transport of medical supplies and continuous monitoring of clinical environments [27–30]. These examples highlight the potential advantages of embodied AI in complex and dynamic clinical scenarios. Despite its potential, the deployment of embodied AI in medicine faces several critical challenges. In embodied perception, limitations stem from insufficient training data and variability in sensory inputs. In embodied decision-making, the complexity of medical reasoning and the need to model uncertainty pose significant obstacles. In embodied action, high-precision control remains highly sensitive to errors.

Given the significance of embodied AI in advancing medical artificial intelligence and the challenges it entails, a systematic review of this field is both timely and essential. While existing studies have examined embodied AI from the perspectives of application domains and specific methodologies [31–36], only one review to date has focused specifically on medical applications, and it addresses clinical use cases without thoroughly analyzing the underlying technical foundations [37]. Moreover, relevant medical datasets that support the development and evaluation of embodied AI have largely been overlooked. To address these gaps, this paper provides a comprehensive survey of embodied AI in medicine, examining its developmental background, analyzing key enabling technologies, and summarizing representative applications and outstanding challenges, with the aim of offering a solid theoretical foundation and clear guidance for future research in this rapidly evolving domain.

The remainder of this survey is organized as follows. Section 2 reviews the development and core components of Embodied AI. Section 3 examines its applications in medicine, while Section 4 introduces relevant datasets and benchmarks for medical Embodied AI. Section 5 discusses the key challenges and forward-looking perspectives in this field, and Section 6 concludes by summarizing the main insights and highlighting the implications of medical Embodied AI for the future of intelligent, interactive healthcare systems.

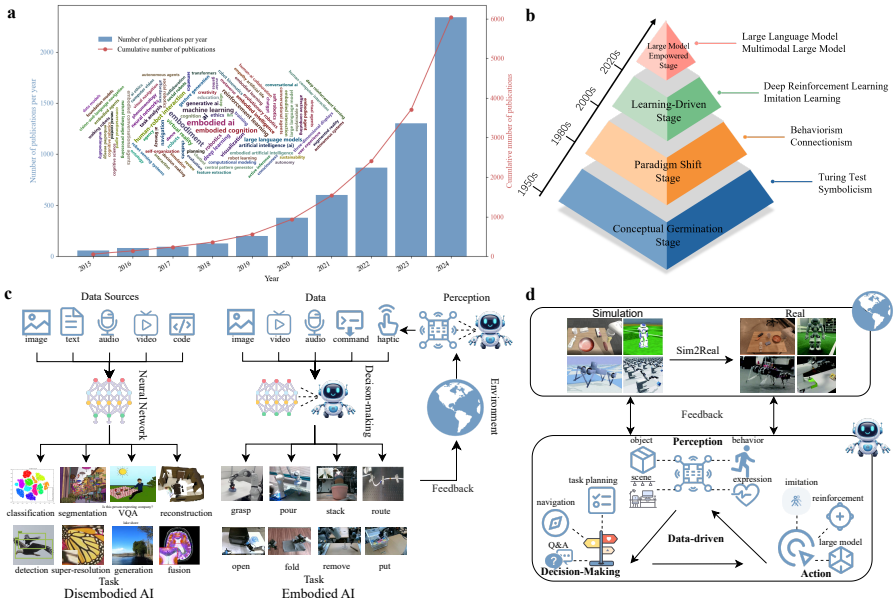

Fig. 1 Foundations of embodied AI. a, Publication volume, temporal trends over the past decade, and representative keywords related to embodied intelligence. The statistics are obtained from Google Scholar using "embodied AI" as the search query. b, The four developmental stages of embodied intelligence, namely the Conceptual Germination Stage, Paradigm Shift Stage, Learning-Driven Stage, and Large Model Empowered Stage. c, A comparison between disembodied intelligence and embodied intelligence. Unlike its disembodied counterpart, embodied intelligence is distinguished by its inherent ability to interact with the environment. d, Core components of embodied intelligence. At the macroscopic level, it consists of agents and their environments; at the technical level, it encompasses embodied perception, embodied decision-making, and embodied action.

2 Embodied AI

As illustrated in Fig.1b, embodied AI has evolved through four stages. The Conceptual Germination Stage in the 1950s, exemplified by the Turing Test and symbolic AI, laid the foundations of artificial intelligence [38]. The Paradigm Shift Stage (1980s–2000s) advanced learning mechanisms and neural networks through behaviorism and connectionism, followed by the Learning-Driven Stage (2000s–2020s), where

deep reinforcement and imitation learning enabled autonomous decision-making. In the Large Model Empowered Stage of the 2020s, large language and multimodal models enhance perceptual, cognitive, and interactive capacities, highlighting the limitations of conventional, disembodied AI and the need for systems that can perceive and act in physical environments.

Conventional expert systems and language models rely primarily on symbolic reasoning and lack direct interaction with the physical environment (Fig.1c), which limits their adaptability, real-time responsiveness, and ability to generalize across tasks. In contrast, embodied AI enables agents to perceive, learn, and interact effectively. As shown in Fig.1d, embodied AI are typically organized around three interrelated components: embodied perception, embodied decision-making, and embodied action, which together form a closed-loop framework with the environment [39–41]. Embodied perception supports multimodal understanding of objects, scenes, and behaviors. Embodied decision-making facilitates planning, navigation, and reasoning using rule-based methods or large models. Embodied action executes strategies through imitation or reinforcement learning, enabling autonomous behavior and interaction (Table 1). In addition, systems leverage Sim2Real transfer to generalize knowledge acquired in simulation to real-world environments [42–45].

2.1 Embodied Perception

Embodied perception forms the foundation for an agent's understanding and adaptation to its environment by extracting meaningful information from multimodal inputs [90, 91]. By processing signals such as images, speech, depth, and touch, agents can build a comprehensive representation of the environment that supports a wide range of tasks, including planning, navigation, interaction, and question answering [92, 93]. Depending on the type of information, embodied perception is generally categorized into four domains: object, scene, behavior, and expression perception.

2.1.1 Object Perception

Object perception is one of the most fundamental and critical tasks in embodied AI, focusing on recognizing, localizing, and representing objects in complex environments through sensory inputs such as RGB images, depth data, and LiDAR signals. Its central goal is to derive semantically meaningful and spatially precise object information, enabling embodied systems to understand and interact with dynamic scenes more effectively [94] [95].

Object perception encompasses a set of core visual understanding tasks, including classification for identifying object categories [46–48], object detection for localizing instances within a scene [49, 50], and semantic or instance segmentation for delineating pixel-level object boundaries [51]. It further extends to 3D perception, which reconstructs geometric structure and estimates object pose from RGB-D data or point clouds [52, 53], as well as open-vocabulary object recognition, which leverages language supervision to achieve category-level generalization beyond fixed training taxonomies [54, 55]. State-of-the-art methods predominantly rely on CNN- or Transformer-based feature extractors, while multimodal fusion strategies are increasingly employed to

Table 1 Overview of the core components, their respective functions, sub-directions, and key tasks in embodied AI.

Components	Function	Sub-Direction	Core Tasks/Method
Embodied Perception	Provides multimodal understanding of the environment.	Object Perception	Object classification [46–48]; object detection [49, 50]; segmentation [51]; 3D object detection [52, 53]; open-vocabulary object recognition [54, 55].
		Scene Perception	Scene classification [56]; spatial semantic segmentation [57]; topological mapping [58, 59]; panoramic perception [60, 61].
		Behavior Perception	Action recognition citebib81,bib82,bib85,bib86; action prediction [62, 63].
		Expression Perception	Speech perception [64]; facial expression recognition [65]; gesture recognition [66, 67]; multimodal intention recognition [68].
Embodied Decision-Making	Converts perception into adaptive strategies.	Task Planning	Symbolic planning [69, 70]; learning-based planning [71]; hybrid planning [72].
		Embodied Navigation	Visual navigation [73]; vision-and-language navigation [74].
		Embodied Question Answering (EQA)	Multimodal-based EQA [75]; large language models-based EQA [76].
Embodied Action	Executes decisions through physical interaction.	Imitation Learning-Based Action	Behavior cloning [77, 78]; Inverse Reinforcement Learning [79, 80].
		Reinforcement Learning-Based Action	Value-based RL [81, 82]; policy-based RL [83, 84]; hierarchical RL [85, 86].
		Large Model-Driven Action	Large language models-driven action [87]; vision–language models-driven action [88]; robot-oriented multimodal-driven action [89].

enhance robustness and cross-domain generalization [96, 97]. The integration of large-scale models such as SAM [98, 99] and DINO [100–102] has further advanced the adaptability and zero-shot capability of object perception systems. These advances directly benefit downstream embodied AI tasks, including grasping and placement, path and motion planning, obstacle avoidance, and language-conditioned control.

2.1.2 Scene Perception

Scene perception is a critical capability in embodied AI for understanding the overall environmental structure and semantic layout. By comprehensively analyzing multimodal perceptual inputs such as images, point clouds, and depth maps, the agent can identify the current scene type, infer navigable areas, recognize functional zones, and perform spatial semantic modeling, thereby enhancing its environmental adaptability and reliability in task execution [103, 104].

Scene perception encompasses a set of fundamental environmental understanding tasks, including scene classification, which identifies the type of environment [56]; spatial semantic segmentation, which annotates functional regions at the pixel or voxel level [57]; topological mapping, which constructs semantic graphs capturing object–space relationships [58, 59]; and panoramic perception, which integrates observations from multiple viewpoints into a unified global representation [60, 61]. It further extends to linguistically grounded scene understanding, which incorporates language or behavioral cues to infer spatial relations and functional semantics, enabling agents to interpret instructions such as "find the tray to the right of the operating table" [105, 106]. State-of-the-art approaches typically rely on CNN- or Transformer-based architectures, often combined with visual SLAM, 3D reconstruction, and graph neural networks to achieve multi-scale and multi-view structural reasoning [107, 108]. These advances substantially strengthen the environmental awareness required for downstream embodied AI tasks, including task planning, navigation, human–machine interaction, and embodied question answering.

2.1.3 Behavior Perception

Behavior perception is essential for understanding dynamic interactions in embodied AI. It focuses on recognizing and analyzing actions and behavior sequences performed by humans or other agents in the environment. By modeling visual streams, skeletal poses, and motion trajectories over time, the agent can identify ongoing behaviors, infer intentions, and predict future actions, thereby enabling more effective collaboration, proactive responses, and contextual reasoning.

Behavior perception includes several key tasks, such as action recognition, which identifies single or composite actions [109–112], and action prediction, which anticipates future behaviors based on partial observations [62, 63]. Current approaches mainly employ spatiotemporal convolutional networks, recurrent or Transformer-based sequential models, and graph neural networks to capture motion patterns and temporal dependencies [113]. Recent studies introduce multimodal fusion by combining inputs such as RGB images, depth data, skeletal information, and audio signals to improve robustness under occlusion, viewpoint changes, and environmental noise [114]. To address the variability and open-ended nature of embodied scenarios, some methods incorporate self-supervised representation learning and online adaptation strategies that enhance generalization to unfamiliar behavior categories and facilitate rapid adjustment to new interaction contexts [115]. These advances provide essential support for downstream embodied AI tasks, including human–robot collaboration, imitation learning, interactive decision-making, and anomaly detection.

2.1.4 Expression Perception

Expression perception is a key competence in embodied AI for interpreting human social signals and affective states. By analyzing nonverbal cues such as vocal intonation, facial expressions, gestures, and gaze, the agent can recognize emotional tendencies, infer underlying intentions, and assess social context. These capabilities complement object- and behavior-level perception and support more fluid, human-centered interaction, ultimately enhancing the agent's situational awareness and interaction reliability [116] [117].

Expression perception involves understanding human affective and communicative cues and includes several core tasks. These tasks encompass speech perception, which identifies spoken content and emotional tone [64], facial expression recognition, which detects basic emotions or subtle variations [65], gesture recognition, which interprets body movements as commands or communicative signals [66, 67], and multimodal intention recognition, which infers higher-level semantics by integrating speech, facial cues, and gestures [68]. Technically, mainstream approaches rely on convolutional neural networks, recurrent neural networks, Transformer-based models, and their multimodal fusion extensions, combined with keypoint detection, facial representation learning, and speech feature encoding to obtain a unified understanding of expressive signals [118]. Recent studies further explore large vision-language models such as CLIP and Flamingo to enhance the generalizability of expression interpretation in unconstrained interactions [119]. These capabilities support natural and context-aware human–machine communication and are increasingly applied to assisted diagnosis, emotion recognition, and proactive service in embodied environments.

2.2 Embodied Decision-Making

Embodied decision-making enables an agent to transform perceptual insights into actionable strategies, forming a critical component of the embodied AI [120]. By reasoning over environmental observations and contextual information, agents can generate adaptive plans that account for dynamic changes, uncertainty, and diverse task objectives. Embodied decision-making generally encompasses three key tasks: task planning, embodied navigation, and embodied question answering.

2.2.1 Task Planning

Task planning is a central capability in embodied AI for generating executable sequences of high-level actions based on environmental perception and task objectives. By organizing actions into coherent plans, the agent can accomplish complex, multi-step tasks while dynamically adapting to environmental changes, resource limitations, and unexpected disturbances. This process requires temporal reasoning and policy adjustment to map abstract goals into structured atomic operations (e.g., "grasp–move–place"), ensuring their logical consistency and feasibility across both spatial and temporal dimensions [121, 122].

Task planning involves generating sequences of actions to achieve specific objectives within embodied environments and can be divided into three main categories. Symbolic planning relies on predefined rules and task models to perform logical

reasoning, offering interpretability and generality but often exhibiting limited robustness in the presence of perceptual noise or dynamic environmental changes [69, 70]. Learning-based planning has gained increasing attention and employs techniques such as reinforcement learning, behavior cloning, and large language models to derive task policies directly from data, thereby improving adaptability to complex scenarios and ambiguous goals [71]. Hybrid approaches integrate the structured expressiveness of symbolic methods with the flexibility of data-driven models [72]. Recent trends further explore task decomposition using natural language instruction parsing and hierarchical policy modeling, which combine high-level planning with low-level control [123]. These advances are fundamental for autonomous operation, collaborative task execution, and semantic instruction interpretation, providing a critical link between goal understanding and action generation.

2.2.2 Embodied Navigation

Embodied navigation is a key capability in embodied AI for enabling autonomous movement and goal-directed path planning based on environmental perception. By integrating spatial understanding, target recognition, and action generation, the agent can navigate dynamic, complex, and semantically rich environments. Unlike traditional navigation methods that focus mainly on localization and obstacle avoidance, embodied navigation emphasizes continuous decision-making and adaptive responses, allowing agents to traverse unstructured spaces reliably from their current position to a desired goal location [124, 125].

Embodied navigation encompasses spatial understanding and goal-directed movement in complex environments and can be structured into three interrelated capabilities. First, mapping and localization enable agents to construct accurate environmental maps from sensor data and estimate their real-time positions [126]. Second, path planning generates and adjusts trajectories, either locally or globally, according to the map and the target destination [127]. Third, goal-directed navigation allows agents to execute actions based on language- or image-specified objectives, such as "go to the blue chair" [128]. To improve adaptability in unfamiliar or dynamic environments, reinforcement learning and imitation learning have been widely applied to learn navigation policies [129, 130]. Recent advances in visual navigation, vision-and-language navigation, and language-instructed navigation have emphasized semantics-driven reasoning over purely geometry-driven approaches [73, 74]. To address challenges such as environmental variability, occlusion, and ambiguous instructions, contemporary methods integrate graph neural networks, Transformer-based models, and large-scale pre-trained models to enhance robustness and generalization in real-world scenarios [131–133]. These navigation capabilities support practical applications including object retrieval, goal-directed exploration, and human–robot collaboration, forming a vital bridge between spatial perception and downstream task execution.

2.2.3 Embodied Question Answering (EQA)

Embodied Question Answering (EQA) is a core capability in embodied AI that combines perception, language understanding, and reasoning. By actively exploring the physical environment and interacting with objects in response to natural language

questions, the agent can retrieve accurate answers while grounding its reasoning in real-world contexts. Unlike passive perception tasks, EQA emphasizes active exploration and decision-making, highlighting the agent's ability to integrate multimodal understanding with goal-directed interaction [134].

Embodied question answering (EQA) involves interpreting natural language queries and executing goal-directed actions to obtain the requested information within interactive environments. A typical EQA pipeline includes question parsing, goal inference, perception-guided exploration, path planning, and answer generation. Agents must comprehend queries such as "What is next to the sink?", identify relevant objects, and interact with the environment to acquire necessary information. This process engages multiple submodules, including natural language understanding, object detection, scene mapping, action selection, and language generation. Current approaches frequently employ multimodal encoding frameworks to jointly represent visual, linguistic, and spatial information, while reinforcement learning and large models are used to learn effective strategies and support semantic reasoning [75]. Recent work has further incorporated large language models to enhance reasoning and language understanding capabilities within EQA systems [76]. These capabilities are essential for navigation, semantic interaction, and goal-directed manipulation, providing a crucial pathway toward high-level interactive intelligence guided by natural language.

2.3 Embodied Action

Embodied action enables an agent to execute decisions and interact effectively with its environment, serving as the final component in the perception–decision–action loop in embodied AI [75]. By integrating perceptual information with decision outputs, agents can perform context-aware, goal-directed behaviors while adapting to dynamic and uncertain conditions. Embodied action is generally categorized into three primary approaches: imitation learning-based, reinforcement learning-based, and large model-driven action.

2.3.1 Imitation Learning-Based Action

Imitation learning is a key paradigm in embodied AI for training agents to perform tasks by learning from expert demonstrations. By observing expert trajectories and mapping states to actions, the agent can replicate skilled behaviors and accomplish goal-directed tasks in similar contexts. This approach is particularly valuable in scenarios requiring high sample efficiency or involving task rules that are difficult to model explicitly, making it widely applicable in embodied action domains [135].

Imitation learning focuses on acquiring policies by observing expert behavior and primarily involves two approaches: Behavior Cloning (BC) and Inverse Reinforcement Learning (IRL) [136]. Behavior Cloning treats imitation as a supervised learning problem, directly mapping observations to actions. While it enables efficient training, it is often sensitive to covariate shift due to reliance on limited demonstration distributions [77, 78]. In contrast, IRL aims to infer the underlying reward function from expert demonstrations and then optimize policies through reinforcement learning, providing stronger generalization to novel scenarios [79, 80]. Recent research has expanded

imitation learning with multimodal demonstrations, cross-task and few-shot adaptation, as well as video- or third-person-based learning, all of which improve robustness and applicability in real-world environments [137, 138]. These capabilities have been successfully applied to robotic manipulation, assistive device control, and surgical assistance systems, establishing imitation learning as a fundamental strategy for transferring human expertise and acquiring effective policies with minimal trial-and-error cost.

2.3.2 Reinforcement Learning-Based Action

Reinforcement Learning (RL) is a core paradigm in embodied AI for optimizing policies through trial-and-error interactions with the environment. By continuously exploring and aiming to maximize long-term cumulative rewards, the agent learns optimal action sequences across diverse states. Unlike imitation learning, RL does not rely on expert demonstrations, providing greater autonomy and exploratory capability, which is particularly advantageous for tasks with unknown or dynamically changing rules [139–141].

In embodied AI, reinforcement learning (RL) is widely applied to manipulation policy learning, dynamic control, and policy transfer across tasks. Common RL algorithms include value-based methods [81, 82], policy-based methods [83, 84], and hybrid approaches [85, 86]. To tackle challenges in high-dimensional state and action spaces, researchers have developed extensions such as hierarchical reinforcement learning, meta-reinforcement learning, and model-based reinforcement learning [142–144]. Due to the high cost and safety concerns of real-world training, simulation-to-real transfer has become an important strategy, where agents are pretrained in simulated environments before deployment. These RL-based approaches have achieved significant success in robotic grasping, assembly, locomotion control, and other embodied action tasks, serving as a critical enabler for transitioning agents from passive execution to adaptive autonomy.

2.3.3 Large Model-Driven Action

With the rapid development of large language models (LLMs) and multimodal foundation models, embodied action is increasingly shifting toward a cognitively driven paradigm. These models provide strong capabilities in knowledge retention, language understanding, and cross-modal reasoning, enabling the translation of high-level task intentions into executable action sequences. By supporting flexible responses to complex instructions, the agent can better comprehend task semantics and move toward embodied systems with greater generality and improved generalization capabilities.

Large model-driven action in embodied AI primarily leverages large language models (LLMs) such as GPT-4 [87] and PaLM-E [145], vision–language models including Flamingo [88] and BLIP-2 [146], or robot-oriented multimodal frameworks such as RT-2 [89] and Code-as-Policies [147]. These models enable end-to-end translation of natural language instructions into low-level action policies or executable control code, allowing seamless mapping from language to action. For instance, RT-2 integrates visual and language inputs using a Transformer architecture to generate control commands for multitask execution, while Code-as-Policies employs language models to

produce robot-executable programs, demonstrating high flexibility and generality. Further enhancements, including the incorporation of world models, memory mechanisms, and tool-use APIs, improve reasoning capabilities and task scalability [148]. Despite challenges such as high deployment costs and potential alignment instability, these large model-driven approaches exhibit strong potential for improving generalization and enabling rapid adaptation to novel tasks and environments.

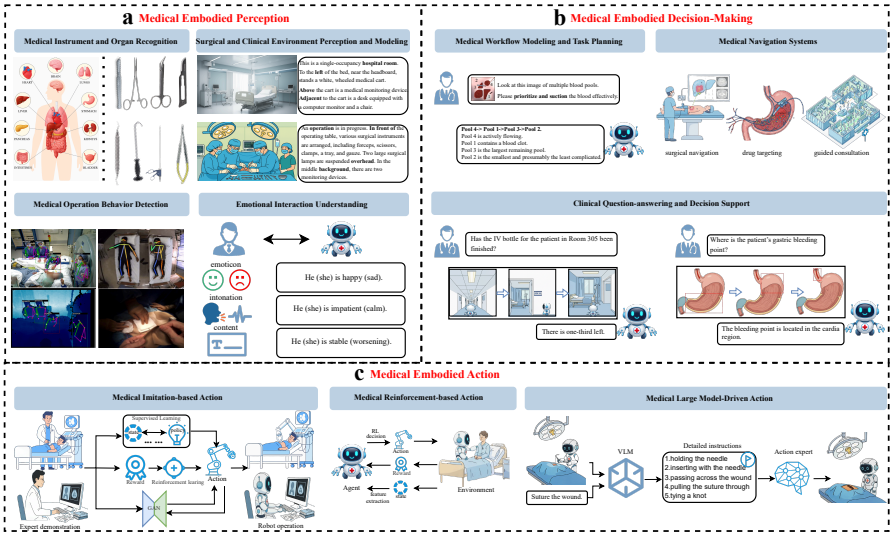

Fig. 2 Embodied AI in medicine. Corresponding to the core components of embodied AI, medical embodied AI encompasses medical embodied perception, medical embodied decision-making, and medical embodied action. a, Medical embodied perception includes medical instrument and organ recognition, perception and modeling of surgical and clinical environments, detection of medical operational behaviors, and understanding of affective and interactive cues. b, Medical embodied decision-making encompasses medical workflow modeling and task planning, medical navigation systems, and clinical question-answering and decision-support mechanisms. c, Medical embodied action consists of imitation-based medical actions, reinforcement-based medical actions, and large-model-driven medical actions.

3 Embodied AI in Medicine

Embodied AI is progressively expanding into the medical domain, driving the evolution of intelligent healthcare systems from static information processing and decision support toward collaborative systems that enable real-time perception, proactive reasoning, and dynamic interaction. To provide a comprehensive overview of the development of embodied AI in medicine, this section focuses on its key applications in medical scenarios: medical embodied perception, medical embodied decision-making, and medical embodied action (Fig.2).

3.1 Medical Embodied Perception

Medical embodied perception is designed to equip agents with the capability to perceive, recognize, and semantically interpret critical elements within medical environments. Compared with open-world environments, medical scenarios are characterized by high object complexity, stringent operational constraints, and densely structured semantic information. As shown in Fig.2a, corresponding to object perception, scene perception, behavior perception, and expression perception, this section focuses on four key aspects: medical instrument and organ recognition, surgical and clinical environment perception and modeling, medical operation behavior detection, and emotional interaction understanding.

3.1.1 Medical Instrument and Organ Recognition

In medical environments, agents primarily perceive surgical instruments and human organs, which are critical for operational safety and diagnostic accuracy [149, 150]. These include diverse tools such as scissors, needle holders, and electrocautery hooks, as well as complex anatomical structures like the liver, lungs, brain, and blood vessels. Recognition is challenging due to dynamic, cluttered scenes, occlusion, blood contamination, unstable lighting, and significant variability in organ morphology. Instruments also undergo rotation, deformation, and partial occlusion, imposing high demands on model robustness and real-time performance.

Current research can be grouped into three main categories: convolution-based image recognition methods, spatio-temporal video-based methods, and multimodal fusion methods. Convolution-based methods are widely used for organ and instrument segmentation in two- and three-dimensional medical images. For example, U-Net [51] and its variants have achieved strong performance in instrument segmentation from laparoscopic images. SwinPA-Net [151] introduces a hierarchical Transformer structure, which enables effective multi-scale feature modeling of complex tissues and lesions. Although these methods are mature and efficient for training, they still have limited ability to handle intraoperative occlusion, dynamic lighting, and rapid tissue deformation. Spatio-temporal or video-based methods capture temporal patterns contained in image sequences, allowing the modeling of instrument trajectories, tissue deformation, and surgical actions. For instance, ST-MTL [152] employs a shared encoder and a spatio-temporal decoder to support joint learning of instrument segmentation and saliency detection. Another study introduces spatio-temporal convolutional layers, which support the simultaneous modeling of tool detection and joint motion estimation [153]. These methods benefit from the continuity of surgical videos and improve recognition robustness, although they remain constrained by the need for large-scale annotated video datasets and high computational demands during real-time inference. Multimodal fusion methods integrate visual, force, depth, and language information to achieve higher-level semantic understanding and stronger generalization. For example, SurgVLM [154] jointly models visual and linguistic features, enabling the system to recognize surgical instruments and anatomical structures

based on textual prompts. Although such methods show improved robustness in complex and dynamic clinical environments, they still face challenges related to sensor synchronization, cross-modal alignment, and computational costs.

Overall, the recognition of surgical instruments and human organs serves as the foundation of embodied perception in medical AI, providing agents with the basic ability to see, distinguish, and interpret clinical scenes. However, the field still faces several challenges, including the difficulty of data acquisition, limited handling of occlusion and deformation, and constraints in model generalization and real-time inference. Future research may focus on multimodal few-shot learning, cross-domain adaptation, lightweight model design, and semantic-aware fusion, which can support the development of agents capable of progressing from perception to understanding and ultimately to action.

3.1.2 Surgical and Clinical Environment Perception and Modeling

In medical embodied perception, the surgical and clinical environment perception and modeling constitute a key component that enables an agent to achieve a comprehensive understanding of its operating space [155] [156]. This task aims to construct structured representations of operating room layouts, device states, clinician–patient interactions, and intraoperative dynamics. Such representations allow embodied AI to perceive and anticipate environmental changes, providing global semantic support for navigation, collaboration, and task planning.

Current research can be broadly divided into three categories: spatial modeling based on three-dimensional reconstruction, relational modeling using graphs and topological structures, and semantic scene modeling driven by large models. Spatial modeling based on three-dimensional reconstruction typically relies on multi-view images, depth information, or point cloud data to recover the geometric structure of surgical scenes and to support continuous updates. For example, NeRF-OR [157] reconstructs high-precision radiance fields from sparse-view RGB-D videos. Deform3DGS [158] builds on this idea by introducing Gaussian splatting and point-cloud initialization, which improve modeling flexibility and accelerate updates. Although these methods provide high geometric accuracy, they remain limited by incomplete viewpoints, occlusions, and substantial computational demands, which restrict their ability to support higher-level semantic reasoning. In contrast, relational modeling based on graphs and topological structures abstracts entities and their interactions in the operating room into semantic scene graphs, enabling a more structured understanding of the environment. 4D-OR [159] uses scene graphs to represent surgical participants, equipment, and their spatial relations, offering a clear framework for downstream reasoning tasks. LABRAD-OR [160] further incorporates temporal information and memory mechanisms, allowing the model to capture the evolving semantics of dynamic operating rooms. These methods are advantageous due to their clear structural representation and interpretability. However, their performance depends heavily on accurate entity detection and relation extraction, and they often struggle to represent implicit semantics in complex environments. With recent advances in vision-language models and embodied foundation models, semantic scene modeling driven by large models

has become an emerging trend. These methods shift the focus from geometric reconstruction to semantic understanding and interaction prediction, leveraging cross-modal reasoning to enhance the depth and generalization of scene analysis. For instance, Spatial-ORMLLM [161] can infer spatial layouts and semantic relations in the operating room using only RGB inputs, providing rich contextual information for task planning and action prediction. Although this category demonstrates strong semantic reasoning and generalization capabilities, it still faces challenges related to high training costs, difficulties in cross-modal alignment, and limited controllability.

Overall, surgical and clinical environment perception and modeling have enabled a shift from purely geometric reconstruction to deeper semantic understanding. However, the field still faces several challenges, including the difficulty of fusing heterogeneous sensory data, conflicts between data annotation and privacy protection, and limitations in both real-time performance and generalization. Future research may focus on unified representation learning that combines graph neural networks with large models, privacy-preserving cross-domain modeling, and lightweight embodied environment models that support real-time semantic updates.

3.1.3 Medical Operation Behavior Detection

In medical embodied perception, detecting surgical and clinical actions is a key component that allows an agent to understand behavior semantics and monitor operations [162]. Medical procedures often involve complex, multi-step action sequences requiring high precision. These actions convey not only task-related information but also the operator's skill, intent, and physical state. Accurate recognition and interpretation of such behaviors enable embodied AI systems to provide real-time feedback and decision support for surgical assistance, skill assessment, remote training, and safety supervision.

Current research on medical operation behavior detection can be grouped into three main categories: vision-based action recognition methods, spatiotemporal modeling methods for surgical phase inference, and multimodal fusion methods for semantic behavior understanding. Vision-based action recognition methods extract spatial and temporal features from surgical videos to identify and classify atomic actions. For example, MGRFormer [163] improves gesture recognition accuracy by modeling interactions between visual cues and kinematic data. Models such as 3D CNNs and SlowFast networks have also been applied to gesture recognition in open surgery, where dynamic features at different temporal scales allow more robust detection of complex actions [164]. These methods do not require additional sensors and are well-suited for vision-driven analysis, although their performance may degrade under occlusions, lighting variations, or heterogeneous operator behaviors. Spatiotemporal modeling methods for surgical phase inference focus on segmenting full procedural workflows and understanding task-level semantics. TransSG [165] models spatial and temporal dependencies through a spatiotemporal Transformer, which supports efficient recognition of surgical gesture sequences. STANet [166] further incorporates multi-scale spatiotemporal features, improving recognition performance across different surgical phases. Although these methods capture temporal patterns and phase-specific characteristics of medical procedures, they still face challenges in multi-task joint modeling

and cross-procedure generalization. Multimodal fusion methods aim to achieve deeper semantic understanding by jointly analyzing visual, haptic, auditory, and physiological signals [167] [168]. Compared with single-modality approaches, these methods capture a broader range of cues related to the operator's perceptual state and surgical intent. They represent a shift from perceptual-level detection toward semantic-level interpretation and enable embodied intelligent agents to perform structured reasoning and contextual understanding of complex medical actions.

Overall, medical operation behavior detection provides a key capability for medical embodied AI, enabling an agent to achieve semantic recognition of procedural actions and to deliver intelligent feedback in dynamic and complex clinical environments. Despite this progress, current research still faces several challenges, including the lack of a unified definition of operative behaviors, the high cost of data annotation, and limited generalization across surgeons and clinical settings. Future work may focus on self-supervised and weakly supervised behavior modeling, operation-level semantic reasoning supported by knowledge graphs, and unified behavior understanding frameworks driven by multimodal embodied foundation models.

3.1.4 Emotional Interaction Understanding

In medical embodied perception, understanding emotional interactions is central to enabling natural communication and contextual awareness among intelligent agents, healthcare staff, and patients [169, 170]. Communication in clinical environments goes beyond semantic content, often involving emotional cues conveyed through vocal tone, facial expressions, body posture, and physiological signals [171, 172]. By accurately recognizing and interpreting these affective and intention-related cues, intelligent systems can support more human-centered, context-adaptive interactions in tasks such as clinical dialogue, gesture-based surgical assistance, and rehabilitation companionship.

Current research on emotional interaction understanding can be grouped into three main categories: audio-visual emotion recognition methods, physiological-behavioral state estimation methods, and language-semantic fusion methods for cognitive understanding. Audio-visual emotion recognition methods identify emotional states by jointly analyzing vocal tone, speech rate, semantic content, and facial expression changes. For example, DEP-former [173] combines audio and facial features and captures dynamic emotional variations to support depression recognition. MSER [174] uses cues from both audio and text and predicts emotion labels with a cross-modal attention mechanism. These methods perform well in scenarios such as monitoring doctor–patient communication and detecting intraoperative emotions, although their performance remains sensitive to noise or occlusion caused by protective equipment. Physiological-behavioral state estimation methods assess the emotional and stress states of healthcare workers or patients by integrating heart rate variability, electrodermal activity, electromyography signals, and body movement trajectories. One dual-stream representation learning framework [175] achieves deep fusion of behavioral and physiological features through feature disentanglement and knowledge transfer, which improves the accuracy of emotion recognition. Another study [176] uses multimodal physiological signals, including EDA and ECG, together with deep learning models to classify multiple emotional states automatically. These methods can reflect

underlying affective responses, yet they remain influenced by sensor noise and individual differences. Language-semantic fusion methods focus on interpreting emotional intent and contextual sentiment in natural language interactions. MedVLM-R1 [177] jointly models textual emotional cues and visual observations, such as patient expressions or clinical scene images, to support emotion-aware clinical question answering and personalized responses. DialogueLLM [178] incorporates contextual and affective knowledge, allowing large language models to achieve dialogue-level emotion recognition and reasoning. These methods provide strong interpretability and adapt well to conversational scenarios, although their performance depends heavily on high-quality language data and may weaken in cases where emotional cues are subtle or linguistically ambiguous.

Overall, emotional interaction understanding provides medical embodied AI with the capacity for empathetic perception and psychological insight. It enables an intelligent agent to recognize human emotional states within complex multimodal environments and respond in a context-appropriate manner. However, current research still faces several challenges, including difficulties in synchronizing heterogeneous signals, the subjectivity of emotion labels, and limited generalization across populations and clinical scenarios. Future work may focus on self-supervised multimodal representation learning, cross-modal emotion transfer, personalized emotion modeling, and unified integration with large embodied models.

3.2 Medical Embodied Decision-Making

Medical embodied decision-making builds on information obtained at the perception level and provides efficient and well-grounded reasoning for clinical tasks. It serves as the core component that enables autonomous behavior and task execution in medical embodied agents [179] [180]. Compared with general-purpose applications, medical tasks often involve complex temporal dependencies, extended procedural chains, and strong reliance on domain knowledge. As shown in Fig.2b, this section discusses three key areas: medical workflow modeling and task planning, medical navigation systems, and clinical question-answering and decision support.

3.2.1 Medical Workflow Modeling and Task Planning

In medical embodied decision-making, medical workflow modeling and task planning are the key intermediate step that connects "observation" with "understanding" and ultimately with "action" [181] [182]. Its goal is to enable an intelligent system to understand complex surgical or diagnostic workflows like a clinician, infer the dependencies between tasks, and generate high-level operational plans. This process requires not only state recognition at the perception level but also temporal modeling, task decomposition, and policy optimization. It forms the fundamental basis that allows medical embodied agents to achieve autonomous decision-making and collaborative execution.

Current research can be grouped into three major categories: supervised stage-modeling methods, temporal graph–based task planning methods, and high-level semantic planning methods driven by language and multimodal models. Supervised stage-modeling methods represent one of the earliest and most mature directions.

These methods often rely on deep convolutional networks and temporal modeling architectures to learn stage segmentation and workflow recognition from surgical videos or sensor signals. For example, Trans-SVNet [183] introduces Transformer architectures into surgical workflow analysis and achieves accurate stage recognition by combining spatial and temporal features. TeCNO [184] improves temporal consistency and robustness through multi-stage temporal convolution networks and hierarchical prediction strategies. Such methods offer simple implementation, stable training, and controllable performance. Their main limitation lies in the fact that structural dependencies between workflow steps remain embedded in network parameters, which prevents explicit modeling and reduces the ability to handle complex task dependencies or long-term logical constraints. To address these limitations, research has increasingly shifted toward temporal graph–based task planning methods. These approaches explicitly represent temporal and semantic relations among tasks, stages, and surgical tools, which enables more structured reasoning and planning. For instance, PATG [185] maps frame-level features into graph nodes and introduces position-aware temporal edges, which strengthens structural modeling and temporal understanding across stages. Other study [186] map detected instrument bounding boxes into graph nodes and construct cross-frame interaction edges to model trajectories and interactions of surgical instruments over time. These methods excel at structured representation and cross-stage reasoning, and they can capture long-range dependencies and explicit task structures. However, their modeling often depends on manually defined graph structures and task priors, which restricts scalability and adaptability. With the rise of large models and multimodal semantic understanding, research has further evolved toward high-level semantic planning methods based on language and multimodal models. These approaches aim to build a unified task-semantic space across visual, linguistic, and knowledge modalities, allowing the model to understand both "what to do" and "why it should be done". For example, SurgVLM [154] learns task decomposition and step-generation capabilities through vision–language pretraining, enabling automatic generation of task sequences and operational goals from textual instructions. LLaVA-Med [187] integrates joint visual–language tuning to support image interpretation and task planning driven by natural-language commands. Compared with traditional structured approaches, this paradigm offers stronger semantic generalization and better cross-task transfer ability. Its challenges include limited interpretability and insufficient integration of medical knowledge constraints.

Overall, medical workflow modeling and task planning have evolved from data-driven learning to structured reasoning and further to semantic-level planning, which enables a shift from simple stage recognition to high-level task understanding. Despite this progress, existing research continues to face challenges such as expensive and privacy-constrained data annotation, limited generalization across clinical settings, and insufficient interpretability and clinical validation of planning results. Future research could focus on multimodal self-supervised learning to reduce reliance on labeled data, on integrating knowledge graphs with language models to strengthen semantic constraints, and on building a closed-loop system that links perception, decision-making, and execution.

3.2.2 Medical Navigation Systems

In medical embodied decision-making, medical navigation systems serve as a key intermediate layer that links spatial perception with action execution. These systems are responsible for core tasks such as localization, registration, and path planning. They support a wide range of applications, including surgical robotics, interventional navigation [188], targeted drug delivery [189], and in-hospital guidance [190] [191]. Although the tasks vary, their shared goal is to enable embodied intelligent agents to achieve precise spatial positioning and safe path decisions.

Current research on medical navigation technologies can be grouped into three major categories: geometric and image registration methods, learning- and optimization-based path planning methods, and multimodal semantic navigation methods. Geometric and image registration methods represent the earliest and most fundamental direction. These methods aim to achieve spatial alignment among the patient, surgical instruments, and preoperative images such as CT, MRI, or ultrasound. Classical systems such as the BrainLab VectorVision Neuronavigation System [192] use optical and electromagnetic hybrid tracking, which provides sub-millimeter localization accuracy and has been validated in various neurosurgical procedures. In addition, augmented-reality-based navigation systems project virtual images onto the surgical field in real time, which improves the surgeon's spatial understanding and operational intuition [193]. Although these methods offer high localization accuracy and strong system stability, they remain vulnerable to tissue deformation, visual occlusion, and stringent real-time constraints. Learning- and optimization-based path planning methods introduce machine learning and reinforcement learning into traditional graph-search frameworks, which allows navigation systems to adapt to dynamic environments. RL-USRegi [194] uses reinforcement learning to achieve radiation-free autonomous ultrasound registration, and demonstrates strong robustness and accuracy in spinal surgery. In vascular intervention, inverse reinforcement learning has been applied to catheter and guidewire navigation, enabling the system to imitate expert behavior and generate optimal control paths [195]. Compared with geometric registration methods, learning-based approaches show better adaptability and generalization in complex and dynamic environments. However, they still face challenges such as high training cost, strong data dependence, and a considerable simulation-to-real transfer gap. Multimodal semantic navigation methods represent a new trend toward more intelligent and semantically grounded navigation. These methods integrate visual, linguistic, knowledge-based, and spatial information, enabling navigation systems to understand task-level semantics and follow high-level instructions. For example, SurgVLM [154] uses vision-language pretraining to perform cross-modal reasoning, which allows the system to convert natural-language instructions from surgeons, such as "advance along the vascular branch to the tumor location," into executable path plans. In broader hospital-level navigation scenarios, NavGPT [196] and NavGPT-2 [197] combine visual scene recognition with semantic map modeling, and can generate indoor guidance paths based on natural-language prompts. Although these methods overcome the modality limitations found in geometric and learning-based approaches and allow navigation systems to act with task-level understanding,

they still require improvements in resolving semantic ambiguity, achieving cross-modal alignment, and ensuring real-time responsiveness.

Overall, medical navigation systems are progressing from a stage that focuses on geometric accuracy and spatial alignment toward a stage that emphasizes semantic understanding and intelligent decision-making. Although these systems still face challenges related to the stability of high-precision registration, the spatiotemporal synchronization of multimodal data, and the interpretability of learning-based models, they have already become an essential component of the medical embodied decision-making framework. Future research could focus on adaptive registration that integrates multimodal information, interpretable path planning that combines reinforcement learning with knowledge-driven constraints, and collaborative navigation systems that incorporate language-based interaction.

3.2.3 Clinical Question-answering and Decision Support

In medical embodied decision-making, clinical question answering and decision support serve as crucial steps that move an intelligent agent from task planning toward decision execution [198] [199] [200]. Their goal is not only to identify key points within surgical or therapeutic workflows, but also to generate interpretable treatment recommendations or action plans based on medical records, imaging data, spoken communication, or other multimodal inputs. This process highlights the agent's ability to collaborate with healthcare professionals in real clinical environments, where knowledge reasoning, language understanding, and decision formulation must be jointly considered.

Current research can be grouped into three major categories: prediction-based decision support methods, language-model-based question answering approaches, and deep decision support systems that rely on multimodal fusion. The first category typically employs machine learning or deep learning to perform risk assessment, treatment response prediction, or complication forecasting using electronic health records, imaging data, and other clinical inputs. For example, a study on colorectal surgery used an AI-based risk prediction model to assist in perioperative pathway planning, which resulted in reductions in both complication rates and overall costs [201]. These methods offer the advantage of providing directly quantifiable decision points, although they often suffer from limited model transparency, weak collaboration with clinical staff, and pronounced "black-box" concerns. The second category focuses on question–answering systems built on language models that support natural language interactions between healthcare professionals and the system. Such approaches can be used for case interpretation, treatment recommendations, or intraoperative decision assistance. For instance, a study in oral and maxillofacial surgery evaluated the role of an AI chatbot in managing complex patients, and the results showed strong performance in both accuracy and completeness [202]. These methods enhance human–machine interaction, yet they remain constrained by the medical specialization, interpretability, and safety of current language models. The final category involves deep decision support systems based on multimodal fusion, which are becoming an important research trend [203] [204]. These systems integrate imaging data, clinical text, behavioral trajectories, and physiological signals to provide comprehensive cognitive support across the "patient–workflow–operation" continuum. Their main

strengths lie in their ability to handle complex scenarios and perform intelligent reasoning. However, challenges remain in data standardization, interpretability, real-time processing, and clinical deployment.

Overall, clinical question answering and decision support provide a crucial bridge that enables medical embodied AI to move from assisting in planning to collaborating in execution. These systems enhance the agent's capacity for language understanding, interactive collaboration, and clinical reasoning. However, significant challenges remain, including the high cost of data acquisition, limited model interpretability, and the difficulty of integrating such systems into clinical workflows while meeting regulatory requirements. Future research may focus on multimodal and few-shot learning, the development of interpretable language–vision decision models, and the construction of a closed loop in which healthcare professionals and intelligent agents jointly contribute to decision execution.

3.3 Medical Embodied Action

Medical embodied action focuses on transforming perceptual and decision outputs into concrete operational actions, serving as a pivotal stage in which an intelligent agent performs medical tasks autonomously [205] [206]. Medical procedures often require high precision, involve substantial risks, and follow strict clinical protocols, which place demanding requirements on the agent's ability to control and execute actions. As shown in Fig.2c, this section discusses three major directions: medical imitation-based action, medical reinforcement-based action, and medical large-model-driven action.

3.3.1 Medical Imitation-based Action

In medical embodied action, such as suturing, needle insertion, soft-tissue dissection, and catheter advancement, the tasks require not only high-precision motor control but also complex physical interactions, nonlinear tissue deformation, and strict safety constraints [207]. Direct reliance on large-scale trial-and-error reinforcement learning is costly and raises ethical and safety concerns. As a result, transferring human surgical skills to robots or embodied agents through expert demonstrations has become a practical and efficient strategy.

Current research on medical imitation-based execution can be grouped into three categories according to the underlying learning paradigm: behavior cloning, inverse reinforcement learning, and hybrid imitation–reinforcement learning. Behavior Cloning (BC) uses supervised learning to directly approximate the state–action mapping demonstrated by experts, allowing the model to imitate the surgeon's operational strategies. It has been widely applied to fundamental surgical skills such as needle grasping, tissue retraction, and knot tying. For example, the Surgical Robot Transformer (SRT) [208] uses a relative action representation that enables systematic verification of BC on the da Vinci research platform, demonstrating strong feasibility and generalization. Intermittent Visual Servoing [209] achieves sample-efficient imitation under visual closed-loop control and significantly improves robustness and success rates in high-precision tasks such as peg transfer. As data complexity and task diversity increase, SuFIA-BC [210] generates synthetic demonstrations within high-fidelity

digital-twin environments, which enhances the generalization quality of vision–action policies. Although BC is efficient and straightforward, it depends heavily on high-quality expert demonstrations with accurate temporal alignment and is prone to distributional shift in complex dynamic environments. In contrast, Inverse Reinforcement Learning (IRL) and Generative Adversarial Imitation Learning (GAIL) aim to recover the implicit reward signals of experts or distinguish expert trajectories through adversarial discriminators, thereby producing more robust policies. For instance, an intra-operative steerable-needle path planner [211] learns an implicit reward function that governs needle motion, enabling optimized needle-path planning during surgery. Model-free adversarial imitation methods [212] have also been applied to automate standard arterial catheter insertion. These methods show promising ability to capture implicit expert intentions in high-contact tasks such as tissue traction and exposure, yet they require complex training procedures, incur high computational costs, and rely on stable trajectory distribution estimation and optimization. The combination of imitation learning and reinforcement learning has emerged as a new trend in embodied action for medical applications. These approaches typically use imitation learning to obtain a safe initial policy that avoids hazardous exploration, while reinforcement learning provides individualized refinement and dynamic adaptation. For example, an ophthalmic robotic assistant learns surgeon-specific preferences through ring-shaped demonstrations, enabling personalized skill adaptation [213]. The Imitation Learning framework for Laparoscope Control (ILLC) with reinforcement learning [214] integrates limited demonstrations with simulation-based fine-tuning and achieves cross-scenario generalization in laparoscopic camera control. This hybrid paradigm offers clear advantages in tasks with limited interaction data and strict safety constraints, and it provides a practical pathway toward safe and adaptive learning for future medical embodied agents.

Overall, imitation learning provides a safe and efficient pathway for transferring human surgical skills to medical embodied agents, yet achieving clinically robust deployment remains challenging. High-quality demonstrations that are multimodal, which may include visual, force, kinematic, and language signals, are difficult to collect and require strict temporal alignment. In addition, the distribution gap between demonstrations and real clinical environments is often substantial. Future research may focus on hierarchical imitation that decomposes complex surgical procedures into learnable subskills, multimodal fusion of expert demonstrations, and integrated frameworks that use imitation for initialization followed by safe reinforcement-based refinement.

3.3.2 Medical Reinforcement-based Action

In embodied action for medical tasks, reinforcement learning (RL) optimizes control policies through continuous interaction with the environment, enabling an agent to achieve adaptive and optimal control in dynamic and complex surgical settings [215]. Unlike imitation learning, which relies on expert demonstrations for policy initialization, RL is driven by reward signals and allows the agent to autonomously explore the mapping between perception, decision-making, and action. This paradigm provides a

crucial pathway for embodied agents to move from imitating experts to learning independently, and it forms an essential foundation for the intelligent execution of complex medical operations.

Common reinforcement learning algorithms can be grouped into three categories: value-based methods, policy-based methods, and hybrid Actor–Critic methods. Value-based approaches estimate state–action value functions to guide policy updates, and they are well-suited for low-dimensional and discrete decision problems. For example, Collaborative Suturing [216] constructs a sparse reward function and applies Q-learning to derive an optimal policy, enabling autonomous learning of handover actions during surgical suturing. These methods offer advantages in convergence and policy interpretability, although they are highly sensitive to reward design and often suffer from low sample efficiency. Policy-based approaches directly optimize a policy function to maximize expected returns, which gives them stronger capabilities for continuous action modeling and more stable learning behavior. For instance, LapGym [217] uses the PPO algorithm as a model-free baseline to analyze task difficulty and policy performance in laparoscopic environments. The A3C [218] has also been employed in virtual vascular intervention settings to enable multi-task navigation and automated instrument control. These methods perform well in continuous control and cross-task transfer, yet they remain dependent on carefully designed reward signals and require substantial computational resources and training samples. Hybrid methods integrate the strengths of value estimation and policy optimization. They employ an Actor–Critic structure that learns both action values and policy distributions, which makes them effective in high-dimensional continuous spaces and partially observable scenarios. For example, AC-SSIL [219] introduces a self-supervised imitation learning technique that retrieves the closest demonstration states and uses them to guide the Actor network, enabling an efficient transition from imitation to reinforcement. CASOG [220] provides another representative approach, as it conservatively estimates the Q-function and smooths gradients, which helps mitigate distribution shift and overfitting while improving stability and sample efficiency in complex surgical skill learning. Despite these advantages, hybrid methods often require careful tuning of dual learning signals and may become unstable when value estimation and policy updates are not well aligned, which limits their robustness in real surgical environments.

Overall, reinforcement learning provides a critical pathway for medical embodied action, as it enables agents to move from passive imitation to active optimization and to refine their policies through exploration and feedback in complex surgical environments. However, its clinical deployment still faces several challenges, which include high sample costs, strict safety constraints, difficulties in defining effective reward functions, and substantial gaps between simulation and real procedures. Future research may focus on multimodal perception fusion, hierarchical and safety-oriented reinforcement learning, and knowledge-guided policy optimization that integrates large medical models, so that autonomous surgical execution systems can become more robust and more interpretable.

3.3.3 Medical Large Model-Driven Action

In medical embodied action, large models, particularly vision–language–action models, are reshaping how agents learn and make decisions [221] [222] [223]. Unlike traditional imitation learning and reinforcement learning, which rely on limited demonstrations or task-specific reward signals, large models are pretrained on extensive multimodal data that include visual inputs, language instructions, and action trajectories. This pretraining grants them strong representational, reasoning, and generalization capabilities, which allow medical embodied agents to perform complex operational tasks even when explicit supervision is limited.

Currently, large model–driven medical embodied action follows three main development directions: multimodal representation–driven perception–action alignment, language-conditioned task planning, and cross-modal transfer with few-shot execution. In the first approach, large models jointly model visual and action modalities to achieve end-to-end mapping from video observation to action generation. For example, SurgicalGPT [224] constructs a multimodal Transformer that integrates explicit surgical video representations with implicit language instructions, enabling automatic segmentation and prediction of surgical phases and action sequences in unannotated scenarios. SurgVLM [154] is pretrained on large-scale paired surgery video–text data, allowing the model to generate executable action sequences from natural language descriptions and thus achieve cross-modal translation from understanding to execution. The second approach emphasizes the use of language instructions or knowledge graphs to guide embodied policy learning, enabling agents to generate interpretable action plans under high-level semantic guidance. For instance, LLaVA-Med [187] and Med-Flamingo [225] employ joint visual–language modeling to allow surgical robots to perform tasks such as instrument alignment and incision path planning based on natural language prompts, demonstrating semantically guided dynamic execution. The third approach focuses on cross-task transfer and few-shot generalization by using parameter-efficient fine-tuning methods, such as LoRA or Adapter, or sim-to-real alignment mechanisms to enable rapid adaptation to new tasks. For example, RoboNurse-VLA [226] can grasp and deliver surgical instruments in real time according to surgeons' verbal instructions, maintaining high operational success rates even when faced with unknown tools or complex scenarios. This illustrates the scalability and generalization potential of embodied agents in open medical environments.

Overall, the introduction of large models has shifted medical embodied action from experience-based learning to knowledge-driven approaches, enabling unified perception, reasoning, and action. However, this field still faces challenges such as the scarcity and privacy constraints of multimodal medical data, limited stability and safety of model reasoning, and restricted cross-modal alignment and few-shot generalization. Future research could focus on constructing medical-specific multimodal corpora, developing interpretable and safety-constrained instruction-following mechanisms, and optimizing large models for human–machine collaboration.

Table 2 Integrated application scenarios of medical embodied AI.

Scenario	Core Functions	Key Technologies	Representative Applications
Surgical Robot	Perception–decision–action integration; precise manipulation; enhanced safety and consistency	Multimodal sensing; real-time imaging feedback; motion scaling; autonomous assistance	da Vinci [227–229]; EMARO [230]; ROSA [231]; PRECEYES [232]; Robodoc [233]; Symani [234]; CyberKnife [235]; Monarch [236]; Flex [237]; CorPath GRX [238]
Intelligent Caregiving and Companion Robot	Daily care support; emotional interaction; contextual understanding; environmental adaptability	Multimodal perception; affective computing; compliant control; telepresence	PARO [239]; AIBO [240]; Pepper [241]; ElliQ [242]; Arash [243]; Robear [244]; Giraff [245]; Telenoid [246]
Immersive Medical Education Platform	Cognitive training; operative skill learning; system-level simulation	VR/AR/MR simulation; force feedback; anatomical modeling; data-driven assessment	Touch Surgery [247]; Body Interact [248]; VirtaMed ArthroS [249]; Vimedix 3.2 [250]; OSSO [251]; 3D Organon VR [252]
Telecollaborative Diagnostic and Treatment System	Remote consultation; distributed collaboration; resource sharing	Multimodal communication; real-time monitoring; cloud-based analytics; telepresence	Teladoc Mini Cart [253]; Mercy Telehealth [254]; VSee [255]; Creyos [256]

3.4 Integrated Application Scenarios in Healthcare

The comprehensive application scenarios of embodied AI in medicine highlight the synergistic integration of perception, decision-making, and action capabilities within real clinical environments. These scenarios represent a crucial direction for advancing intelligent agent technologies toward practical clinical applications. This section provides an in-depth discussion of four representative domains: surgical robot, intelligent caregiving and companion robot, immersive medical education platform, and telecollaborative diagnostic and treatment system (Table 2).

3.4.1 Surgical Robot

Surgical robotic systems represent advanced embodiments of medical embodied AI, integrating perception, decision-making, and action capabilities within clinical operations. By leveraging multimodal sensors to perceive surgical environments and combining intelligent control with real-time imaging feedback, these systems enable precise manipulation and path planning in complex anatomical regions, thereby significantly enhancing surgical safety and consistency. At present, various surgical robotic platforms are widely used in urology, neurosurgery, orthopedics, ophthalmology, and cardiovascular interventions.

The da Vinci Surgical System [227] [228] [229] remains the most mature and widely deployed minimally invasive surgical robot. Its multi-degree-of-freedom robotic

arms and high-definition three-dimensional vision enable highly precise laparoscopic procedures, supporting standardized clinical use across abdominal, thoracic, ophthalmic, and urological surgeries. The EMARO system [230], developed at the Tokyo Institute of Technology, adopts pneumatic actuation together with a head–foot control strategy that enables autonomous endoscope positioning and stable operational control, thereby reducing the need for human surgical assistance. The ROSA® platform [231] is widely used in neurosurgery and spinal procedures; its combination of stereotactic localization and intraoperative navigation provides accurate guidance for electrode implantation and spinal screw placement. The PRECEYES Surgical System [232] enables sub-millimeter retinal manipulation through remote operation and motion-scaling techniques, supporting delicate intraocular procedures with enhanced stability and precision. The Robodoc system [233], an early orthopedic surgical robot, automates femoral resection and prosthesis implantation based on preoperative CT-derived plans. The Symani Surgical System [234] targets microsurgical and reconstructive applications and employs motion scaling and tremor-suppression mechanisms to facilitate fine suturing of blood vessels and nerves. The CyberKnife platform [235] exemplifies radiosurgical robotics by integrating real-time image tracking with a six-degree-of-freedom robotic arm capable of delivering frameless stereotactic radiotherapy. The Auris Monarch Platform [236] incorporates a flexible catheter architecture and video-based navigation to achieve precise localization and biopsy of distal pulmonary lesions during bronchoscopic interventions. The Flex Robotic System [237] provides a snake-like flexible arm that supports transoral access within narrow anatomical structures of the head, neck, and pharynx. The CorPath GRX platform [238] enables remote, high-precision manipulation of guidewires and stents in cardiovascular interventions, with ongoing research exploring the integration of AI-driven trajectory prediction to further enhance procedural safety.

Overall, surgical robotic systems achieve precise manipulation in complex anatomical environments and enhance procedural safety and consistency. Future developments will emphasize deeper modality fusion, adaptive intraoperative reasoning, and more autonomous assistance, enabling broader applications in minimally invasive surgery and microsurgical intervention while meeting stringent requirements for robustness and reliability.

3.4.2 Intelligent Caregiving and Companion Robot

Intelligent caregiving and companion robots serve as a vital link between professional medical services and patients' daily care. Deployed in hospital wards, rehabilitation centers, and eldercare facilities, they help alleviate staff shortages and enhance patients' sense of safety and adherence through continuous companionship and personalized interaction. Compared with general service robots, medical companion robots require greater environmental adaptability, human–robot interaction capability, and contextual awareness, while adhering to strict medical safety and ethical standards. Most adopt humanoid or biomimetic designs that foster emotional engagement, encompassing both home-based devices for individual users and institutional platforms for group care.

Intelligent caregiving and companion robots are generally divided into three categories: emotional companion robots, physical assistance robots, and telepresence companion robots. Emotional companion robots provide psychological support and social engagement through multimodal sensing that integrates voice, vision, and touch, relying on affective computing models and interaction mechanisms that generate personalized emotional responses. PARO [239], the biomimetic seal robot widely adopted for individuals with dementia and cognitive impairment, has shown clear effectiveness in reducing anxiety and alleviating loneliness. AIBO [240] has been introduced into pediatric wards for studies on social interaction, where it supports emotional improvement and encourages communication among children. Pepper [241] combines speech recognition, emotion modeling, and natural dialogue, enabling its use in dementia care, chronic disease education, and cognitive stimulation therapy with consistent benefits for the richness and intelligence of human–robot interaction. ElliQ [242], designed for older adults living independently at home, integrates speech, lighting cues, and touchscreen interaction to provide entertainment, video calls, cognitive exercises, and health reminders. Deployments across eldercare facilities in the United States indicate substantial improvements in well-being and reductions in loneliness. Arash [243], developed for pediatric cancer patients, supports educational and emotional engagement that helps reduce treatment-related stress and enhance adherence. Keepon contributes to emotional regulation and social motivation for children with autism and for older adults through rhythmic movements and responsive visual feedback. Physical assistance robots support patient transfer, posture adjustment, and body stabilization. These systems rely on compliant force control and dynamic balance mechanisms that reduce caregiver workload while improving the safety of clinical and nursing operations. Robear [244] demonstrates smooth and stable motion suitable for lifting and transferring bedridden patients and provides essential postural and mobility support for users with limited motor function. Telepresence companion robots focus on sustaining emotional connection and medical support across physical distances. Giraff [245] integrates telecommunication capabilities with autonomous navigation to assist medical staff in remote ward rounds, alert response, and environmental monitoring, thereby increasing the reach and efficiency of caregiving across multiple units. Telenoid [246] employs a minimalist humanoid design capable of conveying basic facial expressions and vocal cues, enabling human-like communication that provides emotional connection and psychological comfort for older adults living alone or with restricted mobility.

Overall, intelligent caregiving and companion robots are transforming single-task execution to context understanding and emotional interaction. Future developments are expected to focus on multimodal perception fusion, cognitive context modeling, and empathetic human–robot interaction. These advances are anticipated to enable long-term care, rehabilitation assistance, and mental health management that meet strict safety and reliability standards.

3.4.3 Immersive Medical Education Platform

Immersive medical education platforms aim to overcome the limitations of traditional teaching that stem from restricted learning environments, limited resources,

and experience-based instruction. By integrating virtual reality (VR), augmented reality (AR), and mixed reality (MR) technologies, these platforms reconstruct complex anatomical structures, pathological processes, and clinical procedures, enabling students to conduct multisensory and iterative training in virtual settings. They enhance the efficiency and safety of clinical skill acquisition while supporting remote learning, resource sharing, and standardized assessment.

Based on functional design and educational objectives, existing immersive medical teaching systems can be classified into cognitive training, operative skill, and comprehensive simulation platforms. Cognitive training platforms emphasize the visualization and interactivity of clinical reasoning and decision-making processes. Touch Surgery [247] was developed based on the Cognitive Task Analysis framework and integrates virtual reality with mobile computing environments to create a cognitive simulation platform for surgical education. The system decomposes surgical procedures into perceptible and practicable decision units, placing greater emphasis on the development of surgical reasoning than on the repetitive acquisition of motor skills. This design supports preoperative planning and enables adaptive evaluation of individual learning trajectories. Body Interact [248] serves general medical education through interactive virtual patients that allow multiple learners to engage in case discussions, collaborative diagnosis, and emergency decision-making. Operative skill platforms concentrate on hands-on training in anatomy, endoscopy, and ultrasound. VirtaMed ArthroS™ [249] provides a high-fidelity arthroscopy simulation environment that reproduces surgical scenarios through force feedback and visual rendering, and its modular architecture supports validation and certification across multiple joint procedures. Vimedix 3.2 [250] offers virtual ultrasound training for abdominal, transthoracic cardiac, and major vascular examinations, enabling risk-free repetitive practice and the automatic generation of standardized reports that enhance diagnostic accuracy. The OSSO system [251] introduces a data-driven framework for learning the mapping between human body surfaces and internal skeletal structures using dual-energy X-ray absorptiometry data from 2,000 subjects. Its design supports orthopedic structure recognition and reconstruction, forming a foundation for future personalized surgical navigation. Comprehensive simulation platforms combine anatomy visualization with multidisciplinary medical education. 3D Organon VR Anatomy [252] includes fifteen major body systems, more than 10,000 anatomical structures, and over 550 motion modules involving muscles and organs. The platform enables virtual dissection, free rotation, and layered exploration of anatomical relationships, allowing learners to intuitively grasp spatial topology and functional connections within an immersive environment.

Overall, these immersive medical training platforms have established a progressive structure within the medical education system, evolving from cognitive decision-making to procedural skill acquisition and ultimately to system-level understanding. By integrating multimodal perception, virtual simulation, and data-driven assessment algorithms, future systems are expected to merge with artificial intelligence to enable personalized learning path recommendations and adaptive competency evaluations.

3.4.4 Telecollaborative Diagnostic and Treatment System

Telecollaborative diagnostic and treatment systems aim to overcome the constraints of physical distance, enabling efficient collaboration and resource sharing across geographically distributed healthcare institutions. By integrating sensing, communication, and interaction technologies, these systems link clinicians, patients, and distributed intelligent agents, forming an interactive network for collaborative care. In scenarios such as public health emergencies, primary care support, and multidisciplinary consultations, telecollaboration enhances medical accessibility and response efficiency while promoting the equitable distribution and downward extension of high-quality healthcare resources.

Currently, several representative telecollaborative diagnostic and treatment systems have emerged internationally. Teladoc Health [253], a global leader in telemedicine, developed the Mini Cart system, which connects terminal devices with cloud applications via the Internet to enable near face-to-face, real-time clinical interactions. The platform has been widely adopted in chronic disease management, psychological counseling, and postoperative follow-up. It features high-fidelity video communication, an AI-based health assistant, and integration with electronic medical records. The Mercy Telehealth system [254] excels in intensive care and remote monitoring. Its virtual care center connects high-definition cameras to ward terminals, allowing remote physicians and nursing teams to monitor and guide patient care in real time. The system strictly adheres to privacy protection protocols, requiring explicit on-site authorization before entering a patient's ward remotely. In the domain of pathology and laboratory collaboration, the VSee platform [255] provides HIPAA-compliant low-bandwidth video conferencing designed for international pathology consultations and teaching. By integrating microscope cameras with imaging software, VSee enables real-time transmission of histopathological slides to remote specialists, supporting multi-user collaboration and remote diagnostics. Additionally, the Creyos platform [256] specializes in remote cognitive assessment, offering web-based autonomous cognitive testing and quantitative analytics that cover reasoning, memory, and attention domains. It has been utilized in clinical studies to detect post-operative cognitive dysfunction in cardiac surgery patients, establishing a standardized technical pathway for remote neuropsychological evaluation.

Overall, telecollaborative diagnostic and treatment systems are evolving from video conferencing–based consultations to intelligent, interactive collaboration networks. Future development trends include diagnosis and treatment assistance driven by multimodal medical data fusion, collaborative decision-making enabled by deep learning, and embodied-intelligence–assisted remote operation systems.

4 Datasets and benchmark

High-quality data resources serve as a fundamental cornerstone for advancing the development of embodied AI in medicine. This section systematically reviews representative publicly available datasets, categorized according to the three technological layers of perception, decision-making, and action, as well as additional simulation and synthetic data (Fig.3).

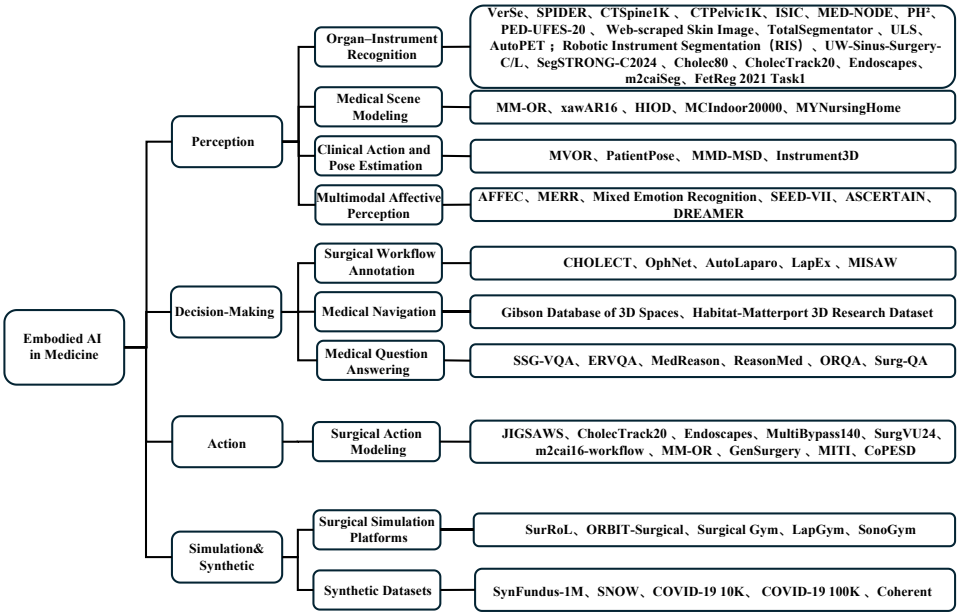

Fig. 3 The datasets of medical embodied AI, which are categorized into perception, decision-making, action, and simulation & synthetic.

4.1 Perception Datasets

4.1.1 Organ–Instrument Recognition Datasets

Organ recognition tasks primarily focus on image segmentation, and numerous publicly available datasets now cover a wide range of organs. For the head and neck region, existing datasets support segmentation of brain tissues, identification of stroke lesions [257], detection of intracranial hemorrhage [258], and fine-grained delineation of high-risk areas such as retinal vessels [259] and fundus lesions [260]. In the thoracic region, datasets enable segmentation of structures including the heart [261], trachea [262], thoracic cavity [263], and breast tumors [264], while also supporting disease classification tasks such as pneumonia and lung cancer. In the abdominal region, common segmentation targets include the liver [265], kidneys, renal tumors, and gastrointestinal organs. For the skeletal system, datasets such as VerSe [266], SPIDER [267], CTSpine1K [268], and CTPelvic1K [269] provide detailed annotations of the spine and spinal canal structures. In the field of dermatology, the ISIC dataset series [270], MED-NODE [271], and PH^2 [272] support tasks related to segmentation and classification of melanoma, whereas PED-UFES-20 [273] and Web-scraped Skin Image datasets [274] are utilized for classification of various skin diseases. Additionally, the TotalSegmentator dataset [275] series offers precise segmentation of multiple organs throughout the body. ULS [276] and AutoPET [277] support automated whole-body tumor segmentation, providing a high-quality data foundation for systemic medical analysis and comprehensive diagnosis.

Medical instrument recognition encompasses tasks such as detection, localization, segmentation, and tracking of surgical tools, serving as a foundational component for surgical automation and intraoperative intelligent assistance. A variety of datasets have been developed and widely adopted for different types of instruments and surgical scenarios. Some datasets are specifically designed for instrument segmentation. For example, the Robotic Instrument Segmentation (RIS) dataset [278] focuses on segmenting typical instrument structures in the da Vinci surgical robot, including key components such as forceps and instrument shafts. The UW-Sinus-Surgery-C/L dataset [279] targets endoscopic sinus surgery and provides detailed instrument segmentation annotations under challenging conditions such as specular reflections and smoke occlusion. The SegSTRONG-C2024 dataset [280] emphasizes robust segmentation of instruments under non-adversarial degradation conditions, including low illumination, bleeding, and smoke. Other datasets include surgical videos or support multiple tasks, treating instrument segmentation as a subtask. The Cholec80 [281] and CholecTrack20 [282] datasets document complete laparoscopic cholecystectomy procedures performed by multiple surgeons and support instrument recognition and multi-class tracking research. The Endoscapes dataset [283] comprises more than 200 laparoscopic surgical videos and supports tasks such as instrument detection, scene segmentation, and intraoperative risk assessment. The m2caiSeg dataset [284] provides semantic segmentation labels for 19 categories of organs and instruments in laparoscopic surgery. The FetReg 2021 Task 1 dataset [285] focuses on segmenting surgical instruments and vascular structures in fetoscopic images, offering crucial data support for instrument recognition in minimally invasive surgical environments.

4.1.2 Medical Scene Modeling Datasets

The detection, segmentation, and modeling of hospital environments aim to enable intelligent understanding of spatial layouts, equipment distribution, and human activities within medical settings such as operating rooms and wards. Several datasets have been developed to support such research tasks across levels ranging from visual perception and semantic understanding to 3D reconstruction. For example, the MM-OR dataset [286] focuses on semantic understanding in high-intensity surgical environments. Collected from real-world operating rooms, it provides multimodal data including RGB-D images, audio, speech transcripts, robot logs, and spatial tracking information. It supports tasks such as panoramic segmentation, semantic scene graph construction, and various downstream applications, and represents the first large-scale multimodal dataset dedicated to operating room scene understanding. The xawAR16 dataset [287], captured in a mixed-reality surgical environment, is designed to evaluate the visual localization capabilities of handheld mobile devices. It offers RGB-D images and precise spatial pose annotations from multiple viewpoints, making it suitable for indoor 3D mapping and augmented reality interaction studies. The HIOD (Hospital Indoor Object Detection) dataset [288] contains 4,417 images with over 50,000 instance annotations across 56 common hospital object categories. It supports multi-class object detection in complex indoor environments. The MCIndoor20000 dataset [289] includes more than 20,000 images of three typical structural elements, including doors, staircases, and hospital signage. It is suitable for research on hospital

environment recognition and navigation. The MYNursingHome dataset [290] focuses on elderly care scenarios and provides 37,500 images covering 25 categories of indoor objects. This dataset can be used to support assistive facility identification and the development of interaction systems for vulnerable populations.

4.1.3 Clinical Action and Pose Estimation Datasets

Clinical action and pose estimation aims to identify and understand the movement patterns and interactive behaviors exhibited by medical personnel during diagnosis, nursing, and surgical procedures. Related datasets primarily focus on tasks such as human pose estimation, action recognition, and multi-agent interaction, thereby advancing the application of embodied AI in modeling medical behaviors. For example, the MVOR (Multi-View Operating Room) dataset [291] collects synchronized RGB-D videos from multiple camera views in real surgical rooms, capturing the activities of surgeons, nurses, and other roles across different surgical phases. It supports 3D pose estimation and multi-target tracking. The PatientPose dataset [292] centers on clinical settings such as epilepsy monitoring units and provides upper-body pose annotations from long-duration patient video recordings. It is integrated with neural recording systems to explore the relationship between neural activity and spontaneous movements, enabling objective assessment of patients' motor abilities. The MMD-MSD dataset [293] combines data from camcorders, video cameras, wrist-worn sensors, and triaxial accelerometers. It is designed for research on the prevention of musculoskeletal disorders by modeling body posture and physiological states during computer use. Additionally, the Instrument3D dataset [294] supports the evaluation of 3D tracking of surgical instruments, offering essential data for understanding fine-grained surgical operations with high precision.

4.1.4 Multimodal Affective Perception Datasets

Multimodal affective perception datasets integrate information from multiple sensory channels such as electroencephalography (EEG), facial expressions, physiological signals, and eye movements, providing a robust data foundation for research in emotion recognition, affective modeling, and human–computer interaction. The AFFEC dataset [295] combines eye-tracking data, facial action units (AUs), galvanic skin response (GSR), and personality trait information, establishing a multimodal emotion classification pipeline designed for affective computing and psychological research. The MERR dataset [296] includes 28,618 coarse-grained and 4,487 fine-grained samples across diverse emotional categories, supporting tasks in multimodal emotion recognition and inference. The Mixed Emotion Recognition dataset [297] records multimodal signals and subjective ratings from 73 participants using carefully curated video stimuli, along with EEG, electrocardiography (ECG), GSR, and facial videos, enabling research in emotion elicitation and modeling of blended emotional states. The SEED-VII dataset [298] focuses on seven emotional categories, including happiness, sadness, and anger, offering EEG and eye-tracking signals with continuous labels to represent emotion intensity, making it suitable for continuous emotion modeling. The ASCERTAIN dataset [299] incorporates EEG, ECG, GSR, and facial expressions to establish the first multimodal link between personality traits and emotional states.

The DREAMER dataset [300] uses low-cost, wireless wearable devices to acquire EEG and ECG signals, along with self-reported labels of valence, arousal, and dominance, aiming to advance the practical application of emotion recognition technologies in real-world settings.

4.2 Decision-Making Datasets

4.2.1 Surgical Workflow Annotation Datasets

Surgical workflow annotation and procedural phase recognition serve as a fundamental basis for enabling automation during surgical procedures. In recent years, several high-quality datasets have been developed to support surgical workflow modeling across a wide range of procedures and task granularities, including laparoscopic, ophthalmic, and minimally invasive surgeries. The CHOLECT dataset series [282] focuses on laparoscopic cholecystectomy, providing triplet annotations of surgical instruments, actions, and anatomical targets in the form of ¡instrument, verb, target¿. It enables fine-grained surgical action recognition and includes spatial instrument trajectories as well as phase annotations. The OphNet dataset [301] comprises 2,278 ophthalmic surgery videos, covering 102 surgical phases and 150 fine-grained actions. With its hierarchical temporal annotations, it supports surgical workflow modeling, temporal localization, and phase prediction tasks. The AutoLaparo dataset [302] centers on hysterectomy and integrates workflow recognition, motion prediction, and image segmentation tasks, offering a comprehensive multimodal resource for image-guided surgical automation. The LapEx dataset [303] targets sleeve gastrectomy procedures, providing activity annotations, scene segmentation, and skill assessment labels, emphasizing its utility in intraoperative context modeling and operator performance evaluation. The MISAW dataset [304] focuses on minimally invasive vascular anastomosis scenarios and integrates synchronized video, kinematic data, and phase-level workflow annotations, highlighting its strengths in multimodal analysis of surgical behavior.

4.2.2 Medical Navigation Datasets

Medical navigation datasets provide essential support for accurate localization, path planning, and environmental understanding in embodied AI applied to healthcare scenarios. In the field of surgical navigation, A Portable 6D Surgical Instrument Magnetic Localization Dataset [305] offers six-degree-of-freedom magnetic localization data in minimally invasive surgery contexts. It includes various motion trajectories such as square, circular, and spiral paths, along with simulated experiments involving knee surgeries and needle biopsies. The data is collected using 16 magnetometers at a sampling rate of 300 Hz, available in both raw and filtered formats. For mixed reality navigation in neurosurgery, the Head Model Collection for Mixed Reality Navigation compiles CT and MRI data from 19 patients with intracranial lesions, along with corresponding three-dimensional anatomical models [306]. This dataset supports high-quality preoperative modeling and intraoperative validation. Beyond surgical applications, broader medical scene navigation and spatial modeling are supported

by the Gibson Database of 3D Spaces [307], which includes three-dimensional models of 572 real-world environments across 1440 floors, covering hospitals, residential areas, and office buildings. The Habitat-Matterport 3D Research Dataset [308] further enriches this domain by providing high-resolution digital twins of 1000 real indoor environments, including 3D meshes, texture maps, and material files, supporting research in medical navigation, spatial understanding, and autonomous planning.

4.2.3 Medical Question Answering Datasets

In advancing the capabilities of large medical language models, question answering and reasoning datasets play an indispensable role, particularly in supporting high-level semantic understanding and decision-making within embodied AI. The SSG-VQA dataset [309], constructed from laparoscopic surgical videos, serves as a visual question answering benchmark. It utilizes segmentation and detection models to extract spatial and action-related information of surgical instruments and anatomical structures, thereby generating structured surgical scene graphs. The ERVQA dataset [310] focuses on emergency room scenarios, offering image-question-answer triplets to evaluate the adaptability of vision-language models in real clinical environments. Through systematic error analysis, it further reveals the inherent complexity of such tasks. The MedReason [311] builds logical inference chains grounded in structured medical knowledge graphs to provide stepwise explanations for over 30,000 question-answer pairs. The ReasonMed [312] integrates multi-agent verification mechanisms to refine 370,000 high-quality reasoning instances from initial outputs generated by large language models. This resource significantly advances medical language modeling and interpretable inference. The ORQA dataset [313] establishes a multimodal question answering benchmark for operating room environments by integrating four public datasets—MVOR, 4D-OR, EgoSurgery, and MM-OR. It incorporates visual, auditory, and structured data to generate 23 types of question-answer pairs that support multitask learning and reasoning in surgical contexts. The Surg-QA [314], the first large-scale instruction-based question answering dataset for surgical videos, includes 44,000 video clips from 2,201 surgical procedures, forming over 102,000 video-question-answer pairs. This dataset enables models to comprehend complex surgical workflows and engage in semantically grounded interactions.

4.3 Action Datasets

In the domain of execution-level research, numerous high-quality datasets have established a solid foundation for surgical action modeling, imitation learning, and complex scene understanding. The JIGSAWS dataset [315], collected using the da Vinci Surgical System, captures kinematic and video data from eight surgeons with varying skill levels performing three standardized tasks: suturing, knot-tying, and needle-passing. The CholecTrack20 dataset [282] focuses on laparoscopic cholecystectomy and provides multi-category, multi-tool, and multi-view tracking annotations, along with detailed labels for surgical phases, scene challenges, and operator identities. It supports comprehensive modeling of surgical workflows and instrument behavior recognition. The

Endoscapes dataset [316] includes 201 real-world laparoscopic surgery videos and covers tasks such as scene segmentation, object detection, and key safety view assessment, advancing capabilities in visual understanding and risk anticipation. MultiBypass140 [317] is a multi-center collection of Roux-en-Y gastric bypass videos that supports hierarchical recognition of surgical phases, steps, and adverse events, emphasizing the importance of procedural generalization and adaptation to complex intraoperative scenarios. The SurgVU24 dataset [318] originates from medical trainees using the da Vinci surgical robot for training and consists of approximately 280 long videos with over 18 million frames. It focuses on instrument recognition and surgical phase classification, providing rich samples for robotic surgery imitation learning and long-term strategy modeling. The m2cai16-workflow dataset [319] targets workflow modeling in laparoscopic cholecystectomy, comprising 41 real cases. As a core component of the M2CAI challenge, it has driven the evaluation and benchmarking of surgical phase recognition algorithms under real-world conditions. MM-OR is a large-scale, high-fidelity multimodal operating room dataset that enables multimodal scene graph generation. It integrates modalities such as RGB-D, audio, speech transcripts, robot logs, and tracking data, supporting semantic understanding and multitask modeling in surgical environments. GenSurgery [320] is built from publicly available laparoscopic and robotic surgical videos, comprising 680 hours of footage across 28 surgical types. It stands as one of the largest surgical vision datasets currently available. The MITI dataset [321] captures multimodal sensor data such as IMU, stereo video, and infrared tracking collected throughout a complete handheld surgical intervention. It provides a standardized benchmark for intraoperative action modeling and SLAM algorithm evaluation. The CoPESD dataset [322] focuses on endoscopic submucosal dissection (ESD), containing 17,679 images and 88,395 hierarchical action annotations extracted from 35 hours of video. It supports fine-grained recognition of complex submucosal manipulations and facilitates the training of vision-language models in surgical contexts.

4.4 Simulation Platforms and Synthetic Datasets

4.4.1 Surgical Simulation Platforms

In the advancement of intelligent surgical robotics, high-fidelity and scalable simulation platforms play a crucial role in supporting the training and evaluation of strategies such as reinforcement learning and imitation learning. In recent years, several open-source surgical simulation systems have been introduced, laying the foundation for the automation of complex surgical tasks and the transfer of learned skills (Table 3).

SurRoL [323] is an open-source platform for surgical robot learning that emphasizes high-fidelity interactive environments and reproducible strategy optimization processes. It integrates physical collision modeling, haptic feedback, and surgical instrument simulation, enabling realistic training for typical tasks such as grasping, cutting, and suturing. The platform also supports demonstration collection via human–robot interaction, facilitating the development of both imitation learning and reinforcement learning methods. ORBIT-Surgical [324] offers photorealistic rendering

Table 3 Simulation platforms for medical embodied AI.

Platform	Core Features	Supported Tasks / Scenarios	Learning Paradigms	Advantages	Limitations
SurRoL [317]	High-fidelity interaction; collision modeling; haptics; instrument simulation	Grasping, cutting, suturing; general surgical operations	Imitation learning; reinforcement learning; human–robot demonstration collection	Strong realism; supports diverse surgical tasks	Moderate visual realism; higher computation cost
ORBIT-Surgical [318]	Photorealistic rendering; precise physics; GPU-parallel engine	Surgical dexterity; active perception	Reinforcement learning; imitation learning	Excellent visual fidelity; efficient GPU simulation	High hardware demand; complex environment setup
Surgical Gym [319]	Fully GPU-based; extremely fast training	Large-scale reinforcement learning training; rapid policy iteration	Reinforcement learning	Very high training speed; ideal for large-scale reinforcement learning	Lower physical/visual realism vs. photorealistic engines
LapGym [320]	RALS-focused; flexible environment tools; RL modules; sensor simulation	Laparoscopic surgery; path planning; human-in-the-loop control	Imitation learning; reinforcement learning; policy generalization	Highly extensible; supports multimodal trajectories	Narrow task library; fidelity depends on user-defined setup
SonoGym [321]	Real CT scans; anatomical labels; ultrasound navigation and intervention	Path planning; bone reconstruction; ultrasound-guided procedures	Reinforcement learning; imitation learning	High anatomical realism; strong for ultrasound tasks	Limited to ultrasound modality; fewer general surgical tasks

and high-precision physical interactions. It enables GPU-parallel training of reinforcement and imitation learning algorithms, making it suitable for research on surgical dexterity and active perception. Surgical Gym [325] is a high-performance, fully GPU-based open-source platform focused on improving training efficiency in surgical robot learning. Compared to conventional platforms, it achieves training speedups ranging from 100 to 5000 times, making it well-suited for rapid large-scale iterations of reinforcement learning algorithms. LapGym [326] targets robot-assisted laparoscopic surgery (RALS) scenarios and provides flexible task environment construction tools and reinforcement learning experiment modules. It supports functions such as path planning, human-in-the-loop control, and sensor simulation. By generating trajectories from multimodal perception data for both robotic and expert operations, it facilitates research in imitation learning and policy generalization. SonoGym [327] is an extended training platform for robotic ultrasound navigation and intervention. Based on real patient CT scans and 3D anatomical labels, it supports complex orthopedic tasks such as path planning, bone surface reconstruction, and ultrasound-guided interventions. The platform also serves as a benchmark for reinforcement learning, safe reinforcement learning, and imitation learning algorithms.

4.4.2 Synthetic Datasets

In the context of limited access to real-world medical data and strict privacy protection requirements, synthetic datasets have emerged as an important supplement for training and evaluating medical artificial intelligence models. In recent years, a number of high-quality synthetic medical datasets have been released, spanning various domains such as fundus imaging, pathology slides, electronic health records (EHRs), and multimodal health data.

SynFundus-1M [328] is currently the largest high-quality synthetic fundus image dataset, comprising over one million images across 11 disease categories. In addition to disease-level annotations, the dataset provides four readability grades for key anatomical regions, supporting fine-grained lesion analysis and image quality modeling. SNOW (Synthetic Nuclei and annOtation Wizard) [329] is a large-scale virtual nuclei dataset designed for breast cancer pathology image segmentation. It is constructed using a Synthetic Image Generator (SIG) and a Nuclei Annotator (NA), and contains 20,000 synthetic images with 1,448,522 labeled nuclei, enabling deep learning models to be trained and evaluated on synthetic histopathological data. COVID-19 10K and COVID-19 100K [330] are synthetic patient record datasets generated using the Synthea platform. Provided in CSV format, these datasets capture COVID-19–related symptoms, diagnoses, treatments, and follow-up processes, offering high-quality structured synthetic data for pandemic modeling and system-level simulation. Coherent DataSet [331] is a comprehensive synthetic multimodal medical dataset that integrates clinical structured data in FHIR format, DICOM medical images, genomic sequences, and physiological monitoring data such as electrocardiograms. By leveraging the FHIR framework, the dataset ensures consistency and linkage across data types, forming a complete virtual health record suitable for training and validating multimodal medical AI models.

5 Challenges and Outlook

Despite recent advances, embodied AI for medical applications still faces significant challenges for safe and reliable deployment in complex clinical environments. This section systematically examines these challenges from three core perspectives: perception, decision-making, and action, as well as potential directions for future research.

5.1 Insufficient Training Data and Perception Discrepancy

The medical embodied perception typically relies on large volumes of high-quality annotated data for model training and performance evaluation. This dependency is particularly pronounced in high-risk tasks such as intraoperative navigation, instrument recognition, tissue segmentation, and lesion detection, where the quantity, annotation precision, and diversity of data directly determine the system's perceptual robustness and operational safety. However, in real-world clinical settings, the acquisition of medical data is constrained by multiple factors, including ethical approvals, privacy protection, and the high cost of clinicians' time. These constraints often result

in sparse task coverage, class imbalance, and limited domain diversity within training datasets. The problem is further exacerbated in scenarios involving rare pathologies, boundary-level lesions, or cross-institutional deployment, where systems are prone to the well-known 'domain gap' between training and deployment environments. This discrepancy can lead to reduced recognition accuracy, model drift, and heightened clinical risk.

To address the issue of data scarcity, a variety of approaches have been proposed to mitigate the dependence on large-scale annotated datasets. Among them, synthetic data generation has emerged as a mainstream strategy. These methods significantly reduce data acquisition costs while enhancing controllability [328] [329]. Although these methods often yield promising results during training, synthetic samples typically lack the complex perturbations found in real-world clinical settings such as instrument occlusions, tissue adhesions, and blood-induced specular reflections. This leads to synthetic overfitting and poor generalization to real patient cases. On another front, domain adaptation techniques have attempted to bridge the domain gap between training and deployment through strategies like distribution alignment and style transfer, thereby improving the system's adaptability in practical applications [332]. However, these approaches often still require access to target domain data and struggle with challenges such as class shift, insufficient distribution overlap, and the presence of unseen classes. More recently, semi-supervised learning and federated learning paradigms have gained traction in medical perception research. The former enhances the utility of unlabeled data through mechanisms like pseudo-label iteration and consistency regularization [333]. The latter enables collaborative modeling across institutions without exposing raw data, thus addressing the dual challenges of domain inconsistency and privacy preservation [334].

Future research directions could focus on building a more open and flexible perception system from the dual perspectives of generalization and adaptability. On the one hand, exploring unified perception model architectures that span multiple modalities, devices, and tasks, referred to as a unified perception backbone, can facilitate collaborative enhancement of multi-task perception through shared knowledge representations. On the other hand, incorporating active learning mechanisms driven by uncertainty can improve data utilization efficiency in scenarios with limited samples, and continual or incremental learning approaches can enable models to adapt to new tasks and environments without forgetting previously acquired knowledge. Moreover, developing a closed-loop perception training framework that integrates data generation, knowledge transfer, privacy preservation, and task generalization is essential. This framework could combine synthetic and real data co-labeling strategies, simulator-driven self-supervised pretraining methods, and multi-resolution fusion architectures that support interpretable reasoning and error awareness. Collectively, these advancements will propel medical embodied perception systems from task-specific solutions toward generalized and adaptable paradigms.

5.2 Semantic Ambiguity and Multimodal Knowledge Fusion Difficulties

In embodied artificial intelligence for medicine, perceptual information extends beyond visual images to include preoperative assessment texts, electronic health records, intraoperative voice commands, haptic feedback, and other heterogeneous modalities. Each of these data sources encapsulates valuable clinical knowledge and operational intent. However, their inherent disparities in spatial resolution, semantic granularity, and temporal alignment pose significant challenges to seamless multimodal integration. These challenges become particularly pronounced in task-driven real-time semantic understanding scenarios, where the lack of standardized modality encoding protocols and adaptive contextual recognition mechanisms often leads to semantic ambiguities and intermodal conflicts. For instance, when a surgeon issues the voice command "needle," failure to correctly disambiguate whether the term refers to a surgical instrument, an operative action, or an anatomical structure may result in erroneous execution and pose serious risks to patient safety. Moreover, interference factors such as background noise in voice signals, ambiguity in textual inputs, and instability in haptic feedback further compound the complexity of multimodal data fusion in medical embodied systems.

Recent research increasingly focuses on the use of structured medical knowledge graphs as foundational resources for semantic completion and reasoning support. These efforts aim to construct a unified semantic space that bridges the gaps among language, vision, and procedural instructions. In parallel, large pre-trained medical language models [154] [177] have shown strong capabilities in aligning and reasoning across text and image modalities, leading to improved performance in standardized medical tasks. However, the external knowledge graphs used in these approaches often face limitations such as delayed updates and uneven coverage across regions and medical specialties [335] [336]. These issues may result in missing or misleading information during knowledge transfer. Furthermore, the performance of language models in specific medical subdomains remains constrained by the distribution and quality of training data, limiting their ability to represent complex surgical workflows and spatial relationships accurately. Most importantly, current modality alignment methods typically rely on offline static mappings or fixed label sets. These approaches are insufficient for adapting to frequent dynamic changes and contextual dependencies that arise during surgical procedures. As a result, the system's capacity to perceive and respond to complex semantic environments in real time remains limited.

Future research could prioritize the development of context-aware dynamic semantic alignment mechanisms to address the limitations of current static mapping approaches. A key objective is to establish multimodal semantic fusion frameworks that are capable of adapting to real-time contextual variations. For instance, combining graph neural networks with self-attention mechanisms can effectively capture complex intermodal dependencies and temporal dynamics, allowing for the modeling and correction of semantic transfer pathways more flexibly and accurately. Moreover, the introduction of causal reasoning and causal modeling can uncover the underlying relationships among different modalities. This strategy enhances the interpretability

and robustness of multimodal integration and reduces the likelihood of misinterpretation caused by noise or missing information. In parallel, reinforcement learning and meta-learning techniques can be employed to design adaptive multimodal perception agents that respond rapidly to intraoperative contextual changes and continuously improve their performance. Finally, building open and continuously evolving medical knowledge ecosystems through collaborative human-AI knowledge curation will support the development of perception systems with deep medical understanding and advanced reasoning capabilities.

5.3 Medical Reasoning Complexity and Uncertainty Modeling

In real-world clinical settings, surgical robots and medical embodied agents operate within highly dynamic and heterogeneous information environments, often characterized by incomplete data and noise interference. For example, intraoperative emergencies such as unexpected bleeding, anatomical variations, or equipment malfunctions require the system to perform rapid and accurate multi-step diagnostic reasoning and decision-making based on limited and partially missing perceptual evidence. These tasks extend far beyond the capabilities of traditional static classification or regression models and fall within the domain of Bayesian reasoning, where prior knowledge constraints and uncertainty quantification play central roles. The primary challenge lies in efficiently deriving optimal decision pathways while maintaining transparency and interpretability throughout the reasoning process. This is essential for fostering trust and acceptance among clinicians and patients.

Most current embodied artificial intelligence in medicine still relies on end-to-end deep reinforcement learning or policy learning approaches [24]. While these methods perform well in idealized simulation environments, they often lack explicit reasoning paths and interpretable intermediate states. As a result, the underlying logic of system decisions remains opaque, limiting the ability to diagnose and correct potential errors. In recent years, the development of multi-step reasoning medical question-answering datasets has facilitated research on stepwise reasoning and multi-hop logical inference. This progress has enhanced the interpretability of reasoning chains and improved clinical traceability. Some studies have integrated graph neural networks with medical knowledge graphs and employed rule-based engines to incorporate domain-specific constraints, leading to improved inference accuracy in complex pathological conditions. Additionally, the application of causal reasoning has provided theoretical foundations for identifying true causal relationships and eliminating spurious correlations, contributing to the development of more robust diagnostic models. However, several limitations remain. Constructing and maintaining comprehensive causal knowledge graphs in the medical domain is time-consuming and often insufficient to capture the full spectrum of pathological diversity. Although graph neural networks and large language models offer powerful representational capabilities, they typically rely on limited annotations and weak supervision signals, making it difficult to fully model rare diseases and individualized pathological features.

Future research could focus on developing hybrid reasoning frameworks that integrate causal knowledge graphs, expert-derived rule systems, and reinforcement learning. Such integration can leverage the complementary strengths of each approach

to enable dynamic generation and real-time correction of reasoning paths. In particular, enhancing human–AI interaction mechanisms by incorporating expert-in-the-loop feedback is essential, allowing models to continuously adapt and self-improve in clinical practice. Moreover, the use of meta-learning and transfer learning strategies can help improve generalization to rare cases and emerging diseases. It is also imperative to establish a reasoning safety assurance framework grounded in the principles of trustworthy artificial intelligence, ensuring that the inference process remains transparent and controllable within ethical, regulatory, and clinical safety constraints. Finally, interdisciplinary collaboration is needed to build an open, evolving medical causal knowledge base. The development of automated knowledge extraction and update techniques will further support the long-term sustainability and clinical applicability of reasoning-based models.

5.4 Lack of Mechanisms for Decision Pathway Generation and Validation

In embodied artificial intelligence in medicine, the effective translation of reasoning outcomes into executable and safe action policies is of critical importance, particularly within dynamic and complex surgical environments and multi-stage diagnostic and therapeutic workflows. Each action carries inherent high risks, and even minor deviations or policy errors may result in severe medical incidents or patient harm. However, most current approaches based on reinforcement learning and imitation learning primarily focus on efficiently generating action sequences under predefined reward mechanisms, while lacking systematic mechanisms for verifying the consequences of policy execution or conducting thorough risk assessments. As a result, it remains challenging to ensure the safety and robustness of these policies in real-world clinical applications.

Existing methods for policy optimization in medical embodied systems primarily rely on two major approaches: expert demonstration and reward shaping [337]. Expert demonstrations guide models by providing high-quality trajectory data that exemplify correct behavior patterns, while reward shaping encourages desired actions through carefully designed reward functions. These approaches have shown promising results in controlled environments, particularly for standardized and repetitive surgical tasks. However, significant limitations persist. The quality of expert demonstrations is often difficult to guarantee as comprehensive and error-free. This becomes especially problematic in the context of novel or rare pathological conditions, where expert knowledge may be insufficient, resulting in models that fail to recognize policy deficiencies. Reward function design typically depends on handcrafted heuristics, making it difficult to account for all potential risks and lacking the flexibility for dynamic adjustment. Moreover, there is a notable absence of well-established "gold standard" policy trajectories for post-hoc verification, leaving current models without mechanisms for self-assessment or error correction of their decisions. This increases the risk exposure of policy deployment in real-world clinical scenarios.

Future research directions may focus on establishing multi-agent adversarial verification frameworks by introducing adversarial agents or counterfactual agents into

virtual surgical environments. These agents can simulate extreme or atypical scenarios to proactively expose and challenge the vulnerabilities of primary policies, thereby encouraging the development of more robust and safer behavioral strategies. In parallel, adopting the concept of counterfactual validation, researchers can construct synthetic verification platforms encompassing rare and potentially hazardous scenarios to enable comprehensive and systematic evaluation of policy performance. The advancement of digital twin technology further reinforces this objective by enabling high-fidelity, dynamically synchronized digital replicas of patients and surgical environments. Such digital twins not only allow for retrospective analysis and replay of historical procedures but also support real-time simulation of potential future risks, facilitating a closed-loop system for policy generation and safety validation. In addition, future work may explore the integration of formal verification methods with deep reinforcement learning to provide mathematical guarantees of policy safety and enable real-time monitoring. This would help ensure the predictability and reliability of learned strategies in high-stakes clinical applications.

5.5 Error Sensitivity in High-Precision Action Control

At the action level of embodied artificial intelligence in medicine, high-precision operations impose extremely stringent requirements on trajectory control and response latency. This is particularly critical in minimally invasive procedures or when operating near highly sensitive anatomical structures, where even millimeter- or sub-millimeter-level deviations can result in tissue damage, functional impairment, or complete surgical failure. Such minute errors often arise from a combination of factors, including limited mechanical stiffness of robotic arms, micro-vibrations in transmission mechanisms, dynamic path variations, sensor latency, and insufficient resolution. Unfortunately, most current execution systems lack comprehensive system-level fault tolerance. They predominantly rely on static motion control algorithms, which are insufficient for enabling a dynamic closed-loop control paradigm encompassing early warning, real-time response, and active compensation. As a result, these systems struggle to adapt to the transient changes and complex perturbations frequently encountered during surgical procedures.

Existing studies aiming to improve execution control accuracy primarily rely on visual servoing closed-loop feedback mechanisms, which adjust robotic arm trajectories in real time based on image error signals. Concurrently, imitation learning and expert demonstration trajectories are employed to generate control policies, leveraging historical data to construct smooth and plausible operation paths that mitigate execution errors. Some approaches further incorporate trajectory smoothing algorithms and outlier detection techniques to enhance the control system's robustness against sudden perturbations. However, these methods are generally constrained by latency discrepancies between perception and control loops. The feedback frequency and sensor response time of visual perception systems often fail to meet the demands of high-speed dynamic operations, resulting in delayed closed-loop control responses or even introducing secondary errors. Additionally, due to limitations in sensor resolution and calibration accuracy, current systems remain insufficient in capturing fine-grained

force and tactile feedback, hindering comprehensive perception and effective response to the complex and rapidly changing intraoperative environment.

To address the aforementioned challenges, future research could focus on developing highly robust control systems that deeply integrate hardware and software components. On the one hand, multimodal perceptual data including visual, force, tactile, temperature, and displacement signals could be comprehensively fused to enhance fine-grained perception of surgical instruments and tissue states, enabling multidimensional dynamic monitoring of the operational process. On the other hand, leveraging embedded computing and edge intelligence technologies can realize high-frequency, localized feedback control loops, significantly reducing response latency and improving the real-time performance and stability of closed-loop control. At the algorithmic level, exploring hybrid control frameworks that combine model predictive control and deep learning based uncertainty estimation can facilitate dynamic prediction and compensation of operational errors. Simultaneously, the development of robust control models with error tracking, dynamic adjustment and safety pruning capabilities is necessary to enable real-time detection and adaptive correction of abnormal events. Furthermore, adaptive methods such as reinforcement learning can be employed to optimize control policies while digital twin technologies provide simulation and validation environments that support safe evaluation and continuous improvement of control strategies.

5.6 Lack of General-Purpose Medical Simulation Platforms

The training and validation of execution capabilities in medical embodied AI currently rely heavily on simulation platforms to facilitate policy iteration, motion pretraining, and risk mitigation testing. However, compared to industrial automation or service robotics, medical scenarios present significantly greater complexity and dynamic variability. Anatomical structures during surgery exhibit substantial individual differences, surgical procedures are diverse and involve complex execution sequences, and varying surgeon styles and personalized habits further increase the challenges of system generalization. Existing simulation platforms are typically tailored to specific surgeries or tasks and lack a universal framework that supports cross-scenario, cross-modal, and cross-task hierarchies. This limitation prevents comprehensive coverage of the diverse challenges and uncertainties inherent in real surgical environments during training. Moreover, these platforms demonstrate limited adaptability to procedural variations, individual surgeon differences, and specialized instrument trajectories, which directly impact the transfer effectiveness and operational stability when transitioning from simulated training to real-world deployment.

In recent years, open-source platforms have made preliminary progress in multi-scenario surgical motion simulation and trajectory replay, promoting the standardization of datasets and the development of an open ecosystem for medical embodied AI training [325] [326]. Concurrently, some studies have attempted to achieve more realistic operation reproduction by integrating surgical videos, motion trajectories, and real-time sensor data, enhancing the capability for personalized procedure modeling and improving the fidelity between simulation and actual surgical operations. Nonetheless, these platforms continue to face fundamental challenges. First, existing physics engines exhibit limited accuracy in simulating complex physical processes such

as tissue elasticity deformation, instrument-tissue contact mechanics, and blood flow, making it difficult to faithfully reproduce the fine-grained interactions within surgical environments. Second, there is a lack of flexible and automated mechanisms for procedural task definition and generation of abnormal states, resulting in insufficient representativeness and coverage of training data. Third, the behavioral models tend to be shallow in hierarchy, impeding the effective decomposition of high-level surgical intents into low-level execution commands, thus restricting advanced policy learning and cross-task generalization capabilities.

Future research could prioritize the development of a universal medical embodied AI simulation platform centered on real-world closed-loop validation to overcome existing limitations. First, it is essential to integrate personalized physiological atlases with multimodal imaging data to construct highly detailed and reconstructable surgical anatomy models that support multi-level task customization and microsurgical operation simulation. Second, advanced procedural scene generation and task scheduling mechanisms could be designed to enable automatic simulation of diverse and dynamic surgical workflows as well as emergent anomalies, facilitating seamless transitions from basic maneuvers to complex procedures and supporting multi-task training. Finally, a closed-loop system for collaborative verification between simulation and physical platforms could be established. Leveraging digital twin technology and online transfer learning, this system would enable rapid iteration and optimization of strategies in high-fidelity simulated environments, while allowing real-time adaptation and fine-tuning on actual execution platforms, thereby ensuring the clinical applicability, safety, and reliability of trained models.

6 Conclusion

Embodied AI introduces a transformative paradigm to healthcare, effectively bridging the critical gap between computational foundation models and the physical clinical world. This review has provided a comprehensive survey of the field, systematically analyzing the core components of perception, decision-making, and action, while cataloging representative medical applications and essential datasets. Despite the promising progress, we also highlighted the significant challenges in clinical settings. By elucidating the current landscape and identifying key bottlenecks, this work aims to serve as a foundational roadmap. It is our hope that this survey will facilitate researchers in addressing these limitations, ultimately accelerating the transition of intelligent agents from theoretical frameworks to practical, reliable assistants in real-world medical workflows.

References

[1] Varghese, C., Harrison, E.M., O'Grady, G., Topol, E.J.: Artificial intelligence in surgery. Nature medicine **30**(5), 1257–1268 (2024)

[2] Thieme, A.H., Zheng, Y., Machiraju, G., Sadee, C., Mittermaier, M., Gertler, M., Salinas, J.L., Srinivasan, K., Gyawali, P., Carrillo-Perez, F., *et al.*: A deep-learning algorithm to classify skin lesions from mpox virus infection. Nature

medicine **29**(3), 738–747 (2023)

[3] Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)

[4] Ma, D., Pang, J., Gotway, M.B., Liang, J.: A fully open ai foundation model applied to chest radiography. Nature, 1–11 (2025)

[5] Liu, F., Zhou, H., Gu, B., Zou, X., Huang, J., Wu, J., Li, Y., Chen, S.S., Hua, Y., Zhou, P., et al.: Application of large language models in medicine. Nature Reviews Bioengineering, 1–20 (2025)

[6] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., *et al.*: Toward expert-level medical question answering with large language models. Nature Medicine **31**(3), 943–950 (2025)

[7] Liu, X., Liu, H., Yang, G., Jiang, Z., Cui, S., Zhang, Z., Wang, H., Tao, L., Sun, Y., Song, Z., *et al.*: A generalist medical language model for disease diagnosis assistance. Nature medicine **31**(3), 932–942 (2025)

[8] Liu, Y., Chen, W., Bai, Y., Liang, X., Li, G., Gao, W., Lin, L.: Aligning cyber space with physical world: A comprehensive survey on embodied ai. IEEE/ASME Transactions on Mechatronics (2025)

[9] Li, J., Xu, Z., Li, N., Zhang, K., Xiong, G., Sun, M., Hou, C., Ji, J., Zhang, F., Zhong, J., *et al.*: Ai-embodied multi-modal flexible electronic robots with programmable sensing, actuating and self-learning. Nature Communications **16**(1), 8818 (2025)

[10] Long, Y., Lin, A., Kwok, D.H.C., Zhang, L., Yang, Z., Shi, K., Song, L., Fu, J., Lin, H., Wei, W., *et al.*: Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. Science Robotics **10**(104), 3093 (2025)

[11] Fiorini, P., Goldberg, K.Y., Liu, Y., Taylor, R.H.: Concepts and trends in autonomy for robot-assisted surgery. Proceedings of the IEEE **110**(7), 993–1011 (2022)

[12] Yao, T., Wang, H., Lu, B., Ge, J., Pei, Z., Kowarschik, M., Sun, L., Seneviratne, L., Qi, P.: Sim2real learning with domain randomization for autonomous guidewire navigation in robotic-assisted endovascular procedures. IEEE Transactions on Automation Science and Engineering (2025)

[13] Yao, T., Xu, Y., Wang, H., Qiu, X., Althoefer, K., Qi, P.: Multi-agent fuzzy reinforcement learning with llm for cooperative navigation of endovascular robotics.

IEEE Transactions on Fuzzy Systems (2025)

[14] Pore, A., Li, Z., Dall'Alba, D., Hernansanz, A., De Momi, E., Menciassi, A., Gelpi, A.C., Dankelman, J., Fiorini, P., Vander Poorten, E.: Autonomous navigation for robot-assisted intraluminal and endovascular procedures: A systematic review. IEEE Transactions on Robotics **39**(4), 2529–2548 (2023)

[15] Song, J., Yang, K., Chen, H., Liu, J., Gu, Y., Hui, Q., Huang, Y., Li, M., Zhang, Z., Cao, T., *et al.*: Vascularpilot3d: Toward a 3d fully autonomous navigation for endovascular robotics. In: 2025 IEEE International Conference on Robotics and Automation (ICRA), pp. 9318–9324 (2025). IEEE

[16] Song, J., Zhang, R., Zhang, W., Zhou, H., Ghaffari, M.: Slam assisted 3d tracking system for laparoscopic surgery. In: 2025 IEEE International Conference on Robotics and Automation (ICRA), pp. 6868–6874 (2025). IEEE

[17] Alla, S., Bheesetty, N., Park, H.J.: Informative path planning for nano-surgical robot adaptive drug delivery. In: 2025 59th Annual Conference on Information Sciences and Systems (CISS), pp. 1–6 (2025). IEEE

[18] Arreola, W., Rivas, J.J., Castrejon, L., Sucar, L.E.: Affective embodied agent for patient assistance in virtual rehabilitation. IEEE Transactions on Affective Computing (2025)

[19] Chengjie, Z., Suiran, Y.: Virtual co-embodiment rehabilitation: An innovative method integrating virtual co-embodiment and action observation therapy in virtual reality rehabilitation. In: 2024 17th International Convention on Rehabilitation Engineering and Assistive Technology (i-CREATe), pp. 1–6 (2024). IEEE

[20] Jiang, Z., Huang, X., Wang, Z., Liu, Y., Huang, L., Luo, X.: Embodied conversational agents for chronic diseases: scoping review. Journal of Medical Internet Research **26**, 47134 (2024)

[21] Diana, M., Marescaux, J.: Robotic surgery. Journal of British Surgery **102**(2), 15–28 (2015)

[22] Di, J., Dugonjic, Z., Fu, W., Wu, T., Mercado, R., Sawyer, K., Most, V.R., Kammerer, G., Speidel, S., Fan, R.E., et al.: Using fiber optic bundles to miniaturize vision-based tactile sensors. IEEE Transactions on Robotics (2024)

[23] Luo, S., Jiang, M., Zhang, S., Zhu, J., Yu, S., Dominguez Silva, I., Wang, T., Rouse, E., Zhou, B., Yuk, H., *et al.*: Experiment-free exoskeleton assistance via learning in simulation. Nature **630**(8016), 353–359 (2024)

[24] Sharifi, M., Tripathi, S., Chen, Y., Zhang, Q., Tavakoli, M.: Reinforcement learning methods for assistive and rehabilitation robotic systems: A survey. IEEE

Transactions on Systems, Man, and Cybernetics: Systems (2025)

[25] Wang, Z., Wu, Q., Zhu, Y., Wang, J., Luo, C., Huang, S., Chen, B.: based therapist skill transfer learning framework for upper-limb rehabilitation exoskeleton. In: 2025 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), pp. 1–6 (2025). IEEE

[26] Ben Abdallah, I., Bouteraa, Y., Alotaibi, A.: Ai-driven hybrid rehabilitation: synergizing robotics and electrical stimulation for upper-limb recovery after stroke. Frontiers in Bioengineering and Biotechnology **13**, 1619247 (2025)

[27] Fragapane, G., Hvolby, H.-H., Sgarbossa, F., Strandhagen, J.O.: Autonomous mobile robots in hospital logistics. In: IFIP International Conference on Advances in Production Management Systems, pp. 672–679 (2020). Springer

[28] Bernhard, L., Schwingenschlögl, P., Hofmann, J., Wilhelm, D., Knoll, A.: Boosting the hospital by integrating mobile robotic assistance systems: a comprehensive classification of the risks to be addressed. Autonomous Robots **48**(1), 1 (2024)

[29] Ding, W., Tian, Q., Xia, Y., Yang, Y., Wang, Y., Zhang, Y.: Research on multirobot collaboration platform for logistic distribution of medical consumables in the operating room. In: Third International Conference on Biomedical and Intelligent Systems (IC-BIS 2024), vol. 13208, pp. 637–642 (2024). SPIE

[30] Palinko, O., Wendlandt, R., Udby, S., Uhing, F., Fog, J.H., Hansen, E., Junge, R.P., Holm, D.G., Kipp, M., Bodenhagen, L.: Interaction matters when it comes to hand disinfection using robots at hospitals. In: International Conference on Social Robotics, pp. 74–85 (2024). Springer

[31] Liang, W., Zhou, R., Ma, Y., Zhang, B., Li, S., Liao, Y., Kuang, P.: Large model empowered embodied ai: A survey on decision-making and embodied learning. arXiv preprint arXiv:2508.10399 (2025)

[32] Xing, W., Li, M., Li, M., Han, M.: Towards robust and secure embodied ai: A survey on vulnerabilities and attacks. arXiv preprint arXiv:2502.13175 (2025)

[33] Sapkota, R., Roumeliotis, K.I., Karkee, M.: Uavs meet agentic ai: A multidomain survey of autonomous aerial intelligence and agentic uavs. arXiv preprint arXiv:2506.08045 (2025)

[34] Feng, T., Wang, X., Jiang, Y.-G., Zhu, W.: Embodied ai: From llms to world models. arXiv preprint arXiv:2509.20021 (2025)

[35] Wang, Y., Chen, S., Li, Z., Shen, T., Wang, K.: Embodied intelligent driving: Key technologies and applications. In: 2024 IEEE 4th International Conference on Digital Twins and Parallel Intelligence (DTPI), pp. 132–137 (2024). IEEE

[36] Wu, Y., Li, D., Chen, Y., Jiang, R., Zou, H.P., Fang, L., Wang, Z., Yu, P.S.: Multi-agent autonomous driving systems with large language models: A survey of recent advances. arXiv preprint arXiv:2502.16804 (2025)

[37] Liu, Y., Cao, X., Chen, T., Jiang, Y., You, J., Wu, M., Wang, X., Feng, M., Jin, Y., Chen, J.: From screens to scenes: A survey of embodied ai in healthcare. Information Fusion **119**, 103033 (2025)

[38] Turing, A.: Computing machinery and intelligence-am turing. Mind **59**(236), 433 (1950)

[39] Liu, J., Shi, X., Nguyen, T.D., Zhang, H., Zhang, T., Sun, W., Li, Y., Vasilakos, A.V., Iacca, G., Khan, A.A., et al.: Neural brain: A neuroscience-inspired framework for embodied agents. arXiv preprint arXiv:2505.07634 (2025)

[40] Liu, H., Guo, D., Cangelosi, A.: Embodied intelligence: A synergy of morphology, action, perception and learning. ACM Computing Surveys **57**(7), 1–36 (2025)

[41] Paolo, G., Gonzalez-Billandon, J., Kégl, B.: Position: a call for embodied ai. In: Forty-first International Conference on Machine Learning (2024)

[42] Wang, B., Meng, X., Wang, X., Zhu, Z., Ye, A., Wang, Y., Yang, Z., Ni, C., Huang, G., Wang, X.: Embodiedreamer: Advancing real2sim2real transfer for policy training via embodied world modeling. arXiv preprint arXiv:2507.05198 (2025)

[43] Jiang, T., Guan, Y., Ma, L., Xu, J., Meng, J., Chen, W., Zeng, Z., Li, L., Wu, D., Chen, R.: Dexsim2real^{2}: Building explicit world model for precise articulated object dexterous manipulation. arXiv preprint arXiv:2409.08750 (2024)

[44] Yardi, Y., Biruduganti, S., Ankile, L.: Bridging the sim2real gap: Vision encoder pre-training for visuomotor policy transfer. arXiv preprint arXiv:2501.16389 (2025)

[45] Liu, G., Deng, Y., Zhao, R., Zhou, H., Chen, J., Chen, J., Xu, R., Tai, Y., Jia, K.: Dexscale: Automating data scaling for sim2real generalizable robot control. In: Forty-second International Conference on Machine Learning

[46] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)

[47] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[48] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern

Recognition, pp. 770–778 (2016)

[49] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)

[50] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)

[51] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[52] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)

[53] Li, S., Chan, A.B.: 3d human pose estimation from monocular images with deep convolutional neural network. In: Asian Conference on Computer Vision, pp. 332–347 (2014). Springer

[54] Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.-F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14393–14402 (2021)

[55] Gao, M., Xing, C., Niebles, J.C., Li, J., Xu, R., Liu, W., Xiong, C.: Open vocabulary object detection with pseudo bounding-box labels. In: European Conference on Computer Vision, pp. 266–282 (2022). Springer

[56] Qiu, C., Zhang, X., Tong, X., Guan, N., Yi, X., Yang, K., Zhu, J., Yu, A.: Few-shot remote sensing image scene classification: Recent advances, new baselines, and future trends. ISPRS Journal of Photogrammetry and Remote Sensing **209**, 368–382 (2024)

[57] Liu, J., Yu, Z., Breckon, T.P., Shum, H.P.: U3ds3: Unsupervised 3d semantic scene segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3759–3768 (2024)

[58] An, D., Wang, H., Wang, W., Wang, Z., Huang, Y., He, K., Wang, L.: Etpnav: Evolving topological planning for vision-language navigation in continuous environments. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

[59] Garg, S., Rana, K., Hosseinzadeh, M., Mares, L., Sünderhauf, N., Dayoub, F., Reid, I.: Robohop: Segment-based topological map representation for open-world visual navigation. In: 2024 IEEE International Conference on Robotics

and Automation (ICRA), pp. 4090–4097 (2024). IEEE

[60] Zheng, X., Liao, C., Weng, Z., Lei, K., Dongfang, Z., He, H., Lyu, Y., Jiang, L., Qi, L., Chen, L., et al.: Panorama: The rise of omnidirectional vision in the embodied ai era. arXiv preprint arXiv:2509.12989 (2025)

[61] Liu, P., Feng, C., Xu, Y., Ning, Y., Xu, H., Shen, S.: Omninxt: A fully open-source and compact aerial robot with omnidirectional visual perception. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10605–10612 (2024). IEEE

[62] Arora, R., Singh, S., Swaminathan, K., Datta, A., Banerjee, S., Bhowmick, B., Jatavallabhula, K.M., Sridharan, M., Krishna, M.: Anticipate & act: Integrating llms and classical planning for efficient task execution in household environments. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 14038–14045 (2024). IEEE

[63] Guo, W., Kingston, Z., Kavraki, L.E.: Castl: Constraints as specifications through llm translation for long-horizon task and motion planning. In: 2025 IEEE International Conference on Robotics and Automation (ICRA), pp. 11957–11964 (2025). IEEE

[64] Kim, M., Kim, H.-I., Ro, Y.M.: Prompt tuning of deep neural networks for speaker-adaptive visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

[65] Gan, Y., Xu, L., Song, S., Tao, X.: Context transformer with multiscale fusion for robust facial emotion recognition. Pattern Recognition, 111720 (2025)

[66] Dai, P., Zhou, J., Ma, J., Zhang, H., Wu, X.: Meta-transfer learning based cross-domain gesture recognition using wifi channel state information. IEEE Transactions on Consumer Electronics (2025)

[67] Kyranou, I., Szymaniak, K., Nazarpour, K.: Emg dataset for gesture recognition with arm translation. Scientific Data **12**(1), 100 (2025)

[68] Yang, Q., Shi, Q., Wang, T., Ye, M.: Uncertain multimodal intention and emotion understanding in the wild. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 24700–24709 (2025)

[69] Silver, T., Chitnis, R., Tenenbaum, J., Kaelbling, L.P., Lozano-Pérez, T.: Learning symbolic operators for task and motion planning. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3182–3189 (2021). IEEE

[70] Fox, M., Long, D.: Pddl2. 1: An extension to pddl for expressing temporal planning domains. Journal of artificial intelligence research **20**, 61–124 (2003)

[71] Liu, Y., Palmieri, L., Koch, S., Georgievski, I., Aiello, M.: Delta: Decomposed efficient long-term robot task planning using large language models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA), pp. 10995–11001 (2025). IEEE

[72] Kwon, M., Kim, Y., Kim, Y.J.: Fast and accurate task planning using neuro-symbolic language models and multi-level goal decomposition. In: 2025 IEEE International Conference on Robotics and Automation (ICRA), pp. 16195–16201 (2025). IEEE

[73] Lei, X., Wang, M., Zhou, W., Li, H.: Gaussnav: Gaussian splatting for visual navigation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

[74] Chen, K., An, D., Huang, Y., Xu, R., Su, Y., Ling, Y., Reid, I., Wang, L.: Constraint-aware zero-shot vision-language navigation in continuous environments. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)

[75] Saxena, S., Buchanan, B., Paxton, C., Liu, P., Chen, B., Vaskevicius, N., Palmieri, L., Francis, J., Kroemer, O.: Grapheqa: Using 3d semantic scene graphs for real-time embodied question answering. arXiv preprint arXiv:2412.14480 (2024)

[76] Majumdar, A., Ajay, A., Zhang, X., Putta, P., Yenamandra, S., Henaff, M., Silwal, S., Mcvay, P., Maksymets, O., Arnaud, S., *et al.*: Openeqa: Embodied question answering in the era of foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16488–16498 (2024)

[77] Foster, D.J., Block, A., Misra, D.: Is behavior cloning all you need? understanding horizon in imitation learning. Advances in Neural Information Processing Systems **37**, 120602–120666 (2024)

[78] Mehta, S.A., Ciftci, Y.U., Ramachandran, B., Bansal, S., Losey, D.P.: Stable-bc: Controlling covariate shift with stable behavior cloning. IEEE Robotics and Automation Letters (2025)

[79] Beliaev, M., Pedarsani, R.: Inverse reinforcement learning by estimating expertise of demonstrators. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 15532–15540 (2025)

[80] Yue, B., Wang, S., Gaurav, A., Li, J., Poupart, P., Liu, G.: Understanding constraint inference in safety-critical inverse reinforcement learning. In: The Thirteenth International Conference on Learning Representations (2025)

[81] Liu, Z. Value-based reinforcement learning. In: Artificial Intelligence for Engineers: Basics and Implementations, pp. 337–355. Springer, ??? (2025)

[82] Watkins, C.J., Dayan, P.: Q-learning. Machine learning **8**(3), 279–292 (1992)

[83] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

[84] Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning, pp. 1928–1937 (2016). PmLR

[85] Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015)

[86] Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International Conference on Machine Learning, pp. 1861–1870 (2018). Pmlr

[87] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

[88] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., *et al.*: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)

[89] Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., *et al.*: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning, pp. 2165–2183 (2023). PMLR

[90] Fan, L., Liang, M., Li, Y., Hua, G., Wu, Y.: Evidential active recognition: Intelligent and prudent open-world embodied perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16351–16361 (2024)

[91] Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., *et al.*: Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19757–19767 (2024)

[92] Du, H., Ren, L., Wang, Y., Cao, X., Sun, C.: Advancements in perception system with multi-sensor fusion for embodied agents. Information Fusion, 102859 (2024)

[93] Sun, Y., Cheng, N., Zhang, S., Li, W., Yang, L., Cui, S., Liu, H., Sun, F., Zhang, J., Di, G., et al.: Tactile data generation and applications based on visuo-tactile sensors: A review. Information Fusion, 103162 (2025)

[94] Braud, R., Giagkos, A., Shaw, P., Lee, M., Shen, Q.: Robot multimodal object perception and recognition: Synthetic maturation of sensorimotor learning in embodied systems. IEEE Transactions on Cognitive and Developmental Systems **13**(2), 416–428 (2020)

[95] Dutta, A., Burdet, E., Kaboli, M.: Predictive visuo-tactile interactive perception framework for object properties inference. IEEE Transactions on Robotics (2025)

[96] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (2002)

[97] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[98] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., *et al.*: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)

[99] Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)

[100] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660 (2021)

[101] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)

[102] Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)

[103] Wang, W., Duan, C., Peng, Z., Liu, Y., Zhou, B.: Embodied scene understanding for vision language models via metavqa. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 22453–22464 (2025)

[104] Zhou, Y., Huang, L., Bu, Q., Zeng, J., Li, T., Qiu, H., Zhu, H., Guo, M., Qiao, Y., Li, H.: Embodied understanding of driving scenarios. In: European Conference on Computer Vision, pp. 129–148 (2024). Springer

[105] Fu, R., Liu, J., Chen, X., Nie, Y., Xiong, W.: Scene-llm: Extending language model for 3d visual reasoning. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2195–2206 (2025). IEEE

[106] Jia, B., Chen, Y., Yu, H., Wang, Y., Niu, X., Liu, T., Li, Q., Huang, S.: Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In: European Conference on Computer Vision, pp. 289–310 (2024). Springer

[107] Sun, L.C., Bhatt, N.P., Liu, J.C., Fan, Z., Wang, Z., Humphreys, T.E., Topcu, U.: Mm3dgs slam: Multi-modal 3d gaussian splatting for slam using vision, depth, and inertial measurements. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 10159–10166 (2024). IEEE

[108] Wei, X., Yu, R., Sun, J.: Learning view-based graph convolutional network for multi-view 3d shape analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 7525–7541 (2022)

[109] Dentamaro, V., Gattulli, V., Impedovo, D., Manca, F.: Human activity recognition with smartphone-integrated sensors: A survey. Expert Systems with Applications **246**, 123143 (2024)

[110] Yin, H., Sinnott, R.O., Jayaputera, G.T.: A survey of video-based human action recognition in team sports. Artificial intelligence review **57**(11), 293 (2024)

[111] Cob-Parro, A.C., Losada-Gutiérrez, C., Marrón-Romera, M., Gardel-Vicente, A., Bravo-Munoz, I.: A new framework for deep learning video based human action recognition on the edge. Expert Systems with Applications **238**, 122220 (2024)

[112] Zhou, Y., Yan, X., Cheng, Z.-Q., Yan, Y., Dai, Q., Hua, X.-S.: Blockgcn: Redefine topology awareness for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2049–2058 (2024)

[113] Tang, Y., Wang, W., Zhang, C., Liu, J., Zhao, Y.: Learnable feature augmentation framework for temporal action localization. IEEE Transactions on Image Processing **33**, 4002–4015 (2024)

[114] Liu, D., Meng, F., Mi, J., Ye, M., Li, Q., Zhang, J.: Sam-net: Semantic-assisted multimodal network for action recognition in rgb-d videos. Pattern Recognition, 111725 (2025)

[115] Zhang, J., Wan, Z., Hu, L., Lin, S., Wu, S., Shan, S.: Collaboratively self-supervised video representation learning for action recognition. IEEE Transactions on Information Forensics and Security (2025)

[116] Zhao, G., Li, X., Li, Y., Pietikäinen, M.: Facial micro-expressions: An overview. Proceedings of the IEEE **111**(10), 1215–1235 (2023)

[117] Karnati, M., Seal, A., Bhattacharjee, D., Yazidi, A., Krejcar, O.: Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. IEEE Transactions on Instrumentation and Measurement **72**, 1–31 (2023)

[118] Wu, C., Cai, Y., Liu, Y., Zhu, P., Xue, Y., Gong, Z., Hirschberg, J., Ma, B.: Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. arXiv preprint arXiv:2505.20511 (2025)

[119] Liu, S., Mao, X., Zhao, S., Li, P., Xu, T., Chen, E.: Mer-clip: Au-guided vision-language alignment for micro-expression recognition. IEEE Transactions on Affective Computing (2025)

[120] Jin, W., Du, H., Zhao, B., Tian, X., Shi, B., Yang, G.: A comprehensive survey on multi-agent cooperative decision-making: Scenarios, approaches, challenges and perspectives. arXiv preprint arXiv:2503.13415 (2025)

[121] Guo, H., Wu, F., Qin, Y., Li, R., Li, K., Li, K.: Recent trends in task and motion planning for robotics: A survey. ACM Computing Surveys **55**(13s), 1–36 (2023)

[122] Zhao, Z., Cheng, S., Ding, Y., Zhou, Z., Zhang, S., Xu, D., Zhao, Y.: A survey of optimization-based task and motion planning: From classical to learning approaches. IEEE/ASME Transactions on Mechatronics (2024)

[123] Yang, Z., Garrett, C., Fox, D., Lozano-Pérez, T., Kaelbling, L.P.: Guiding long-horizon task and motion planning with vision language models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA), pp. 16847–16853 (2025). IEEE

[124] Liu, Y., Liu, L., Zheng, Y., Liu, Y., Dang, F., Li, N., Ma, K.: Embodied navigation. Science China Information Sciences **68**(4), 1–39 (2025)

[125] Wu, Y., Zhang, P., Gu, M., Zheng, J., Bai, X.: Embodied navigation with multimodal information: A survey from tasks to methodology. Information Fusion **112**, 102532 (2024)

[126] Deng, T., Shen, G., Xun, C., Yuan, S., Jin, T., Shen, H., Wang, Y., Wang, J., Wang, H., Wang, D., *et al.*: Mne-slam: Multi-agent neural slam for mobile robots. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 1485–1494 (2025)

[127] Hu, G., Huang, F., Shu, B., Wei, G.: Mahaco: Multi-algorithm hybrid ant colony optimizer for 3d path planning of a group of uavs. Information Sciences **694**, 121714 (2025)

[128] Sun, J., Wu, J., Ji, Z., Lai, Y.-K.: A survey of object goal navigation. IEEE Transactions on Automation Science and Engineering **22**, 2292–2308 (2024)

[129] Huang, Z., Yang, Z., Krupani, R., Şenbaşlar, B., Batra, S., Sukhatme, G.S.: Collision avoidance and navigation for a quadrotor swarm using end-to-end deep reinforcement learning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 300–306 (2024). IEEE

[130] Lin, X., Karapetyan, N., Joshi, K., Liu, T., Chopra, N., Yu, M., Tokekar, P., Aloimonos, Y.: Uivnav: Underwater information-driven vision-based navigation via imitation learning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 5250–5256 (2024). IEEE

[131] Xiao, Z., Li, P., Liu, C., Gao, H., Wang, X.: Macns: A generic graph neural network integrated deep reinforcement learning based multi-agent collaborative navigation system for dynamic trajectory planning. Information Fusion **105**, 102250 (2024)

[132] Xie, W., Jiang, H., Zhu, Y., Qian, J., Xie, J.: Naviformer: A spatio-temporal context-aware transformer for object navigation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 14708–14716 (2025)

[133] Song, X., Chen, W., Liu, Y., Chen, W., Li, G., Lin, L.: Towards long-horizon vision-language navigation: Platform, benchmark and method. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 12078–12088 (2025)

[134] Li, Z., Yu, H., Ding, Y., Li, Y., He, Y., Akhtar, N.: Embodied intelligence for 3d understanding: A survey on 3d scene question answering. arXiv preprint arXiv:2502.00342 (2025)

[135] Zare, M., Kebria, P.M., Khosravi, A., Nahavandi, S.: A survey of imitation learning: Algorithms, recent developments, and challenges. IEEE Transactions on Cybernetics (2024)

[136] Osa, T., Pajarinen, J., Neumann, G., Bagnell, J.A., Abbeel, P., Peters, J., *et al.*: An algorithmic perspective on imitation learning. Foundations and Trends® in Robotics **7**(1-2), 1–179 (2018)

[137] Si, W., Wang, N., Harris, R., Yang, C.: Deep multimodal imitation learning-based framework for robot-assisted medical examination. IEEE Transactions on Industrial Electronics (2025)

[138] Kareer, S., Patel, D., Punamiya, R., Mathur, P., Cheng, S., Wang, C., Hoffman, J., Xu, D.: Egomimic: Scaling imitation learning via egocentric video. In: 2025 IEEE International Conference on Robotics and Automation (ICRA), pp. 13226–13233 (2025). IEEE

[139] Tang, C., Abbatematteo, B., Hu, J., Chandra, R., Martín-Martín, R., Stone, P.: Deep reinforcement learning for robotics: A survey of real-world successes. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 28694–28698 (2025)

[140] Wang, X., Wang, S., Liang, X., Zhao, D., Huang, J., Xu, X., Dai, B., Miao, Q.: Deep reinforcement learning: A survey. IEEE Transactions on Neural Networks and Learning Systems **35**(4), 5064–5078 (2022)

[141] Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. Journal of artificial intelligence research **4**, 237–285 (1996)

[142] Nayyar, R.K., Srivastava, S.: Autonomous option invention for continual hierarchical reinforcement learning and planning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 19642–19650 (2025)

[143] Huang, S., Sun, C., Wang, R.-Q., Pompili, D.: Toward adaptive and coordinated transportation systems: A multi-personality multi-agent meta-reinforcement learning framework. IEEE Transactions on Intelligent Transportation Systems (2025)

[144] Medany, M., Piglia, L., Achenbach, L., Mukkavilli, S.K., Ahmed, D.: Model-based reinforcement learning for ultrasound-driven autonomous microrobots. Nature Machine Intelligence, 1–15 (2025)

[145] Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., et al.: Palm-e: An embodied multimodal language model (2023)

[146] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, pp. 19730–19742 (2023). PMLR

[147] Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., Zeng, A.: Code as policies: Language model programs for embodied control. arXiv preprint arXiv:2209.07753 (2022)

[148] Fung, P., Bachrach, Y., Celikyilmaz, A., Chaudhuri, K., Chen, D., Chung, W., Dupoux, E., Gong, H., Jégou, H., Lazaric, A., et al.: Embodied ai agents: Modeling the world. arXiv preprint arXiv:2506.22355 (2025)

[149] Lei, Y., Fu, Y., Wang, T., Qiu, R.L., Curran, W.J., Liu, T., Yang, X.: Deep

learning in multi-organ segmentation. arXiv preprint arXiv:2001.10619 (2020)

[150] Ahmed, F.A., Yousef, M., Ahmed, M.A., Ali, H.O., Mahboob, A., Ali, H., Shah, Z., Aboumarzouk, O., Al Ansari, A., Balakrishnan, S.: Deep learning for surgical instrument recognition and segmentation in robotic-assisted surgeries: a systematic review. Artificial Intelligence Review **58**(1), 1 (2024)

[151] Du, H., Wang, J., Liu, M., Wang, Y., Meijering, E.: Swinpa-net: Swin transformer-based multiscale feature pyramid aggregation network for medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems **35**(4), 5355–5366 (2022)

[152] Islam, M., Vibashan, V., Lim, C.M., Ren, H.: St-mtl: Spatio-temporal multitask learning model to predict scanpath while tracking instruments in robotic surgery. Medical Image Analysis **67**, 101837 (2021)

[153] Colleoni, E., Moccia, S., Du, X., De Momi, E., Stoyanov, D.: Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. IEEE Robotics and Automation Letters **4**(3), 2714–2721 (2019)

[154] Zeng, Z., Zhuo, Z., Jia, X., Zhang, E., Wu, J., Zhang, J., Wang, Y., Low, C.H., Jiang, J., Zheng, Z., et al.: Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. arXiv preprint arXiv:2506.02555 (2025)

[155] Li, Z., Shaban, A., Simard, J.-G., Rabindran, D., DiMaio, S., Mohareri, O.: A robotic 3d perception system for operating room environment awareness. arXiv preprint arXiv:2003.09487 (2020)

[156] Erol, G., Güngör, A., Sevgi, U.T., Gülsuna, B., Doğruel, Y., Emmez, H., Türe, U.: Creation of a microsurgical neuroanatomy laboratory and virtual operating room: a preliminary study. Neurosurgical Focus **56**(1), 6 (2024)

[157] Gerats, B.G., Wolterink, J.M., Broeders, I.A.: Nerf-or: neural radiance fields for operating room scene reconstruction from sparse-view rgb-d videos. International journal of computer assisted radiology and surgery **20**(1), 147–156 (2025)

[158] Yang, S., Li, Q., Shen, D., Gong, B., Dou, Q., Jin, Y.: Deform3dgs: Flexible deformation for fast surgical scene reconstruction with gaussian splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 132–142 (2024). Springer

[159] Özsoy, E., Örnek, E.P., Eck, U., Czempiel, T., Tombari, F., Navab, N.: 4d-or: Semantic scene graphs for or domain modeling. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 475–485 (2022). Springer

[160] Özsoy, E., Czempiel, T., Holm, F., Pellegrini, C., Navab, N.: Labrad-or: Lightweight memory scene graphs for accurate bimodal reasoning in dynamic operating rooms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 302–311 (2023). Springer

[161] He, P., Zhang, Z., Zhang, Y., Zhao, X., Peng, S.: Spatial-ormllm: Improve spatial relation understanding in the operating room with multimodal large language model. arXiv preprint arXiv:2508.08199 (2025)

[162] Demir, K.C., Schieber, H., Weise, T., Roth, D., May, M., Maier, A., Yang, S.H.: Deep learning in surgical workflow analysis: a review of phase and step recognition. IEEE Journal of Biomedical and Health Informatics **27**(11), 5405–5417 (2023)

[163] Feghoul, K., Maia, D.S., El Amrani, M., Daoudi, M., Amad, A.: Mgrformer: A multimodal transformer approach for surgical gesture recognition. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG), pp. 1–10 (2024). IEEE

[164] Men, Y., Luo, J., Zhao, Z., Wu, H., Luo, F., Zhang, G., Yu, M.: Surgical gesture recognition in open surgery based on 3dcnn and slowfast. In: 2024 IEEE 7th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), vol. 7, pp. 429–433 (2024). IEEE

[165] Ma, L., Kang, H., Magnenat-Thalmann, N., Wac, K.: Transsg: A spatial-temporal transformer for surgical gesture recognition. In: Computer Graphics International Conference, pp. 151–165 (2024). Springer

[166] Jia, B., Wang, W., Tian, X., Wang, X.: Stanet: A surgical gesture recognition method based on spatiotemporal fusion. Annals of the New York Academy of Sciences (2025)

[167] Cristina, S., Despotovic, V., Pérez-Rodríguez, R., Aleksic, S.: Audio-and video-based human activity recognition systems in healthcare. IEEE Access **12**, 8230–8245 (2024)

[168] Van Amsterdam, B., Funke, I., Edwards, E., Speidel, S., Collins, J., Sridhar, A., Kelly, J., Clarkson, M.J., Stoyanov, D.: Gesture recognition in robotic surgery with multimodal attention. IEEE transactions on medical imaging **41**(7), 1677–1687 (2022)

[169] Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE transactions on affective computing **13**(3), 1195–1215 (2020)

[170] Li, Y., Wei, J., Liu, Y., Kauttonen, J., Zhao, G.: Deep learning for micro-expression recognition: A survey. IEEE Transactions on Affective Computing **13**(4), 2028–2046 (2022)

[171] Zhang, L., Qian, Y., Arandjelović, O., Zhu, T., Xiao, H.: Multimodal latent emotion recognition from micro-expression and physiological signal. Pattern Recognition, 111963 (2025)

[172] Zhang, F., Liu, Y., Yu, X., Wang, Z., Zhang, Q., Wang, J., Zhang, Q.: Towards facial micro-expression detection and classification using modified multimodal ensemble learning approach. Information Fusion **115**, 102735 (2025)

[173] Ye, J., Yu, Y., Lu, L., Wang, H., Zheng, Y., Liu, Y., Wang, Q.: Dep-former: Multimodal depression recognition based on facial expressions and audio features via emotional changes. IEEE Transactions on Circuits and Systems for Video Technology (2024)

[174] Khan, M., Gueaieb, W., El Saddik, A., Kwon, S.: Mser: Multimodal speech emotion recognition using cross-attention with deep fusion. Expert Systems with Applications **245**, 122946 (2024)

[175] Gao, H., Cai, Z., Wang, X., Wu, M., Liu, C.: Multimodal fusion of behavioral and physiological signals for enhanced emotion recognition via feature decoupling and knowledge transfer. IEEE journal of biomedical and health informatics (2025)

[176] Kumar, P.S., Govarthan, P.K., Gadda, A.A.S., Ganapathy, N., Ronickom, J.F.A.: Deep learning-based automated emotion recognition using multimodal physiological signals and time-frequency methods. IEEE Transactions on Instrumentation and Measurement **73**, 1–12 (2024)

[177] Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 337–347 (2025). Springer

[178] Zhang, Y., Wang, M., Wu, Y., Tiwari, P., Li, Q., Wang, B., Qin, J.: Dialoguellm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations. arXiv preprint arXiv:2310.11374 (2023)

[179] Garcia-Vidal, C., Sanjuan, G., Puerta-Alcalde, P., Moreno-García, E., Soriano, A.: Artificial intelligence to support clinical decision-making processes. EBioMedicine **46**, 27–29 (2019)

[180] Loftus, T.J., Tighe, P.J., Filiberto, A.C., Efron, P.A., Brakenridge, S.C., Mohr, A.M., Rashidi, P., Upchurch, G.R., Bihorac, A.: Artificial intelligence and surgical decision-making. JAMA surgery **155**(2), 148–158 (2020)

[181] Mangano, F.G., Admakin, O., Lerner, H., Mangano, C.: Artificial intelligence and augmented reality for guided implant surgery planning: a proof of concept.

Journal of Dentistry **133**, 104485 (2023)

[182] Aspland, E., Gartner, D., Harper, P.: Clinical pathway modelling: a literature review. Health Systems **10**(1), 1–23 (2021)

[183] Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.-A.: Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 593–603 (2021). Springer

[184] Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 343–352 (2020). Springer

[185] Kadkhodamohammadi, A., Luengo, I., Stoyanov, D.: Patg: position-aware temporal graph networks for surgical phase recognition on laparoscopic videos. International Journal of Computer Assisted Radiology and Surgery **17**(5), 849–856 (2022)

[186] Zhang, F.X., Al Moubayed, N., Shum, H.P.: Towards graph representation learning based surgical workflow anticipation. In: 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 01–04 (2022). IEEE

[187] Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)

[188] Mezger, U., Jendrewski, C., Bartels, M.: Navigation in surgery. Langenbeck's archives of surgery **398**(4), 501–514 (2013)

[189] Hasan, A.M., Cavalu, S., Kira, A.Y., Hamad, R.S., Abdel-Reheim, M.A., Elmorsy, E.A., El-Kott, A.F., Morsy, K., AlSheri, A.S., Negm, S., et al.: Localized drug delivery in different gastrointestinal cancers: navigating challenges and advancing nanotechnological solutions. International Journal of Nanomedicine, 741–770 (2025)

[190] Øllgaard, J.A., Bundgaard, K., Sorag Kjaer, M., Bodenhagen, L., Palinko, O.: Utilizing a social robot as a greeter at a children's hospital. In: International Conference on Social Robotics, pp. 131–144 (2024). Springer

[191] Hughes, N., Pinchin, J., Brown, M., Shaw, D.: Navigating in large hospitals. In: 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 1–9 (2015). IEEE

[192] Gumprecht, H.K., Widenka, D.C., Lumenta, C.B.: Brainlab vectorvision neuronavigation system: technology and clinical experiences in 131 cases. Neurosurgery **44**(1), 97–104 (1999)

[193] Liu, P., Li, C., Xiao, C., Zhang, Z., Ma, J., Gao, J., Shao, P., Valerio, I., Pawlik, T.M., Ding, C., *et al.*: A wearable augmented reality navigation system for surgical telementoring based on microsoft hololens. Annals of biomedical engineering **49**(1), 287–298 (2021)

[194] Li, A., Han, J., Zhao, Y., Meng, M.Q.-H., Liu, L.: Rl-usregi: Autonomous ultrasound registration for radiation-free spinal surgical navigation using reinforcement learning. IEEE Transactions on Automation Science and Engineering (2025)

[195] Robertshaw, H., Karstensen, L., Jackson, B., Granados, A., Booth, T.C.: Autonomous navigation of catheters and guidewires in mechanical thrombectomy using inverse reinforcement learning. International Journal of Computer Assisted Radiology and Surgery **19**(8), 1569–1578 (2024)

[196] Zhou, G., Hong, Y., Wu, Q.: Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 7641–7649 (2024)

[197] Zhou, G., Hong, Y., Wang, Z., Wang, X.E., Wu, Q.: Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In: European Conference on Computer Vision, pp. 260–278 (2024). Springer

[198] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., *et al.*: Toward expert-level medical question answering with large language models. Nature Medicine **31**(3), 943–950 (2025)

[199] Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z.: Medical visual question answering: A survey. Artificial Intelligence in Medicine **143**, 102611 (2023)

[200] Jin, Q., Yuan, Z., Xiong, G., Yu, Q., Ying, H., Tan, C., Chen, M., Huang, S., Liu, X., Yu, S.: Biomedical question answering: a survey of approaches and challenges. ACM Computing Surveys (CSUR) **55**(2), 1–36 (2022)

[201] Rosen, A.W., Ose, I., Gögenur, M., Andersen, L.P.K., Bojesen, R.D., Vogelsang, R.P., Rose, M.H., Steenfos, P.W., Hansen, L.B., Spuur, H.S., et al.: Clinical implementation of an ai-based prediction model for decision support for patients undergoing colorectal cancer surgery. Nature Medicine, 1–12 (2025)

[202] Şişman, A.Ç., Acar, A.H.: Artificial intelligence-based chatbot assistance in clinical decision-making for medically complex patients in oral surgery: a comparative

study. BMC Oral Health **25**(1), 351 (2025)

[203] Patil, A., Patil, V., Sankpal, S., Patankar, T.S., Bhute, H.: Multimodal decision support system for improved diagnosis and healthcare decision making. Journal of Biology and Health Science (2025)

[204] Gong, L., Yang, J., Han, S., Ji, Y.: Medblip: A multimodal method of medical question-answering based on fine-tuning large language model. Computerized Medical Imaging and Graphics, 102581 (2025)

[205] Schmidgall, S., Opfermann, J.D., Kim, J.W., Krieger, A.: Will your next surgeon be a robot? autonomy and ai in robotic surgery. Science Robotics **10**(104), 0187 (2025)

[206] Attanasio, A., Scaglioni, B., De Momi, E., Fiorini, P., Valdastri, P.: Autonomy in surgical robotics. Annual Review of Control, Robotics, and Autonomous Systems **4**(1), 651–679 (2021)

[207] Peloso, A., Damiano, R., Zhang, X., Bicchi, A., Votta, E., De Momi, E.: Imitation learning for path planning in cardiac percutaneous interventions. IEEE Transactions on Biomedical Engineering (2025)

[208] Kim, J.W., Zhao, T.Z., Schmidgall, S., Deguet, A., Kobilarov, M., Finn, C., Krieger, A.: Surgical robot transformer (srt): Imitation learning for surgical tasks. arXiv preprint arXiv:2407.12998 (2024)

[209] Paradis, S., Hwang, M., Thananjeyan, B., Ichnowski, J., Seita, D., Fer, D., Low, T., Gonzalez, J.E., Goldberg, K.: Intermittent visual servoing: Efficiently learning policies robust to instrument changes for high-precision surgical manipulation. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 7166–7173 (2021). IEEE

[210] Moghani, M., Nelson, N., Ghanem, M., Diaz-Pinto, A., Hari, K., Azizian, M., Goldberg, K., Huver, S., Garg, A.: Sufia-bc: Generating high quality demonstration data for visuomotor policy learning in surgical subtasks. arXiv preprint arXiv:2504.14857 (2025)

[211] Segato, A., Di Marzo, M., Zucchelli, S., Galvan, S., Secoli, R., De Momi, E.: Inverse reinforcement learning intra-operative path planning for steerable needle. IEEE Transactions on Biomedical Engineering **69**(6), 1995–2005 (2021)

[212] Chi, W., Dagnino, G., Kwok, T.M., Nguyen, A., Kundrat, D., Abdelaziz, M.E., Riga, C., Bicknell, C., Yang, G.-Z.: Collaborative robot-assisted endovascular catheterization with generative adversarial imitation learning. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 2414–2420 (2020). IEEE

[213] Gomaa, A., Mahdy, B., Kleer, N., Krüger, A.: Towards a surgeon-in-the-loop ophthalmic robotic apprentice using reinforcement and imitation learning. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 6939–6946 (2024). IEEE

[214] Li, B., Wei, R., Xu, J., Lu, B., Yee, C.H., Ng, C.F., Heng, P.-A., Dou, Q., Liu, Y.-H.: 3d perception based imitation learning under limited demonstration for laparoscope control in robotic surgery. In: 2022 International Conference on Robotics and Automation (ICRA), pp. 7664–7670 (2022). IEEE

[215] Yu, C., Liu, J., Nemati, S., Yin, G.: Reinforcement learning in healthcare: A survey. ACM Computing Surveys (CSUR) **55**(1), 1–36 (2021)

[216] Varier, V.M., Rajamani, D.K., Goldfarb, N., Tavakkolmoghaddam, F., Munawar, A., Fischer, G.S.: Collaborative suturing: A reinforcement learning approach to automate hand-off task in suturing for surgical robots. In: 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1380–1386 (2020). IEEE

[217] Scheikl, P.M., Gyenes, B., Younis, R., Haas, C., Neumann, G., Wagner, M., Mathis-Ullrich, F.: Lapgym-an open source framework for reinforcement learning in robot-assisted laparoscopic surgery. Journal of Machine Learning Research **24**(368), 1–42 (2023)

[218] Meng, F., Guo, S., Zhou, W., Chen, Z.: Evaluation of an autonomous navigation method for vascular interventional surgery in virtual environment. In: 2022 IEEE International Conference on Mechatronics and Automation (ICMA), pp. 1599–1604 (2022). IEEE

[219] Liu, J., Andres, A., Jiang, Y., Luo, X., Shu, W., Tsaftaris, S.A.: Surgical task automation using actor-critic frameworks and self-supervised imitation learning. arXiv preprint arXiv:2409.02724 (2024)

[220] Li, H., Zhou, X.-H., Xie, X.-L., Liu, S.-Q., Feng, Z.-Q., Hou, Z.-G.: Casog: Conservative actor–critic with smooth gradient for skill learning in robot-assisted intervention. IEEE Transactions on Industrial Electronics **71**(7), 7722–7731 (2023)

[221] Min, Z., Lai, J., Ren, H.: Innovating robot-assisted surgery through large vision models. Nature Reviews Electrical Engineering, 1–14 (2025)

[222] Schmidgall, S., Cho, J., Zakka, C., Hiesinger, W.: Gp-vls: A general-purpose vision language model for surgery. arXiv preprint arXiv:2407.19305 (2024)

[223] Moglia, A., Georgiou, K., Cerveri, P., Mainardi, L., Satava, R.M., Cuschieri, A.: Large language models in healthcare: from a systematic review on medical examinations to a comparative analysis on fundamentals of robotic surgery

online test. Artificial Intelligence Review **57**(9), 231 (2024)

[224] Seenivasan, L., Islam, M., Kannan, G., Ren, H.: Surgicalgpt: end-to-end language-vision gpt for visual question answering in surgery. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 281–290 (2023). Springer

[225] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H), pp. 353–367 (2023). PMLR

[226] Li, S., Wang, J., Dai, R., Ma, W., Ng, W.Y., Hu, Y., Li, Z.: Robonurse-vla: Robotic scrub nurse system based on vision-language-action model. arXiv preprint arXiv:2409.19590 (2024)

[227] D'Ettorre, C., Mariani, A., Stilli, A., Baena, F.R., Valdastri, P., Deguet, A., Kazanzides, P., Taylor, R.H., Fischer, G.S., DiMaio, S.P., *et al.*: Accelerating surgical robotics research: A review of 10 years with the da vinci research kit. IEEE Robotics & Automation Magazine **28**(4), 56–78 (2021)

[228] Freschi, C., Ferrari, V., Melfi, F., Ferrari, M., Mosca, F., Cuschieri, A.: Technical review of the da vinci surgical telemanipulator. The International Journal of Medical Robotics and Computer Assisted Surgery **9**(4), 396–406 (2013)

[229] Surgical, I.: da vinci. surgical system. http://www. intusurg. com/html/davinci. html (2013)

[230] Tadano, K., Kawashima, K.: A pneumatic laparoscope holder controlled by head movement. The International Journal of Medical Robotics and Computer Assisted Surgery **11**(3), 331–340 (2015)

[231] Chenin, L., Peltier, J., Lefranc, M.: Minimally invasive transforaminal lumbar interbody fusion with the rosatm spine robot and intraoperative flat-panel ct guidance. Acta neurochirurgica **158**(6), 1125–1128 (2016)

[232] Edwards, T., Xue, K., Meenink, H., Beelen, M., Naus, G., Simunovic, M., Latasiewicz, M., Farmery, A., De Smet, M., MacLaren, R.: First-in-human study of the safety and viability of intraocular robotic surgery. Nature biomedical engineering **2**(9), 649–656 (2018)

[233] Bargar, W.L., Bauer, A., Börner, M.: Primary and revision total hip replacement using the robodoc (r) system. Clinical Orthopaedics and Related Research (1976-2007) **354**, 82–91 (1998)

[234] Dastagir, N., Obed, D., Tamulevicius, M., Dastagir, K., Vogt, P.M.: The use of the symani surgical system® in emergency hand trauma care. Surgical innovation **31**(5), 460–465 (2024)

[235] Kilby, W., Dooley, J., Kuduvalli, G., Sayeh, S., Maurer Jr, C.: The cyberknife® robotic radiosurgery system in 2010. Technology in cancer research & treatment **9**(5), 433–452 (2010)

[236] Graetzel, C.F., Sheehy, A., Noonan, D.P.: Robotic bronchoscopy drive mode of the auris monarch platform. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 3895–3901 (2019). IEEE

[237] Remacle, M., MN Prasad, V., Lawson, G., Plisson, L., Bachy, V., Vorst, S.: Transoral robotic surgery (tors) with the medrobotics flex™ system: first surgical application on humans. European Archives of Oto-Rhino-Laryngology **272**(6), 1451–1455 (2015)

[238] Britz, G.W., Panesar, S.S., Falb, P., Tomas, J., Desai, V., Lumsden, A.: Neuroendovascular-specific engineering modifications to the corpath grx robotic system. Journal of neurosurgery **133**(6), 1830–1836 (2019)

[239] Hung, L., Liu, C., Woldum, E., Au-Yeung, A., Berndt, A., Wallsworth, C., Horne, N., Gregorio, M., Mann, J., Chaudhury, H.: The benefits of and barriers to using a social robot paro in care settings: a scoping review. BMC geriatrics **19**(1), 232 (2019)

[240] Tanaka, K., Makino, H., Nakamura, K., Nakamura, A., Hayakawa, M., Uchida, H., Kasahara, M., Kato, H., Igarashi, T.: The pilot study of group robot intervention on pediatric inpatients and their caregivers, using 'new aibo'. European Journal of Pediatrics **181**(3), 1055–1061 (2022)

[241] Kushniruk, A.W., Nair, H.S., Borycki, E.M.: The pepper robot in healthcare: A scoping review. Envisioning the Future of Health Informatics and Digital Health, 101–105 (2025)

[242] Broadbent, E., Loveys, K., Ilan, G., Chen, G., Chilukuri, M., Boardman, S., Doraiswamy, P., Skuler, D.: Elliq, an ai-driven social robot to alleviate loneliness: progress and lessons learned. The Journal of Aging Research & Lifestyle **13**, 22–28 (2024)

[243] Meghdari, A., Shariati, A., Alemi, M., Vossoughi, G.R., Eydi, A., Ahmadi, E., Mozafari, B., Amoozandeh Nobaveh, A., Tahami, R.: Arash: A social robot buddy to support children with cancer in a hospital environment. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine **232**(6), 605–618 (2018)

[244] Khan, Z.H., Siddique, A., Lee, C.W.: Robotics utilization for healthcare digitization in global covid-19 management. International journal of environmental research and public health **17**(11), 3819 (2020)

[245] González-Jiménez, J., Galindo, C., Ruiz-Sarmiento, J.: Technical improvements

of the giraff telepresence robot based on users' evaluation. In: 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, pp. 827–832 (2012). IEEE

[246] Ogawa, K., Nishio, S., Koda, K., Taura, K., Minato, T., Ishii, C.T., Ishiguro, H.: Telenoid: Tele-presence android for communication. In: ACM SIGGRAPH 2011 Emerging Technologies, pp. 1–1 (2011)

[247] Clark, R.E., Feldon, D.F., Van Merrienboer, J.J., Yates, K.A., Early, S.: Cognitive task analysis. In: Handbook of Research on Educational Communications and Technology, pp. 577–593. Routledge, ??? (2008)

[248] Kolesnyk, M.Y.: The first experience of using the body interact simulation interactive training platform as a part of interns' attestation (2020)

[249] Gallagher, K., Bahadori, S., Antonis, J., Immins, T., Wainwright, T.W., Middleton, R.: Validation of the hip arthroscopy module of the virtamed virtual reality arthroscopy trainer. Surgical technology international **34**, 430–436 (2019)

[250] Vasilev, V., Kondrichina, S.N.: Possibilities for using the vimedix 3.2 virtual simulator to train ultrasound specialists. Digital Diagnostics **5**(1), 41–52 (2024)

[251] Keller, M., Zuffi, S., Black, M.J., Pujades, S.: Osso: Obtaining skeletal shape from outside. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 20492–20501 (2022)

[252] Lilly, J.: 3d organon vr anatomy. Journal of the Medical Library Association: JMLA **110**(2), 276 (2022)

[253] Teladoc health. accessed: 14/11/2025: https://www.teladochealth.com/.

[254] Klingensmith, L., Knodel, L.: Mercy virtual nursing: An innovative care delivery model. Nurse Leader **14**(4), 275–279 (2016)

[255] Guo, D., Liu, W., Zhang, X., Zhao, M., Zhu, B., Hou, T., He, H.: Duck egg white–derived peptide vsee (val-ser-glu-glu) regulates bone and lipid metabolisms by wnt/β-catenin signaling pathway and intestinal microbiota. Molecular nutrition & food research **63**(24), 1900525 (2019)

[256] Kirby, E.D., Beyst, B., Beyst, J., Brodie, S.M., D'Arcy, R.C.: A retrospective, observational study of real-world clinical data from the cognitive function development therapy program. Frontiers in Human Neuroscience **18**, 1508815 (2024)

[257] Hernandez Petzsche, M.R., De La Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.-L., Kofler, F., Ezhov, I., *et al.*: Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset.

Scientific data **9**(1), 762 (2022)

[258] Li, X., Luo, G., Wang, K., Wang, H., Liu, J., Liang, X., Jiang, J., Song, Z., Zheng, C., Chi, H., et al.: The state-of-the-art 3d anisotropic intracranial hemorrhage segmentation on non-contrast head ct: The instance challenge. arXiv preprint arXiv:2301.03281 (2023)

[259] Tian, Y., Shi, M., Luo, Y., Kouhana, A., Elze, T., Wang, M.: Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. arXiv preprint arXiv:2311.02189 (2023)

[260] Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences **501**, 511–522 (2019)

[261] Li, L., Zimmer, V.A., Schnabel, J.A., Zhuang, X.: Atrialjsqnet: a new framework for joint segmentation and quantification of left atrium and scars incorporating spatial and shape information. Medical image analysis **76**, 102303 (2022)

[262] Zhanga, M., Wua, Y., Zhanga, H., Qina, Y., Zhenga, H., Tangc, W., Arnoldq, C., Peic, C., Yuc, P., Nand, Y., et al.: Multi-site, multi-domain airway tree modeling (atm'22): A public benchmark for pulmonary airway segmentation (2023)

[263] Lambert, Z., Petitjean, C., Dubray, B., Kuan, S.: Segthor: Segmentation of thoracic organs at risk in ct images. In: 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA), pp. 1–6 (2020). Ieee

[264] Luo, G., Xu, M., Chen, H., Liang, X., Tao, X., Ni, D., Jeong, H., Kim, C., Stock, R., Baumgartner, M., et al.: Tumor detection, segmentation and classification challenge on automated 3d breast ultrasound: The tdsc-abus challenge. arXiv preprint arXiv:2501.15588 (2025)

[265] Soler, L., Hostettler, A., Agnus, V., Charnoz, A., Fasquel, J.-B., Moreau, J., Osswald, A.-B., Bouhadjar, M., Marescaux, J.: 3d image reconstruction for comparison of algorithm database. URL: https://www. ircad. fr/research/data-sets/liver-segmentation-3d-ircadb-01 **13** (2010)

[266] Sekuboyina, A., Husseini, M.E., Bayat, A., Löffler, M., Liebl, H., Li, H., Tetteh, G., Kukačka, J., Payer, C., Štern, D., *et al.*: Verse: a vertebrae labelling and segmentation benchmark for multi-detector ct images. Medical image analysis **73**, 102166 (2021)

[267] Graaf, J.W., Hooff, M.L., Buckens, C.F., Rutten, M., Susante, J.L., Kroeze, R.J., Kleuver, M., Ginneken, B., Lessmann, N.: Lumbar spine segmentation in mr images: a dataset and a public benchmark. Scientific Data **11**(1), 264 (2024)

[268] Deng, Y., Wang, C., Hui, Y., Li, Q., Li, J., Luo, S., Sun, M., Quan, Q., Yang, S., Hao, Y., et al.: Ctspine1k: A large-scale dataset for spinal vertebrae segmentation in computed tomography. arXiv preprint arXiv:2105.14711 (2021)

[269] Liu, P., Han, H., Du, Y., Zhu, H., Li, Y., Gu, F., Xiao, H., Li, J., Zhao, C., Xiao, L., *et al.*: Deep learning to segment pelvic bones: large-scale ct datasets and baseline models. International Journal of Computer Assisted Radiology and Surgery **16**(5), 749–756 (2021)

[270] Gutman, D., Codella, N.C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., Halpern, A.: Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1605.01397 (2016)

[271] Giotis, I., Molders, N., Land, S., Biehl, M., Jonkman, M.F., Petkov, N.: Mednode: A computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert systems with applications **42**(19), 6578–6585 (2015)

[272] Mendonça, T., Celebi, M., Mendonca, T., Marques, J.: Ph2: A public database for the analysis of dermoscopic images. Dermoscopy image analysis **2** (2015)

[273] Pacheco, A.G., Lima, G.R., Salomao, A.S., Krohling, B., Biral, I.P., De Angelo, G.G., Alves Jr, F.C., Esgario, J.G., Simora, A.C., Castro, P.B., *et al.*: Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data in brief **32**, 106221 (2020)

[274] Islam, T., Hussain, M.A., Chowdhury, F.U.H., Riazul Islam, B.: A web-scraped skin image database of monkeypox, chickenpox, smallpox, cowpox, and measles. biorxiv, 2022–08 (2022)

[275] Wasserthal, J., Breit, H.-C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., *et al.*: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), 230024 (2023)

[276] Grauw, M., Scholten, E.T., Smit, E.J., Rutten, M.J., Prokop, M., Ginneken, B., Hering, A.: The uls23 challenge: A baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography. Medical Image Analysis **102**, 103525 (2025)

[277] Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. Scientific Data **9**(1), 601 (2022)

[278] Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.-H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument

segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)

[279] Lin, S., Qin, F., Li, Y., Bly, R.A., Moe, K.S., Hannaford, B.: Lc-gan: Image-to-image translation based on generative adversarial network for endoscopic images. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 2914–2920 (2020). IEEE

[280] Ding, H., Zhang, Y., Lu, T., Liang, R., Shu, H., Seenivasan, L., Long, Y., Dou, Q., Gao, C., Leng, Y., et al.: Segstrong-c: Segmenting surgical tools robustly on non-adversarial generated corruptions–an endovis' 24 challenge. arXiv preprint arXiv:2407.11906 (2024)

[281] Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)

[282] Nwoye, C.I., Elgohary, K., Srinivas, A., Zaid, F., Lavanchy, J.L., Padoy, N.: Cholectrack20: A multi-perspective tracking dataset for surgical tools. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 8942–8952 (2025)

[283] Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., et al.: The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. arXiv preprint arXiv:2312.12429 (2023)

[284] Maqbool, S., Riaz, A., Sajid, H., Hasan, O.: m2caiseg: Semantic segmentation of laparoscopic images using convolutional neural networks. arXiv preprint arXiv:2008.10134 (2020)

[285] Bano, S., Casella, A., Vasconcelos, F., Moccia, S., Attilakos, G., Wimalasundera, R., David, A.L., Paladini, D., Deprest, J., De Momi, E., et al.: Fetreg: Placental vessel segmentation and registration in fetoscopy challenge dataset. arXiv preprint arXiv:2106.05923 (2021)

[286] Özsoy, E., Pellegrini, C., Czempiel, T., Tristram, F., Yuan, K., Bani-Harouni, D., Eck, U., Busam, B., Keicher, M., Navab, N.: Mm-or: A large multimodal operating room dataset for semantic understanding of high-intensity surgical environments. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 19378–19389 (2025)

[287] Rodas, N.L., Barrera, F., Padoy, N.: See it with your own eyes: Markerless mobile augmented reality for radiation awareness in the hybrid room. IEEE Transactions on Biomedical Engineering **64**(2), 429–440 (2016)

[288] Hu, D., Li, S., Wang, M.: Object detection in hospital facilities: A comprehensive dataset and performance evaluation. Engineering Applications of Artificial Intelligence **123**, 106223 (2023)

[289] Bashiri, F.S., LaRose, E., Peissig, P., Tafti, A.P.: Mcindoor20000: A fully-labeled image dataset to advance indoor objects detection. Data in brief **17**, 71–75 (2018)

[290] Ismail, A., Ahmad, S.A., Soh, A.C., Hassan, M.K., Harith, H.H.: Mynursinghome: A fully-labelled image dataset for indoor object classification. Data in Brief **32**, 106268 (2020)

[291] Srivastav, V., Issenhuth, T., Kadkhodamohammadi, A., Mathelin, M., Gangi, A., Padoy, N.: Mvor: A multi-view rgb-d operating room dataset for 2d and 3d human pose estimation. arXiv preprint arXiv:1808.08180 (2018)

[292] Chen, K., Gabriel, P., Alasfour, A., Gong, C., Doyle, W.K., Devinsky, O., Friedman, D., Dugan, P., Melloni, L., Thesen, T., *et al.*: Patient-specific pose estimation in clinical environments. IEEE journal of translational engineering in health and medicine **6**, 1–11 (2018)

[293] Markova, V., Ganchev, T., Filkova, S., Markov, M.: Mmd-msd: A multimodal multisensory dataset in support of research and technology development for musculoskeletal disorders. Algorithms **17**(5), 187 (2024)

[294] Wu, J., Chen, Z., Xu[1], M.: Surgtrack: Cad-free 3d tracking of real-world surgical instruments. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2024 Workshops: ISIC 2024, iMIMIC 2024, EARTH 2024, DeCaF 2024, Held in Conjunction with MICCAI 2024, Marrakesh, Morocco, October 6–10, 2024, Proceedings, vol. 15274, p. 168 (2025). Springer Nature

[295] Sekiavandi, M.J., Dixen, L., Fimland, J., Desu, S.K., Zserai, A.-B., Lee, Y.S., Barrett, M., Burelli, P.: Advancing face-to-face emotion communication: A multimodal dataset (affec). arXiv preprint arXiv:2504.18969 (2025)

[296] Cheng, Z., Cheng, Z.-Q., He, J.-Y., Wang, K., Lin, Y., Lian, Z., Peng, X., Hauptmann, A.: Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. Advances in Neural Information Processing Systems **37**, 110805–110853 (2024)

[297] Yang, P., Liu, N., Liu, X., Shu, Y., Ji, W., Ren, Z., Sheng, J., Yu, M., Yi, R., Zhang, D., *et al.*: A multimodal dataset for mixed emotion recognition. Scientific Data **11**(1), 847 (2024)

[298] Jiang, W.-B., Liu, X.-H., Zheng, W.-L., Lu, B.-L.: Seed-vii: A multimodal dataset of six basic emotions with continuous labels for emotion recognition. IEEE Transactions on Affective Computing (2024)

[299] Subramanian, R., Wache, J., Abadi, M.K., Vieriu, R.L., Winkler, S., Sebe, N.: Ascertain: Emotion and personality recognition using commercial sensors. IEEE Transactions on Affective Computing **9**(2), 147–160 (2016)

[300] Katsigiannis, S., Ramzan, N.: Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. IEEE journal of biomedical and health informatics **22**(1), 98–107 (2017)

[301] Hu, M., Xia, P., Wang, L., Yan, S., Tang, F., Xu, Z., Luo, Y., Song, K., Leitner, J., Cheng, X., *et al.*: Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding. In: European Conference on Computer Vision, pp. 481–500 (2024). Springer

[302] Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.-H., Dou, Q., Liu, Y.: Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 486–496 (2022). Springer

[303] Derathé, A., Reche, F., Guy, S., Charrière, K., Trilling, B., Jannin, P., Moreau-Gaudry, A., Gibaud, B., Voros, S.: Lapex: A new multimodal dataset for context recognition and practice assessment in laparoscopic surgery. Scientific Data **12**(1), 342 (2025)

[304] Huaulmé, A., Sarikaya, D., Le Mut, K., Despinoy, F., Long, Y., Dou, Q., Chng, C.-B., Lin, W., Kondo, S., Bravo-Sánchez, L., *et al.*: Micro-surgical anastomose workflow recognition challenge report. Computer Methods and Programs in Biomedicine **212**, 106452 (2021)

[305] Wu, Z., Tong, D., Xie, H., Sun, L., Fan, X., Yang, Z.: A portable 6d surgical instrument magnetic localization system with dynamic error correction. IEEE Sensors Journal (2025)

[306] Qi, Z., Jin, H., Xu, X., Wang, Q., Gan, Z., Xiong, R., Zhang, S., Liu, M., Wang, J., Ding, X., *et al.*: Head model dataset for mixed reality navigation in neurosurgical interventions for intracranial lesions. Scientific Data **11**(1), 538 (2024)

[307] Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9068–9079 (2018)

[308] Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., et al.: Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. arXiv preprint arXiv:2109.08238 (2021)

[309] Yuan, K., Kattel, M., Lavanchy, J.L., Navab, N., Srivastav, V., Padoy, N.: Advancing surgical vqa with scene graph knowledge. International journal of computer assisted radiology and surgery **19**(7), 1409–1417 (2024)

[310] Ray, S., Gupta, K., Kundu, S., Kasat, P.A., Aditya, S., Goyal, P.: Ervqa: A dataset to benchmark the readiness of large vision language models in hospital environments. arXiv preprint arXiv:2410.06420 (2024)

[311] Wu, J., Deng, W., Li, X., Liu, S., Mi, T., Peng, Y., Xu, Z., Liu, Y., Cho, H., Choi, C.-I., et al.: Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. arXiv preprint arXiv:2504.00993 (2025)

[312] Sun, Y., Qian, X., Xu, W., Zhang, H., Xiao, C., Li, L., Zhao, D., Huang, W., Xu, T., Bai, Q., *et al.*: Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 26457–26478 (2025)

[313] Özsoy, E., Pellegrini, C., Bani-Harouni, D., Yuan, K., Keicher, M., Navab, N.: Orqa: A benchmark and foundation model for holistic operating room modeling. arXiv preprint arXiv:2505.12890 (2025)

[314] Li, J., Skinner, G., Yang, G., Quaranto, B.R., Schwaitzberg, S.D., Kim, P.C., Xiong, J.: Llava-surg: towards multimodal surgical assistant via structured surgical video learning. arXiv preprint arXiv:2408.07981 (2024)

[315] Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjar, B., Yuh, D.D., *et al.*: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: MICCAI Workshop: M2cai, vol. 3, p. 3 (2014)

[316] Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., Mutter, D., Marescaux, J., Dallemagne, B., et al.: The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. arXiv preprint arXiv:2312.12429 (2023)

[317] Lavanchy, J.L., Ramesh, S., Dall'Alba, D., Gonzalez, C., Fiorini, P., Müller-Stich, B.P., Nett, P.C., Marescaux, J., Mutter, D., Padoy, N.: Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. International journal of computer assisted radiology and surgery **19**(11), 2249–2257 (2024)

[318] Zia, A., Berniker, M., Nespolo, R., Perreault, C., Wang, Z., Mueller, B., Schmidt, R., Bhattacharyya, K., Liu, X., Jarc, A.: Surgical visual understanding (surgvu) dataset. arXiv preprint arXiv:2501.09209 (2025)

[319] Stauder, R., Ostler, D., Kranzfelder, M., Koller, S., Feußner, H., Navab, N.:

The tum lapchole dataset for the m2cai 2016 workflow challenge. arXiv preprint arXiv:1610.09278 (2016)

[320] Schmidgall, S., Kim, J.W., Jopling, J., Krieger, A.: General surgery vision transformer: A video pre-trained foundation model for general surgery. arXiv preprint arXiv:2403.05949 (2024)

[321] Hartwig, R., Ostler, D., Rosenthal, J.-C., Feußner, H., Wilhelm, D., Wollherr, D.: Miti: Slam benchmark for laparoscopic surgery. arXiv preprint arXiv:2202.11496 (2022)

[322] Wang, G., Xiao, H., Zhang, R., Gao, H., Bai, L., Yang, X., Li, Z., Li, H., Ren, H.: Copesd: A multi-level surgical motion dataset for training large vision-language models to co-pilot endoscopic submucosal dissection. In: Proceedings of the 33rd ACM International Conference on Multimedia, pp. 12636–12643 (2025)

[323] Xu, J., Li, B., Lu, B., Liu, Y.-H., Dou, Q., Heng, P.-A.: Surrol: An open-source reinforcement learning centered and dvrk compatible platform for surgical robot learning. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1821–1828 (2021). IEEE

[324] Yu, Q., Moghani, M., Dharmarajan, K., Schorp, V., Panitch, W.C.-H., Liu, J., Hari, K., Huang, H., Mittal, M., Goldberg, K., *et al.*: Orbit-surgical: An open-simulation framework for learning surgical augmented dexterity. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 15509–15516 (2024). IEEE

[325] Schmidgall, S., Krieger, A., Eshraghian, J.: Surgical gym: A high-performance gpu-based platform for reinforcement learning with surgical robots. In: 2024 IEEE International Conference on Robotics and Automation (ICRA), pp. 13354–13361 (2024). IEEE

[326] Scheikl, P.M., Gyenes, B., Younis, R., Haas, C., Neumann, G., Wagner, M., Mathis-Ullrich, F.: Lapgym-an open source framework for reinforcement learning in robot-assisted laparoscopic surgery. Journal of Machine Learning Research **24**(368), 1–42 (2023)

[327] Ao, Y., Moghani, M., Mittal, M., Prajapat, M., Wu, L., Giraud, F., Carrillo, F., Krause, A., Fürnstahl, P.: Sonogym: High performance simulation for challenging surgical tasks with robotic ultrasound. arXiv preprint arXiv:2507.01152 (2025)

[328] Shang, F., Fu, J., Yang, Y., Huang, H., Liu, J., Ma, L.: Synfundus-1m: a high-quality million-scale synthetic fundus images dataset with fifteen types of annotation. arXiv preprint arXiv:2312.00377 (2023)

[329] Ding, K., Zhou, M., Wang, H., Gevaert, O., Metaxas, D., Zhang, S.: A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer. Scientific Data **10**(1), 231 (2023)

[330] Walonoski, J., Klaus, S., Granger, E., Hall, D., Gregorowicz, A., Neyarapally, G., Watson, A., Eastman, J.: Synthea™ novel coronavirus (covid-19) model and synthetic data set. Intelligence-based medicine **1**, 100007 (2020)

[331] Walonoski, J., Hall, D., Bates, K.M., Farris, M.H., Dagher, J., Downs, M.E., Sivek, R.T., Wellner, B., Gregorowicz, A., Hadley, M., *et al.*: The "coherent data set": combining patient data and imaging in a comprehensive, synthetic health record. Electronics **11**(8), 1199 (2022)

[332] Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)

[333] Wang, D., Zhang, Y., Zhang, K., Wang, L.: Focalmix: Semi-supervised learning for 3d medical image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3951–3960 (2020)

[334] Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A., Abidin, A.Z., Liu, A., Costa, A.B., Wood, B.J., Tsai, C.-S., *et al.*: Federated learning for predicting clinical outcomes in patients with covid-19. Nature medicine **27**(10), 1735–1743 (2021)

[335] Sun, Z., Yin, H., Chen, H., Chen, T., Cui, L., Yang, F.: Disease prediction via graph neural networks. IEEE Journal of Biomedical and Health Informatics **25**(3), 818–826 (2020)

[336] Wu, H., Shi, W., Choudhary, A., Wang, M.D.: Clinical decision making under uncertainty: a bootstrapped counterfactual inference approach. BMC Medical Informatics and Decision Making **24**(1), 275 (2024)

[337] Zhang, Y., Chen, Z., Mu, D., Chen, J., Hua, C.: Adaptive visual servoing of robotic systems with closed structure and limited image-space position measurement capacity. IEEE Transactions on Instrumentation and Measurement (2025)