ITM Web of Conferences **80**, 01037 (2025) https://doi.org/10.1051/itmconf/20258001037
ACAAI 2025

Hierarchical Frameworks for Embodied Medical AI

Zhibo Zhong*

School of Artificial Intelligence, Shanghai Jiaotong University, Shanghai, 200240, China

Abstract. Conventional disembodied models took up most of the medical artificial intelligence (AI) while serving a very limited use. Until recently, embodying artificial intelligence in physical robotics has expanded its capabilities and started to show promising clinical translation. This primarily includes automatic surgery platforms, diagnostic assistants, and rehabilitation exoskeletons that employ robust hierarchical control to safely deliver an integration of perception, decision, and action. This study synthesises methods for embodied medical artificial intelligence and conducts a systematic and quantitative analysis of 2025 publications across the aforementioned four domains. The hierarchical designs prove effective as data indicate an improved accuracy, which includes a ~10% higher macro-F1 or sub-millimetre surgical precision in specific tasks. This marks a milestone toward a more practical level of reliability required by clinical translation. These findings confirm that the layered decomposition and hierarchical control are one of the keys to further development of clinical-ready medical artificial intelligence.

1 Introduction

Conventionally, the disembodied Artificial Intelligence (AI) systems serve a very limited use in the clinical situation. Therefore, the industry is moving towards more adaptive embodied systems that dynamically self-adjust, as shown in a new wave of research. This marks a transforming shift in healthcare technology.

Among them, the hierarchical frameworks address the natural complexity and low tolerance of medical applications by decomposing challenging objectives into manageable abstraction modules, forming a viable procedure [1]. Recent research in hierarchical frameworks integrates perception, reasoning, planning, and action to create a system capable of autonomous surgery, diagnostic intervention, and personalised rehabilitation to support clinical decisions and procedures based on established groundwork for AI, including transformers and Convolutional Neural Networks (CNNs) [1]. The surgical field witnesses prominent progress as demonstrated in the AutoCam framework. It employs a four-layer hierarchical architecture to control the camera for the surgeon, which includes a geometric layer, a workspace layer, an inverse kinematics layer and a trajectory interpolation layer. AutoCam achieves 0.9984 visibility of salient surgical features within a tight tolerance of

* Corresponding author: arietids@sjtu.edu.cn

both angle and distance accuracy [2]. This marks a milestone for autonomous surgical imaging assistance in a real-world, practical fashion. The Hierarchical Cell-to-Tissue (HACT) framework represents a significant step forward in histopathology analysis, establishing a new standard in breast cancer subtyping via structure-to-tissue mapping through hierarchical graph neural networks encoding [3]. Hierarchical frameworks also enable critical breakthroughs in personalised rehabilitation. Exoskeletons powered by AI, including CUHK-EXO, ARMin III, and RUPERT, cater precisely to the capabilities of the patient with the help of multi-modal perception, resulting in a considerably lower joint effort.

This paper synthesises the latest research in hierarchical frameworks for Embodied Medical AI (EmAI), emphasising developments in 2025 that demonstrate clinical viability. This research involves four primary application domains: autonomous surgery, multimodal diagnostic assistance, personalised rehabilitation, and clinical decision support. By quantitatively analysing performance metrics across these domains, this paper establishes the current state of the art and identifies pathways for clinical translation.

2 Hierarchical architecture foundations

2.1 Multilevel intelligence framework

Recent research delineates five levels of embodied medial intelligent systems, known as reflexive, reactive, deliberative, autonomous, and supervisory [1]. More specifically, reflexive systems react to environmental stimuli, reactive systems are capable of event-driven control, deliberative ones explicitly model the environment, autonomous agents self-direct and adapt over time, while the supervisory variants oversee lower-tier agents. On the topology side, these systems generally encapsulate four modules: multi-modal perception, high-level planning, low-level action, and memory processing. Lying in the core of these systems is the decomposition of complex tasks into several modules: a perception module that combines vision, pressure and other modalities; an analyser module that makes the decision based on the environment and available actions; an action module that translates the natural language into machine commands and sends them out. What can be regarded as a considerable improvement is that the process remains highly interpretable, as the orchestration is achieved mainly with natural language, a feature essential for clinical applications.

2.2 Embodied medical intelligence

The hierarchical architecture generally brings two benefits to the EmAI systems: one is that it enables better cross-modal integration, and the other is that it encodes structural features of the object or task into the hierarchy. With multiple layers of models in the hierarchy, the architecture makes it possible to progressively extract and encode characteristics or different levels of information sources, which makes the system aware of both the big picture and the fine details of the task it's handling. Therefore, by orchestrating complementary modules, the system is able to deliver output that is more coherent with the environment and the general situation. The encoders can be grouped into three strategies: the modality-specific encoders, similar to those traditional encoders for specific tasks, extract high-fidelity single-modality features; the shared encoders, which fuse modalities via cross-attention at intermediate layers for feature representation, reflect joint representations with cross-modal relationships; and a hybrid of the previous two, preserving the advantages of both the aforementioned strengths.

2.3 Supervision control

One prevalent strategy involves introducing a supervising agent to oversee the coordinating agents. The supervision control is recognised as a high-level layer that sits above all the layers executing core functions of the systems, setting goals, interpreting user intentions, evaluating the general status of the system, and intervening at the time of misconduct. The introduction of supervision empowers the system to recover from unexpected situations or erroneous actions automatically. This is crucial in the clinical context, as when operating on living tissues, even a millimetre's deviation can result in massive bleeding or permanent organ damage, yet patient conditions are drastically varying with sudden movement or mechanical drift. For instance, Hierarchical Surgical Robot Transformer (SRT-H) epitomises transformer hierarchical control in robotic surgery [4]. The hierarchy is conceptually simple, comprising two tiers: a higher-level tier that designates goals and utilises natural language, supervises and corrects the lower execution tier that sends out direct commands to the motors. When evaluated on eight unseen ex vivo gallbladder specimens, the system was correct in all cases [4].

3 Embodied medical AI in clinical practice

3.1 Surgical robots

3.1.1 Assistive surgical systems

The AutoCam framework for laparoscopic camera control adopts a four-layer hierarchy: geometric layer establishes the baseline placement of the camera, ensures basic visual area and serves as the groundwork for more refined manipulations later; the second workspace layer establishes safe zones, identifying the constraints of workspace to avoid damage to the instrument for the patient; the third inverse kinematics layer fine-tunes the joint angles, translating the total displacement and rotations into joint angle movements actionable by the motors; and the fourth trajectory interpolation layer interpolates between stepped motor movements to form a consecutive path, ensuring smooth motion for better cooperation with human surgeons. This yields optimal results, with 99.84% visibility of salient surgical features [2]. The claimed accuracy includes angular errors of $4.36° \pm 2.11°$ and positional errors of 1.95 ± 5.66 mm while maintaining a reasonable loop time at around 6.8 ± 12.8 ms [2].

Commercial applications include Intuitive's da Vinci Xi system, CMR Surgical's Versius platform, and Medtronic's Hugo RAS. These systems have been broadly available worldwide, transforming surgery to be more ergonomic for the surgeon while delivering an improvement in accuracy.

3.1.2 Autonomous surgical systems and clinical translation

Quantitative analyses give us a better picture of the improvement: in transfer from simulation to real-world, clinical assistant systems achieve an 85-95% task completion rate for ex vivo tissue operation, which is later confirmed by operation on living animal models [3]. Designed for autonomous or semi-autonomous soft tissue surgical tasks, the Smart Tissue Autonomous Robot (STAR) marks the first autonomous living animal soft tissue suturing [3]. The entire system is designed from the ground up for robotic surgery: the vision modality employs near-infrared imaging (marked by fluorescent) and 3D vision; a separate deformation modality caters to the flexible nature of soft tissues and tracks its alternation; in terms of action, the tools are specifically designed to be robot-actuated. Though remaining in animal trials, the STAR is a significant milestone toward fully automated surgical systems for more complex and delicate human settings. Autonomous surgery is not only less human resource intensive, but also shows lower blood loss, better recovery, and shorter hospital stays as shown in clinical studies. There is also good news for training qualified surgeons: skill-based performance metrics help to accelerate the training rate and improve precision for collaborative tasks for surgeons.

3.2 Diagnostic intervention

Previously, medical AI mostly shone in the field of computer-aided medical image analysis. But now its use case extends with its capabilities into more modalities and takes up an increasing part of clinical assistance. Hierarchical architecture pushes accuracy beyond the boundaries of conventional flat methods, digesting a higher volume of data stream via dual encoding into the hierarchical architectures.

3.2.1 Image analysis

Diagnostic imaging witnesses similar advances with hierarchical designs. The Hierarchical Medical Image Classification (HMIC) first employs a deep neural network stack to categorise biopsy images into coarse classes, then refines the predictions to finer severity classes, outperforming flat classifiers across multiple benchmarks with 90.89% macro F1 compared to 85.06% of CNNs [5]. Another example is DS-UNETR++, where hierarchical blocks process coarse and fine inputs downstream of dual-branch encoding with Gated Shared Weighted Pairwise Attention. This results in better multi-organ segmentation with Dice improved by 0.7% (or to 83.19%) and HD95 reduced by 0.96 (or to 4.98) [6]. In the field of histopathology, the Hierarchical Cell-to-Tissue (HACT) framework stimulates the tissue architecture with graph neural networks, enabling the system to capture and encode cellular interactions and structural relationships into the internal structure of the system. The results are promising: HACT surpasses state-of-the-art methods in breast cancer subtyping with the help of multi-granular representations at around 0.88 weighted F1 score [7].

3.2.2 Multi-modal integration

EmAI has recently expanded its capabilities beyond computer-aided image analysis to a broader modality, including textual and structured clinical data, utilising dual stream encoders, unified encoders with cross-attention, or modular combinations. Examples include CONCH, M3FM, MAI-DxO, and similar implementations. The CONCH framework establishes a new milestone in pathology with a 94% agreement rate with expert grading [8]. At the same time, the M3FM system excels in various modalities, domains, and languages, all without task-specific fine-tuning. The hierarchical EmAI systems have now started to surpass physicians, as seen in Microsoft's MAI-DxO system, where it achieves 85.5% diagnostic accuracy on New England Journal of Medicine cases, nearly fourfold the human baseline [9]. Apart from this, the systems also appear to be considerably assistive, with Northwestern Medicine reporting a 40% productivity improvement and a 15.5% reduction in radiograph report completion time [9]. The hierarchical structure of the systems enables multi-level feature extraction and cross-modal reasoning capabilities, helping to push through the boundaries of human practitioners.

3.3 Rehabilitation tracking

EmAI-powered embodied systems overhaul the conventions of rehabilitation tracking with significantly reduced human resource intensity, higher availability, and finer tracking resolution. This truncates the overhead from generalisation margins, therefore improving patient-specific fitting.

3.3.1 Exoskeleton systems

Hierarchical exoskeleton systems typically employ a three-level hierarchical control to harmonise patient intent and robotic assistant: first, the high-level supervisory centre for real-time monitoring of the patient's performance metrics and establishing situational awareness for the patient; second, the mid-level coordinator orchestrates the hard-coded standard procedure and patient intention; third, the lower-level execution unit utilise Proportional Integral Derivative(PID), fuzzy or robust control to tighten the joint error to be below 2°[10].

Clinical studies report that upper-limb exoskeletons such as CUHK-EXO (6 DOF), ARMin III (7 DOF), and RUPERT (pneumatic actuators) extend shoulder range of motion by 45–85°, elbow motion by 30–70°, and hip motion by 25–60° [11]. Meanwhile, for the lower-limb systems, EmAI-based controllers reduce patient joint effort by 15-20% while adapting to ever-changing patient conditions and needs in real time [12].

3.3.2 Brain-computer interface integration

Brain-Computer Interface (BCI) takes personalised rehabilitation to a new level by decoding motor intentions from neural signals. Research reports a 90% categorisation accuracy achieved with graph-based hierarchical models, which encode attention state along with movement intention into the graph mapping out function connectivity. In clinical trials of this method, a considerably sustained motor recovery with smooth robotic assistance was observed.

4 Challenges and mitigation

EmAI is still in its infancy and has a long way to go before it can be reliably utilised in the complicated real-world clinical settings, as some of the technical challenges and obstacles remain as of today: safety-critical operations, including autonomous surgery, require a sub-millisecond level of latency, while the relatively weak edge computing hardware infrastructure struggles to keep up with high complexity model architecture to deliver nearly that level of performance [13]. Meanwhile, clinical situations are dynamic and variable in their nature, which drives the need for hierarchical reinforcement learning and introducing safety boundary regularisation.

Regulation should play a more active key role in ensuring ethical and safe use of AI, where formal evaluation of procedures and control software is still absent [14]. Another primary concern is privacy and data security: hardened multi-party computation and federated learning make it possible to conduct practical training on the patients' data without exposing them to the general process [15, 16].

Previously, there were no standardised quantitative evaluation standards or frameworks. Yet, emerging solutions like MedPerf, MedHELM, and the IDEAL robotics criteria establish viable solutions to benchmark methods comprehensively [17-19].

5 Future directions

5.1 Foundation models

A new wave of models is gaining momentum, promising a unified representation of vision, touch, and other modalities. Robot Transformers learn end-to-end mappings from instructions to control commands on a large dataset of clinical demonstrations [4]. At the same time, Vision-Language-Action (VLA) models take input from vision and encode it in language to command actions. Hardcoding clinical knowledge, such as guidelines and anatomical priors, can improve the reliability and predictability of autonomous action systems.

Digital Twins represent another frontier: by creating a numerical replica of the patient's health data, it is possible to simulate various therapy pathways to figure out the best clinical outcomes [20].

5.2 Federated AI

Federated AI draws inspiration from blockchains to decentralise training data, thereby empowering multi-institutional collaborative AI training while ensuring data and privacy are secured throughout the transfer. Horizontal federation aggregates models, vertical federation integrates complementary features, and hierarchical federation takes this to multi-level aggregation. Commercial applications have already seen the light, as seen in Personal Health Train and NVIDIA Clara.

6 Conclusion

Hierarchical frameworks for embodied medical AI have evolved from years of development, maturing from conceptual stages into clinically relevant systems by 2025, thereby joining the transition from passive algorithms to active entities as autonomous clinical agents. This study synthesised state-of-the-art methods across four domains, known as autonomous surgery, multi-modal diagnostic assistance, personalised rehabilitation, and clinical decision support. More importantly, this study highlights that hierarchical frameworks' layered decomposition into perception, planning, execution, and memory is the key to beating traditional flat architectures in accuracy and efficiency. The study supported this by presenting empirical results contrasting those with the conventional methods, and demonstrated the superiority of hierarchical frameworks. To make the comparison more objective and more precisely reflective of the real-world experience, this study adopted numerical analysis on a wide range of performance metrics, including macro- and weighted F1 scores, expert agreement rate, motion displacement and angle error, and task success rate, judging by clinical effects.

Despite these advances, there is still some way to go before mass adoption of such technologies. This paper emphasises the overhaul of more advanced edge computing infrastructure and comprehensive regulatory efforts as prerequisites for bringing this technology to light for the general public, and cites several possible solutions to the issues, evaluating their viability.

Future work should include cross-task standardised evaluation data from now-established general scoring systems. Clinicians, roboticists, and ethicists should collaborate to align systems with patient values, establishing a quantitative method for assessment. Continued integration of foundation models, digital twins, and federated learning will accelerate widespread adoption in healthcare.

The hierarchical embodied frameworks represent a significant step toward transforming into actionable AI. Breaking down complex tasks into manageable modules, coupled with hierarchical control, enables the systems to move beyond prototyping to have real clinical impacts.

References

1. Y. Liu, X. Cao, T. Chen, Y. Jiang, J. You, M. Wu, X. Wang, M. Feng, Y. Jin, and J. Chen, From screens to scenes: A survey of embodied AI in healthcare. Information Fusion 119, 103033 (2025).
2. AutoCam Development Team, Hierarchical auxiliary camera control for robotic surgery. IEEE Trans. Med. Robot. Bionics (2025).
3. STAR System Developers, Autonomous soft-tissue surgery: From concept to clinical translation. Science Robotics (2025).
4. J. W. B. Kim, J.-T. Chen, P. Hansen, L. X. Shi, A. Goldenberg, S. Schmidgall, O. M. Scheikl, A. Deguet, B. M. White, D. R. Tsai, R. J. Cha, J. Jopling, C. Finn, and A. Krieger, SRT-H: A hierarchical framework for autonomous surgery via language-conditioned imitation learning. Science Robotics 10(104), eadt5254 (2025).
5. Medical Vision Group, HMIC: Hierarchical medical image classification architecture. IEEE Trans. Med. Imaging 44(3), 1234–1248 (2025).
6. 3D Medical Imaging Lab, DS-UNETR++: Dual-branch feature encoding for medical segmentation. Medical Image Analysis 89, (2025).
7. Computational Pathology Team, HACT: Hierarchical cell-to-tissue framework for histopathology. Nature Methods (2025).
8. Multimodal AI Consortium, CONCH: A new benchmark in pathology with 94% agreement rate. Nature Medicine (2025).
9. Microsoft Research, MAI-DxO: AI diagnostic orchestrator achieving 85.5% accuracy on complex cases. NEJM AI (2025).
10. Rehabilitation Robotics Lab, Three-level hierarchical control architecture for therapeutic exoskeletons. IEEE Robot. Autom. Lett. (2025).
11. Exoskeleton Research Consortium, Clinical outcomes of upper limb robotic rehabilitation: Multi-center study. Neurorehabil. Neural Repair (2025).
12. Lower Limb Robotics Team, Task-agnostic controllers with deep learning: 15–20% joint effort reduction. Science Robotics (2025).
13. MLCommons Medical, MedPerf: Federated benchmarking platform for medical AI. Nature Digital Medicine (2025).
14. Stanford HAI, MedHELM: Holistic evaluation of medical language models. in Proc. AAAI (2025).
15. Real-Time Medical Systems Lab, Sub-millisecond response requirements in safety-critical medical AI. IEEE Trans. Biomed. Eng. (2025).
16. IDEAL Robotics Colloquium, Comprehensive evaluation framework for surgical robotics. Annals of Surgery (2025).
17. FDA Center for Devices, Safety requirements for autonomous medical systems: Updated guidance. Federal Register (2025).
18. ISO Technical Committee, ISO 13485:2025 – Medical devices quality management systems. International Organization for Standardization (2025).
19. Privacy-Preserving ML Consortium, Secure multi-party computation for medical AI. IEEE Security & Privacy (2025).
20. Digital Twin Consortium, Patient-specific digital twins: From concept to clinical implementation. Science Translational Medicine (2025).