2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

UltraHiT: A Hierarchical Transformer Architecture for Generalizable Internal Carotid Artery Robotic Ultrasonography

Teng Wang[1,*], Haojun Jiang[1,*,§], Yuxuan Wang[2,*], Zhenguo Sun[3], Xiangjie Yan[1], Xiang Li[1] and Gao Huang[1,†]

***Abstract*— Carotid ultrasound is crucial for the assessment of cerebrovascular health, particularly the internal carotid artery (ICA). While previous research has explored automating carotid ultrasound, none has tackled the challenging ICA. This is primarily due to its deep location, tortuous course, and significant individual variations, which greatly increase scanning complexity. To address this, we propose a Hierarchical Transformer-based decision architecture, namely UltraHiT, that integrates high-level variation assessment with low-level action decision. Our motivation stems from conceptualizing individual vascular structures as morphological variations derived from a standard vascular model. The high-level module identifies variation and switches between two low-level modules: an adaptive corrector for variations, or a standard executor for normal cases. Specifically, both the high-level module and the adaptive corrector are implemented as causal transformers that generate predictions based on the historical scanning sequence. To ensure generalizability, we collected the first large-scale ICA scanning dataset comprising 164 trajectories and 72K samples from 28 subjects of both genders. Based on the above innovations, our approach achieves a 95% success rate in locating the ICA on unseen individuals, outperforming baselines and demonstrating its effectiveness. Project website: `https://ultrahit-thu.github.io/UltraHiT/`.**

I. INTRODUCTION

Ultrasonography is the preferred method for assessing carotid artery health due to its real-time imaging, radiation-free, and cost-effectiveness features. The procedure requires operators to precisely adjust the probe's angle and position to obtain optimal imaging planes, demanding not only anatomical knowledge and interpretation skills for ultrasound images but also operational proficiency. Training an experienced sonographer typically takes several years, resulting in a shortage of professionals in underdeveloped regions. Moreover, ultrasound scanning is a highly repetitive and intensive task, where prolonged operation may lead to fatigue and compromised image quality Fig. 1(a). These factors drive the development of autonomous ultrasound robots.

Carotid ultrasound robots are expected, with perception and decision-making capabilities, to autonomously adjust the probe based on real-time imaging, thereby obtaining clear transverse and longitudinal views of the vessels. Current researches primarily focus on the automated scanning of the common carotid artery (CCA). Specifically, researchers [1]–[8] attempted to embed the knowledge and skills which required for carotid ultrasound into sophisticated rules, known as rule-based methods. By analyzing image changes resulting from specific probe adjustment actions, they establish a set of mapping rules from image features to motion decisions. Nevertheless, their reliance on an identical vascular model and neglect of anatomical variability limit their generalization ability. The ICA's substantial variability makes creating an exhaustive rule set nearly impossible.

Inspired by recent advances in deep learning, another line of research [10]–[18] attempts to adopt learning-based methods for better generalization capability. Learning-based methods include reinforcement and imitation learning. Reinforce-

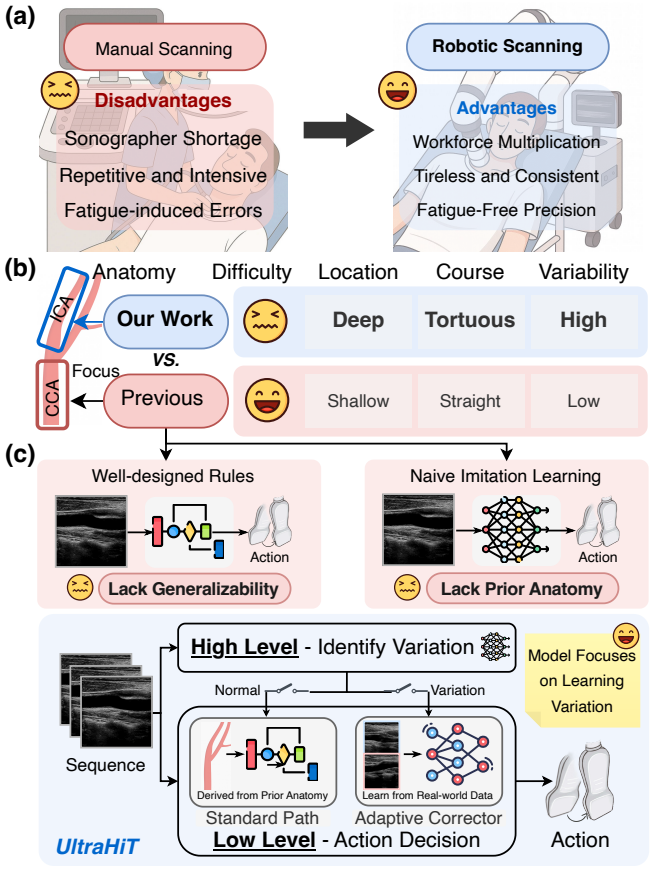

Fig. 1: **Overview**. (a) Robotic scanning vs. manual scanning. (b) Characteristics of ICA vs. CCA. (c) Our hierarchical transformer architecture vs. previous works.

This work is supported in part by the National Key Research and Development Program of China under Grant 2024YFB4708200 and the Scientific Research Innovation Capability Support Project for Young Faculty under Grant ZYGXQNJSKYCXNLZCXM-I20.

AI assistance: The background of Fig. 1(a) was generated using ChatGPT.

[1]Department of Automation, BNRist, Tsinghua University, Beijing, China. [2]School of Computer Science and Technology, Xidian University, Xi'an, China. [3]Beijing Academy of Artificial Intelligence, Beijing, China.

[*]Equal contributions. [†]Corresponding author.

[§]Haojun Jiang guided this work.

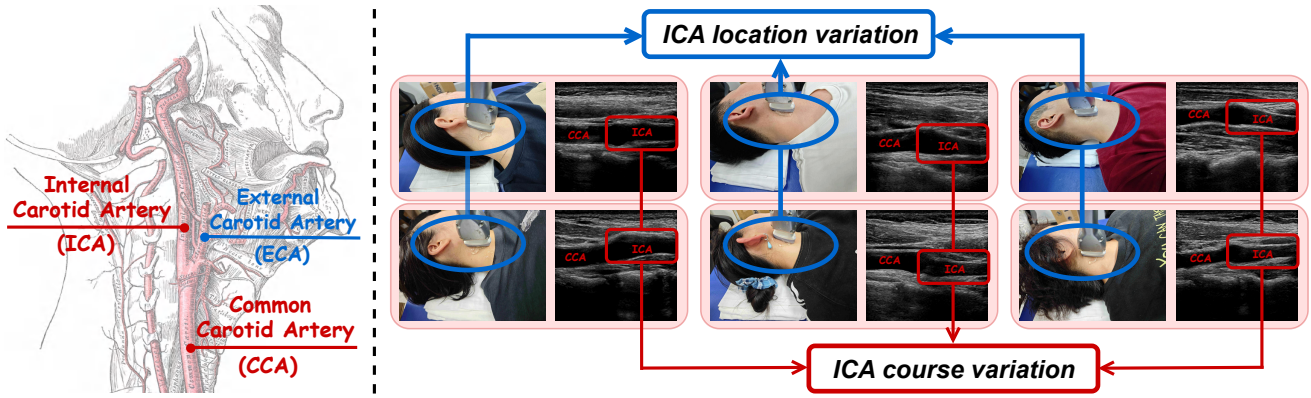

Fig. 2: **Internal carotid artery's anatomy and variability**. The left panel shows the standard anatomy of the ICA [9], while the right panel demonstrates the significant variability in the position and course of the ICA within the population.

ment learning method with either virtual vessel models [15] or 3D CT data of organs [16], [17], [19] is used to simulate the environment. However, while the former suffers from poor generalization due to limited anatomical variations, the latter relies on costly and hard-to-scale 3D CT data. Another category of methods [10]–[13] uses imitation learning, which collects expert operation data from real-world individuals and trains models to imitate expert decisions. Recently, work [10] demonstrated that the imitation learning strategy exhibits good generalization potential.

Despite these developments, existing automation efforts [1], [3]–[5], [10], [13], [15] have primarily focused on the CCA, overlooking the clinically more significant and challenging ICA (Fig. 1b). To enhance carotid robots, this paper targets the ICA—a primary brain blood pathway extending from the CCA (Fig. 2 left). Therefore, the ICA represents a critical window for evaluating cerebral blood supply, and obtaining its longitudinal view is particularly essential in carotid artery examinations. However, ultrasonographic scanning of the ICA is more challenging than that of the CCA, primarily due to significant individual variations in: (1) vascular course, and (2) positional anatomy. Fig. 2 illustrates ICA conditions in six subjects: ultrasound images reveal differences in its course, while camera images show variations in its location. This requires the ultrasound robot to possess strong adaptive capabilities, enabling it to recognize individual anatomical differences in real-time and make adjustment decisions, which presents a significant challenge.

To address this, as shown in Fig. 1c, we propose a **Hi**erarchical **T**ransformer-based decision architecture named **UltraHiT**, which integrates high-level variation identification with low-level action decision. Our core idea is to treat individual vascular structures as variations derived from a standard vascular template. We first construct a high-level state assessment module to identify whether vascular variations are present (*i.e.*, whether adaptive adjustment is needed). Furthermore, based on the high-level semantic signal, the system activates one of two low-level decision executors: **(1)** A standardized path executor, based on anatomical knowledge from the standard ICA vascular model. This standard vascular model inherently encapsulates common anatomical knowledge across populations, embodying a knowledge-enhanced approach that handles basic and normal scenarios effectively. Equipped with this prior knowledge, the learning-based component is freed from learning basic structural information and can be tailored to focus on capturing vascular variations, thereby enhancing its generalization capability. **(2)** An adaptive corrector, trained through imitation learning on large-scale data to capture potential individual variations.

Specifically, the high-level module and the adaptive corrector are both implemented as causal transformers. They are trained independently, each with its own distinct supervisory signal tailored to their specific tasks. Furthermore, we posit that incorporating historical scan data provides richer structural information about the subject, thereby facilitating more informed decision-making. Consequently, the model input is extended to include a sequence of historical scans rather than a single image [1], [4], [10], [15]. Finally, to train the model and ensure its generalizability, we collected the first large-scale ICA scanning dataset comprising 164 trajectories and 72K samples from 28 subjects of both genders.

To validate the effectiveness of our model, we conducted real-world experiments by testing the scanning success rate on unseen subjects of both gender. We compared our approach with rule-based and imitation learning methods, and the results demonstrate the efficacy of our proposed model. Furthermore, failure case analysis shows that our model does not exhibit issues such as complete deviation from the target, mis-targeting, or timeout before task completion, which are commonly observed in baseline models. Additionally, our method demonstrates stronger robustness under challenging conditions, including poor initial states and subject movement. In summary, the contributions are three fold:

- We are the first to achieve autonomous scanning of the ICA longitudinal section as required in clinical practice, extending the capabilities of carotid robots.
- We propose a novel hierarchical transformer-based decision architecture named UltraHiT, which effectively handles anatomical variations in the ICA.
- We have collected the first expert demonstration dataset for ICA scanning, comprising 164 trajectories and 72K samples from 28 subjects of both genders.

II. Related Works

A. Rule-based Ultrasound System

One category of researchers employs rule-based methods [1]–[8]. For instance, some works [1], [2], [4] designed visual servoing methods based on the relationship between vascular images and the angle/position of the probe. Yan et al. [3], [8] utilized external cameras to identify neck structures and planned scanning trajectories based on the recognition results. Such approaches are generally built upon a standard vascular model that captures common structural features across the population, thereby achieving a certain level of adaptability. However, due to significant variability in individual vasculature, the generalization capability of rule-based approaches is fundamentally limited.

B. Learning-based Ultrasound System

Another category of work is the learning-based approach that has emerged in recent years [10]–[18]. For instance, Bi et al. [15] created a virtual vascular simulation environment for training reinforcement learning models, but the vessel models are idealized and lack realistic anatomical variations. In contrast, others [16], [17], [19] used individual 3D CT scan data to construct the simulation environment—such data encompass real-world variations, yet are costly to acquire and difficult to scale. Although reinforcement learning holds potential for learning optimal scanning policies, significant challenges in simulating realistic environments hinder its further development. Another line of research follows the imitation-learning approach. For instance, Jiang et al. [10] collected a large-scale expert scanning dataset to enable imitation learning and demonstrated its effectiveness in scanning individuals with carotid plaques, highlighting the potential of the method; Deng et al. [11] and Droste et al. [12] also validated the feasibility of this technology in abdominal and fetal imaging, respectively. While such methods rely entirely on learning common structures and individual variations from data, the common anatomical features have already been well-established as medical knowledge. Relearning them from data results in lower data utilization efficiency. The strength of learning-based methods lies in their ability to effectively model variations. Therefore, the focus of learning should be placed on capturing individual-specific variability. Furthermore, existing methods have primarily focused on the automatic scanning of the common carotid artery (CCA). Exploration of the internal carotid artery (ICA), which is more challenging and clinically significant, remains insufficient.

III. Method

To address significant individual variations in ICA scanning, we propose a **Hi**erarchical **T**ransformer-based decision architecture, named **UltraHiT**. Our approach conceptualizes the scanning process as a sequence of high-level state assessments and low-level action decisions. This hierarchical design allows the system to leverage a knowledge-based standard scanning path for common anatomies while employing a data-driven adaptive model to handle complex variations. The overall architecture is shown in Fig. 3.

A. High-Level State Assessment Module

Corrective Gate. The Corrective Gate is designed to assess the current scanning trajectory and determine whether vascular variations are present, indicating the need for adaptive correction. It is implemented as a Causal Transformer [20] which takes a sequence of historical data H_t (including past images and actions) as input to generate a binary policy switch signal $g_t \in \{0, 1\}$:

$$g_t = \mathcal{H}_{\text{gate}}(\mathcal{T}_{\text{gate}}(H_t)) \tag{1}$$

where $\mathcal{T}_{\text{gate}}$ represents the Causal Transformer backbone and $\mathcal{H}_{\text{gate}}$ is the dedicated Policy Switch Head. $g_t = 1$ activates the Adaptive Corrector, while $g_t = 0$ activates the Standardized Path Executor. The detailed architecture of the Causal Transformer $\mathcal{T}_{\text{gate}}$ will be elaborated in Sec. III-C.

Stop Model. The Stop Model is responsible for identifying the completion of scanning stages. As shown in Fig. 3(c), it processes only the current ultrasound image $\mathbf{I}_t \in \mathbb{R}^{C\times H\times W}$ to generate a stage switch signal. We utilize a ResNet architecture, Φ_{ResNet}, pre-trained on ImageNet, as the feature extractor, followed by a linear layer to make the final decision. The process can be formulated as:

$$s = \text{Softmax}(\mathbf{W}_{\text{stop}} \cdot \Phi_{\text{ResNet}}(\mathbf{I}_t) + \mathbf{b}_{\text{stop}}) \tag{2}$$

where $\mathbf{W}_{\text{stop}}$ and $\mathbf{b}_{\text{stop}}$ are the weight and bias of the final linear layer, and s is the probability distribution over stage completion signals.

B. Low-Level Action Decision Module

Standardized Path Executor. As illustrated in Fig. 3d, the Standardized Path Executor is a deterministic policy based on anatomical prior knowledge of carotid arteries. The ICA runs posterior-laterally while the ECA runs anterior-medially (Fig. 2(left)). At the CCA bifurcation, the ICA is more lateral, making it the first to be seen when the probe rotates clockwise around the z-axis. Based on this anatomical configuration, we define the action a_t in each state s_t as:

$$a_t = \begin{cases} \text{Forward}, & \text{if } s_t \in \text{Stage1}, \\ \text{+Yaw}, & \text{if } s_t \in \text{Stage2}. \end{cases} \tag{3}$$

where "Forward" advances along the local x-axis "+Yaw" rotates clockwise around the z-axis (Fig. 3d). This routine policy provides an effective and reliable scanning strategy based on anatomical knowledge, thereby alleviating the learning burden on data-driven corrective model.

Adaptive Corrector. The Adaptive Corrector is the core component for handling anatomical variability. It is a data-driven policy learned from expert demonstrations. Its goal is to generate corrective actions that bring a deviating scanning path back to the optimal trajectory. It utilizes a Causal Transformer architecture [20] to map the state history H_t to a corrective policy π_t^{corr}:

$$\pi_t^{\text{corr}} = \mathcal{H}_{\text{corr}}(\mathcal{T}_{\text{corr}}(H_t)) \tag{4}$$

where $\mathcal{T}_{\text{corr}}$ is the Causal Transformer and $\mathcal{H}_{\text{corr}}$ is the actor head that outputs a probability distribution over the action

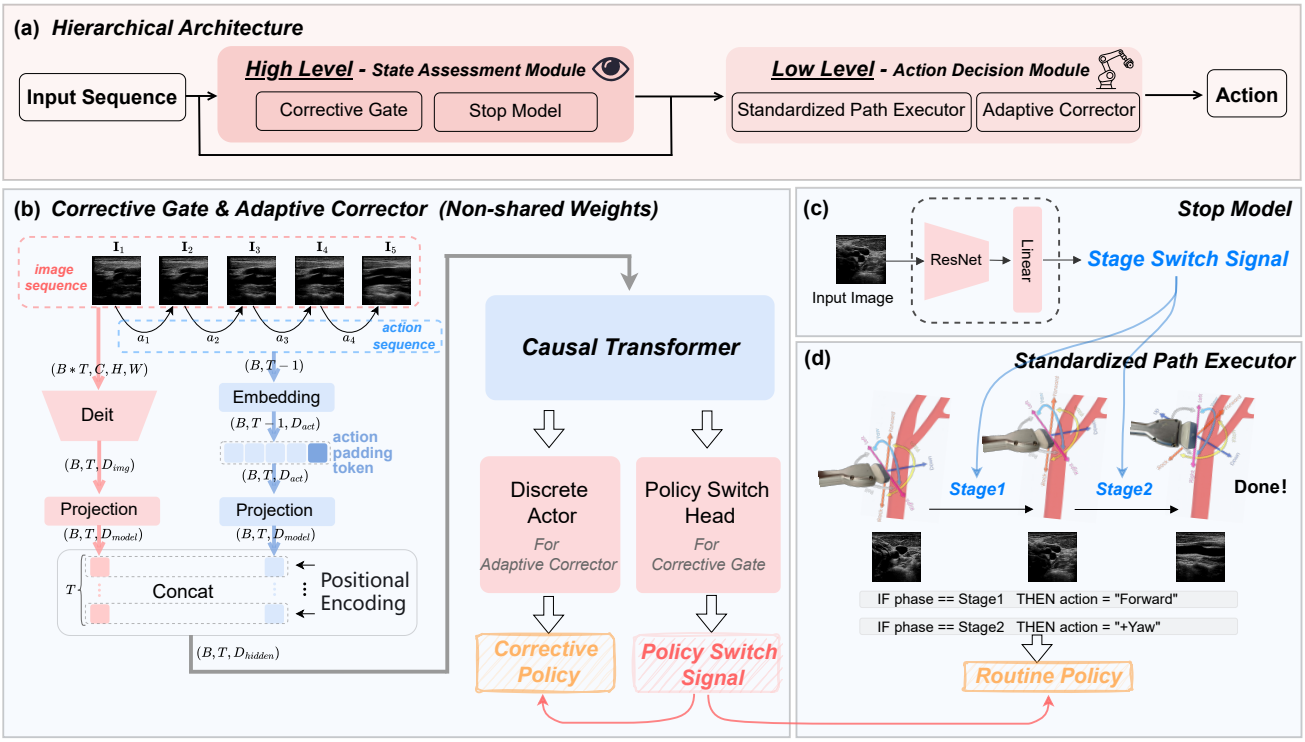

Fig. 3: **Hierarchical transformer architecture**. (a) Overview of hierarchical architecture. The high-level module makes semantic decisions, while the low-level module executes physical actions in the real world. (b) The corrective gate and adaptive corrector, process image-action sequences through a causal transformer. (c) The stop model architecture. (d) The standardized path executor, a knowledge-based policy designed using anatomical prior knowledge.

space. The action with the highest probability is then selected for execution. The action space consists of 12 discrete actions, corresponding to 12 different movement directions. These movements include translations and rotations along the x, y, and z axes of the probe's coordinate system.

C. Causal Transformer for Decision Making

Both the Corrective Gate and Adaptive Corrector are based on a Causal Transformer backbone (Fig. 3(b)), which models temporal dependencies across ultrasound images and past actions to inform future decisions.

Input Representation and Embedding. The input consists of a sequence of the last T ultrasound images $\mathcal{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_T\}$ and $T-1$ past actions $\mathcal{A} = \{a_1, \ldots, a_{T-1}\}$.

Each image $\mathbf{I}_t$ is encoded via a Vision Transformer (DeiT, Φ_{DeiT}) into a feature vector, then projected into $\mathbb{R}^{D_{\text{model}}}$:

$$\mathbf{f}_t^{\text{img}} = \Phi_{\text{DeiT}}(\mathbf{I}_t) \in \mathbb{R}^{D_{\text{img}}} \tag{5}$$

$$\mathbf{x}_t^{\text{img}} = \mathbf{f}_t^{\text{img}} \mathbf{W}_{\text{img\_proj}} \in \mathbb{R}^{D_{\text{model}}} \tag{6}$$

where $\mathbf{W}_{\text{img\_proj}} \in \mathbb{R}^{D_{\text{img}} \times D_{\text{model}}}$ is a projection matrix.

Past actions are embedded using a learnable matrix $\mathbf{E}_{\text{act}}$. An action padding token $\mathbf{a}_{\text{pad}}$ is appended for length alignment, and the sequence is projected via an MLP $\Phi_{\text{act\_proj}}$:

$$\mathbf{e}_t = \begin{cases} \mathbf{E}_{\text{act}}[a_t] & \text{if } 1 \le t \le T-1 \\ \mathbf{a}_{\text{pad}} & \text{if } t = T \end{cases} \tag{7}$$

$$\mathbf{x}_t^{\text{act}} = \Phi_{\text{act\_proj}}(\mathbf{e}_t) \in \mathbb{R}^{D_{\text{model}}} \tag{8}$$

Sequence Modeling with Causal Transformer. At each step t, image and action features are concatenated into a token $\mathbf{z}_t \in \mathbb{R}^{D_{\text{hidden}}}$ ($D_{\text{hidden}} = 2 \cdot D_{\text{model}}$). Learnable positional embeddings $\mathbf{P}$ are added:

$$\mathbf{H}_0 = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T]^T + \mathbf{P}_{1:T} \in \mathbb{R}^{T \times D_{\text{hidden}}} \tag{9}$$

The sequence $\mathbf{H}_0$ is then processed by L Causal Transformer blocks. Each block consists of a Causal Multi-Head Self-Attention (MHSA) layer and an MLP with pre-normalization. For the l-th block:

$$\mathbf{H}'_l = \text{LayerNorm}(\mathbf{H}_{l-1}) \tag{10}$$

$$\mathbf{H}''_l = \text{MHSA}(\mathbf{H}'_l) + \mathbf{H}_{l-1} \tag{11}$$

$$\mathbf{H}_l = \text{MLP}(\text{LayerNorm}(\mathbf{H}''_l)) + \mathbf{H}''_l \tag{12}$$

The causal MHSA uses h heads. For each head i, queries $\mathbf{Q}_i$, keys $\mathbf{K}_i$, and values $\mathbf{V}_i$ are derived via linear projections. The attention scores are computed using scaled dot-product attention with a causal mask $\mathbf{M}$:

$$\text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = \text{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d_k}} + \tilde{\mathbf{M}}\right)\mathbf{V}_i \tag{13}$$

The causal mask $\tilde{\mathbf{M}}$ is a lower-triangular matrix where $\tilde{M}_{ij} = 0$ if $j \le i$ and $\tilde{M}_{ij} = -\infty$ otherwise, ensuring that the prediction at time step t only depends on past inputs.

The outputs of all heads are concatenated and projected back to the hidden dimension:

$$\text{MHSA}(\mathbf{H}'_l) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \tag{14}$$

where $\text{head}_i = \text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i)$ and $\mathbf{W}^O \in \mathbb{R}^{D_{\text{hidden}} \times D_{\text{hidden}}}$ is the output projection matrix.

Output Heads. After the final transformer block, we apply layer normalization and use only the output token at the last time step $\mathbf{h}_{\text{final}}$ for decision-making. This vector encapsulates the information from the entire history. This final hidden state is then fed into task-specific heads:

- For the Adaptive Corrector, a discrete actor head outputs the logits for the corrective action policy:

$$\pi_{\text{corr}} = \text{Softmax}(\mathbf{h}_{\text{final}}\mathbf{W}_{\text{actor}} + \mathbf{b}_{\text{actor}}) \tag{15}$$

- For the Corrective Gate, a policy switch head outputs the logits for the gate signal:

$$g_{\text{logits}} = \mathbf{h}_{\text{final}}\mathbf{W}_{\text{gate}} + \mathbf{b}_{\text{gate}} \tag{16}$$

The Causal Transformer backbone weights are not shared between the Corrective Gate and Adaptive Corrector, enabling task specialization. Both models are trained independently using separate supervisory signals.

D. Robot Control Algorithm

We employ Cartesian impedance control [21], following the same strategy of [10], to balance accuracy and compliance during scanning. The joint dynamics of the Franka robotic arm can be described as:

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{C}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{g}(\mathbf{q}) = \tau + \tau_{\text{ext}}, \tag{17}$$

where $\mathbf{M}$, $\mathbf{C}$, and $\mathbf{g}$ denote the mass matrix, Coriolis/centrifugal forces, and gravity vector, respectively. τ is the commanded torque, and τ_{ext} represents external torques arising from probe–neck contact.

The impedance controller generates torques:

$$\begin{aligned}\tau = \mathbf{J}^T(-\mathbf{K}\tilde{\mathbf{x}} - \mathbf{D}\dot{\mathbf{x}}) \\ + (\mathbf{I} - \mathbf{J}^T\mathbf{J}^{+T})(-\mathbf{K}_n\dot{\mathbf{q}} - \mathbf{D}_n\tilde{\mathbf{q}}) + \mathbf{C}\dot{\mathbf{q}} + \mathbf{g},\end{aligned} \tag{18}$$

where $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_d$ is the Cartesian pose error, $\mathbf{J}$ is the Jacobian, $\mathbf{K}$ and $\mathbf{D}$ are task-space stiffness and damping matrices, $\mathbf{K}_n$, $\mathbf{D}_n$ provide null-space regulation, and the superscript $+$ indicates the pseudo-inverse.

Substituting into the dynamics yields the closed-loop form:

$$\mathbf{M}\ddot{\mathbf{q}} + \mathbf{J}^T(\mathbf{K}\tilde{\mathbf{x}} + \mathbf{D}\dot{\mathbf{x}}) + (\mathbf{I} - \mathbf{J}^T\mathbf{J}^{+T})(\mathbf{K}_n\dot{\mathbf{q}} + \mathbf{D}_n\tilde{\mathbf{q}}) = \tau_{\text{ext}}. \tag{19}$$

Mapping to Cartesian space:

$$\mathbf{J}\mathbf{M}\ddot{\mathbf{q}} + \mathbf{J}\mathbf{J}^T(\mathbf{K}\tilde{\mathbf{x}} + \mathbf{D}\dot{\mathbf{x}}) = \mathbf{F}_{\text{ext}}, \tag{20}$$

where $\mathbf{F}_{\text{ext}} = \mathbf{J}^{+T}\tau_{\text{ext}}$ is the contact force. In quasi-static contact ($\ddot{\mathbf{q}} \approx 0, \dot{\mathbf{x}} \approx 0$), this simplifies to:

$$\mathbf{F}_{\text{ext}} = \mathbf{K}\tilde{\mathbf{x}}, \tag{21}$$

indicating that large pose errors may generate excessive contact forces, posing potential safety risks.

To improve safety, we adopt an error-dependent stiffness:

$$k = \begin{cases} k_{\text{normal}}, & f_{\text{ext}} < \bar{f}_{\text{ext}}, \\ \bar{\mathbf{f}}_{\text{ext}}/\tilde{\mathbf{x}}, & f_{\text{ext}} \geq \bar{f}_{\text{ext}}, \end{cases} \tag{22}$$

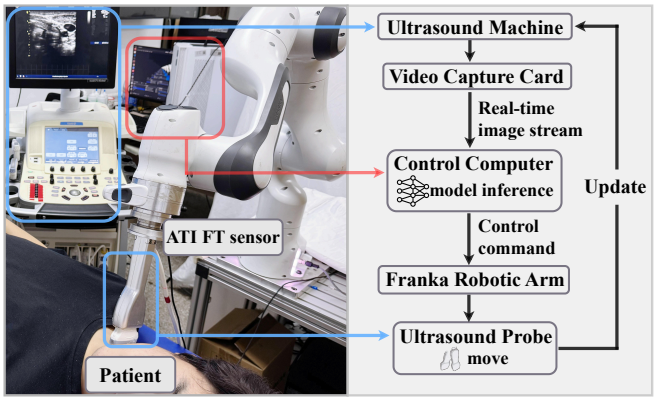

Fig. 4: Hardware and control configuration of the system.

where $\bar{\mathbf{f}}_{\text{ext}}$ is a predefined safe force threshold vector. This adjustment reduces stiffness when excessive forces are detected, limiting contact pressure on the patient.

IV. Experiments

A. Datasets and Implementation Details

Datasets. The dataset was collected by sonographers with over 10 years of experience using a GE Vivid E7 ultrasound device equipped with a 9L probe. During scanning, we synchronously recorded the ultrasound frame and the action taken at each time step. The data collection details follow the work [10]. The corpus contains multiple scans, each scan is a sequence of image–action pairs $\{(\mathbf{I}_t, a_t)\}_{t=1}^{T}$. Stage 1 uses data requested from [10], which includes 233 scan trajectories with 88,712 image–action pairs from 81 subjects (76 for training, 5 for validation). Stage 2 includes 164 **newly** collected trajectories with 72,279 image–action pairs from 28 subjects (25 for training, 3 for validation). This study was approved by the institutional ethics committee, and written informed consent was obtained from all participants.

Model Architecture. The UltraHiT is implemented in PyTorch. The Causal Transformer backbones for both the Corrective Gate ($\mathcal{T}_{\text{gate}}$) and Adaptive Corrector ($\mathcal{T}_{\text{corr}}$) utilize a model dimension $D_{\text{model}} = 256$ and consist of $L = 4$ layers. Each layer contains a Multi-Head Self-Attention mechanism with $h = 8$ attention heads and a MLP with an expansion ratio of 4. Image features are extracted using a pre-trained DeiT-Tiny backbone, with the resulting $D_{\text{img}} = 192$ dimensional features projected to D_{model} via a linear layer. Action embeddings use $D_{\text{act}} = 128$ dimensions before projection. The concatenated image-action features form tokens of dimension $D_{\text{hidden}} = 512$. Both models are trained independently using the cross-entropy loss function.

Training Strategy. We use the Adam optimizer with an initial learning rate of 1×10^{-4} and a cosine learning-rate schedule, weight decay of 1×10^{-3}, and a batch size of 256. All three models are trained with the same recipe for 10 epochs using four NVIDIA A100 GPUs.

Robotic System Configuration. We built a robotic ultrasound platform as illustrated in Fig. 4. The system employs a Franka Emika Panda arm as the manipulator and a GE Vivid E7 ultrasound device fitted with a 9L probe as the imaging

TABLE I: **Real-world experiment results.** We report: Success Rate (Final-ICA) – termination on a high-quality ICA longitudinal view; Success Rate (Pass-ICA) – attainment of such a view at any point; Feature Cosine Similarity – similarity to expert scans; Number of Corrections – count of non-routine actions; Subject Comfort Score – participant comfort (0–10).

Method	Success Rate (Final-ICA)	Success Rate (Pass-ICA)	Feature Cosine Similarity	Scan Time (s) Mean (min–max)	Number of Corrections Mean (min–max)	Subject Comfort Score Mean (min–max)
Rule-based	20% (2/10)	20% (2/10)	0.4895	73.3 (59.6–82.2)	0 (0–0)	4.8 (3–6)
Rule-based + Explore	20% (2/10)	40% (4/10)	0.6124	74.5 (54.0–90.9)	5 (0–10)	5.2 (5–6)
E2E, Single-frame [10]	35% (7/20)	50% (10/20)	0.6201	84.8 (51.9–180)	5.4 (0–38)	7.8 (7–9)
Hier, Single-frame	50% (10/20)	70% (14/20)	0.6186	95.0 (53.8–180)	9.3 (0–37)	7.0 (5–9)
Hier, Sequential (Ours)	**80% (16/20)**	**95% (19/20)**	**0.7470**	77.6 (57.4–109.9)	4.5 (0–16)	7.6 (6–9)

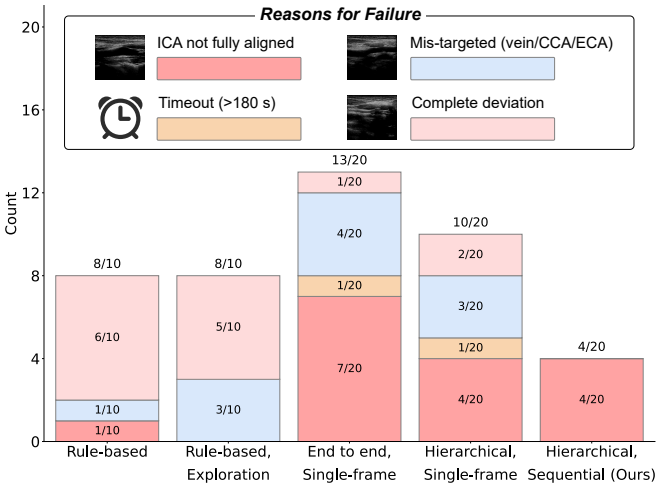

Fig. 5: Failure reasons for different method.

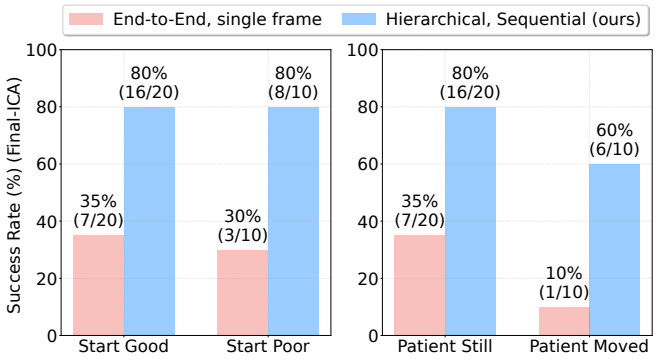

Fig. 6: Robustness evaluation under atypical conditions.

source. The probe is rigidly mounted to the end effector with an ATI Mini 40 force/torque sensor for contact force feedback. Ultrasound video is captured with an Acasis video capture card and streamed to the control computer for real-time inference. Before each experiment, an operator applies acoustic gel to the probe surface and roughly places the probe on the participant's right neck region. The initial positioning requires only approximate placement and does not demand expert precision. For safety, both hardware and software limits are implemented to prevent unintended motion. The operator can trigger an emergency stop at any time to ensure participant safety. Each discrete action output by the system occurs at a frequency of approximately 1.5 Hz, while the cartesian impedance controller runs at 1 kHz.

B. Real-World Experiment

Comparison with Baselines. Five volunteers who were not included in the training set (3 male and 2 female) participated in the real-world evaluation. We compared our model against four baselines. "Rule-based" uses only the Standardized Path Executor without correction. "Rule-based + Explore" augments the Standardized Path Executor with predefined exploratory motions: after 45 routine steps (a population-average estimate to reach the vicinity of the ICA longitudinal view) in Stage 2, the probe emulates the exploratory fine-tuning maneuvers of sonographers by performing 12 directional searches in anatomically reasonable directions, moving two steps per direction, to locate the ICA longitudinal view. "End-to-end, Single-frame [10]" is a monolithic network predicting motion directly from a single frame. "Hierarchical, Single-frame" uses our hierarchical architecture but relies on single-frame input only. All of these baselines use the same stop model as ours.

For "Rule-based" and "Rule-based + Explore", each subject performed two trials per method; other methods were tested in four trials per subject. To ensure fair comparison, the probe was manually positioned at a good initial pose: the z-axis of the probe's coordinate frame was approximately perpendicular to the neck surface, and the common carotid artery was roughly centered in the image. Note that this controlled starting position is only for fairness—Section IV-C later examines robustness under poor initialization.

We evaluated six metrics: (1) Success Rate (Final-ICA) – the probe terminates on a high-quality ICA longitudinal view; (2) Success Rate (Pass-ICA) – the probe passes through a high-quality ICA longitudinal view at any time; (3) Feature Cosine Similarity – for each scan, we computed the cosine similarity between all frames' features and the expert's final ICA view, taking the sequence maximum as its score. These maxima were averaged across sequences for each method. Higher values indicate closer agreement with expert scans. Features were extracted using the encoder of the pre-trained Stop Model; (4) Scan Time – duration of each scan, excluding manually aborted trials after complete carotid drift; (5) Number of Corrections – count of corrective actions per scan, also excluding aborted trials; (6) Subject Comfort Score – self-reported comfort level (0–10) after each scan.

The results are summarized in Tab. I. Our method achieves

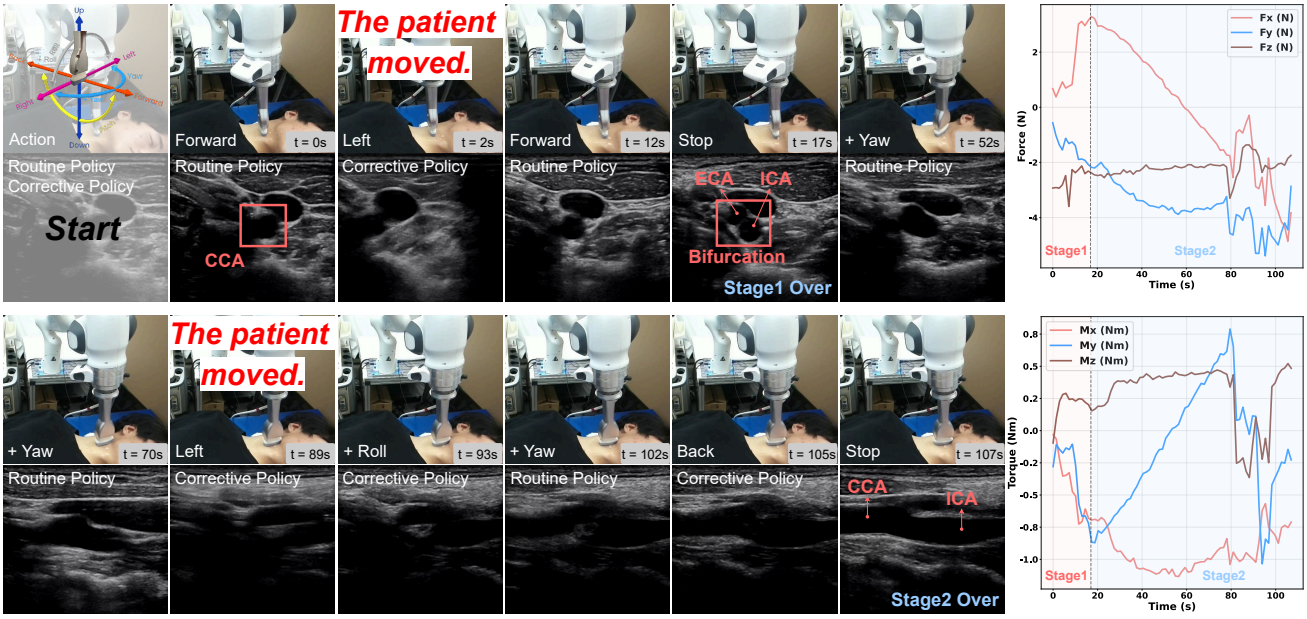

Fig. 7: A representative scanning sequence where the patient moved twice during scan. Our model accurately applied corrective policy to recover and successfully aligned the ICA longitudinal view. The right panels plot the real-time force and torque measurements on the ultrasound probe throughout Stage 1 and Stage 2.

the highest success rates (80% Final-ICA, 95% Pass-ICA) and feature similarity (0.7470) among all baselines, with competitive scan time and moderate corrections. It also received a high comfort rating (mean = 7.6). Despite requiring deep mandibular insertion, our method's precise, impedance-controlled motions minimized discomfort, demonstrating both effectiveness and safety.

Failure Analysis. Failure reasons for each method are summarized in Fig. 5. During the real-world experiments, "Rule-based" and "Rule-based + Explore" performed poorly due to difficulty handling anatomical variability of the ICA, with most failures being Complete deviation requiring manual abort. The "End-to-end, Single-frame" model often predicted imprecise motions near the target, resulting in ICA not fully aligned or Mis-targeted vessels (vein/CCA/ECA). It also frequently entered motion loops—repeatedly outputting opposite actions—causing increased scan time and Timeout (> 180s) failures. The "Hierarchical, Single-frame" model improved corrective prediction but still suffered from action loops and occasional Timeouts. Our method avoided motion loops through sequence input and achieved higher accuracy. Its four failures were all minor ICA not fully aligned cases.

C. Robustness Analysis

In practical robotic ultrasound, challenges such as poor initialization, patient movement, coughing, or external disturbances are common and hinder clinical adoption. We evaluated robustness under two conditions: *poor initialization* and *patient moved during scan*. Tests were conducted on two unseen volunteers, comparing our method with the "End-to-End, single-frame" baseline (five trials per condition). *Start Good*: CCA centered with acceptable quality; *Start Poor*: CCA faint and near the edge with low quality. *Patient Still*: neck stationary; *Patient Moved*: deliberate moderate neck shifts once in Stage 1 and once in Stage 2. As shown in Fig. 6, initial placement quality had negligible impact on our method, achieving similar Final-ICA success under both start conditions. With patient motion, our method maintained 60%success versus the baseline's 10%, demonstrating superior recovery from disturbances.

Fig. 7 illustrates a case where the subject moves twice; our model activates corrective policy and realigns accurately, confirming robust adaptation suitable for clinical use.

D. Offline Evaluation

We reports offline validation results for the three models on both stages in Table II. Compared with the single-frame baseline, our sequence-based approach achieves consistent improvements in Accuracy, Precision, Recall, and F1 for both the Corrective Gate and the Adaptive Corrector, indicating that temporal context contributes to more reliable decision making. The Stop Model uses only single-frame input, as it is sufficient for judgment. Experimental results show that it achieves high accuracy and a balanced precision-recall performance in both stages.

We also presents an ablation study on sequence length $L \in \{3, 5, 7\}$ in Fig. 8. For both stages, longer sequences generally improve accuracy up to a length of 5, after which the benefit saturates or slightly decreases. Balancing computational cost and predictive performance, we select a sequence length of 5 for all real-world experiments.

V. CONCLUSION

In this work, we present UltraHiT, a hierarchical transformer-based architecture that achieves, for the first time, autonomous scanning of the ICA longitudinal section

TABLE II: **Offline validation results.** We report Accuracy, Precision, Recall, and F1 for each model; mPrec., mRec., and mF1 denote macro-averages over all classes. "Single-frame" indicates the single-frame input paradigm.

Method	Corrective Gate				Adaptive Corrector				Stop Model			
	Acc.	Prec.	Rec.	F1	Acc.	mPrec.	mRec.	mF1	Acc.	Prec.	Rec.	F1
Stage 1: Locate the bifurcation of ICA and ECA on the transverse section												
Single-frame	95.07%	69.37%	73.15%	0.7121	87.67%	94.07%	77.36%	0.8299	93.28%	84.94%	92.98%	0.8878
Ours	96.04%	78.05%	79.15%	0.7860	90.22%	93.30%	78.20%	0.8260				
Stage 2: Switch to the longitudinal section and locate the ICA												
Single-frame	84.51%	70.62%	89.47%	0.7893	66.72%	45.64%	63.57%	0.5897	93.20%	84.69%	90.89%	0.8768
Ours	88.07%	76.52%	91.05%	0.8316	75.09%	73.69%	66.15%	0.6621				

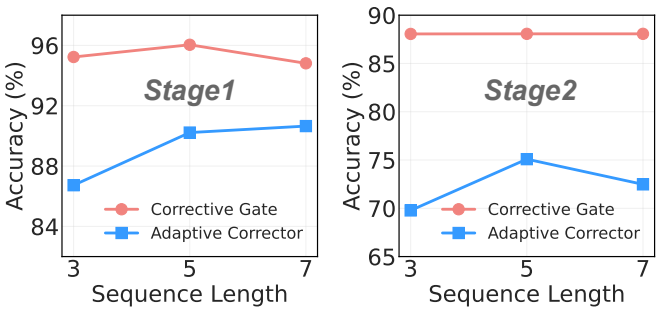

Fig. 8: Ablation study on sequence length.

as required in clinical practice. By integrating a high-level variation assessment module with two specialized low-level executors—a knowledge-based standard executor and a data-driven adaptive corrector—our method effectively handles significant anatomical variations in the ICA. Experimental results demonstrate that UltraHiT achieves a 95% success rate on unseen subjects and shows strong robustness in challenging conditions. This work extends the capability of robotic ultrasound to more complex vascular structures and provides a promising framework for handling anatomical variability in medical robotics.

REFERENCES

[1] Q. Huang, B. Gao, and M. Wang, "Robot-assisted autonomous ultrasound imaging for carotid artery," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–9, 2024.

[2] Z. Wang, Y. Han, B. Zhao, H. Xie, L. Yao, B. Li, M. Q.-H. Meng, and Y. Hu, "Autonomous robotic system for carotid artery ultrasound scanning with visual servo navigation," *IEEE Transactions on Medical Robotics and Bionics*, 2024.

[3] X. Yan, S. Luo, Y. Jiang, M. Yu, C. Chen, S. Zhu, G. Huang, S. Song, and X. Li, "A unified interaction control framework for safe robotic ultrasound scanning with human-intention-aware compliance," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 14 004–14 011.

[4] Y. Huang, W. Xiao, C. Wang, H. Liu, R. Huang, and Z. Sun, "Towards fully autonomous ultrasound scanning robot with imitation learning based on clinical protocols," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3671–3678, 2021.

[5] D. Huang, Y. Bi, N. Navab, and Z. Jiang, "Motion magnification in robotic sonography: Enabling pulsation-aware artery segmentation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 6565–6570.

[6] R. Goel, F. Abhimanyu, K. Patel, J. Galeotti, and H. Choset, "Autonomous ultrasound scanning using bayesian optimization and hybrid force control," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8396–8402.

[7] A. Duan, M. Victorova, J. Zhao, Y. Sun, Y. Zheng, and D. Navarro-Alarcon, "Ultrasound-guided assistive robots for scoliosis assessment with optimization-based control and variable impedance," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8106–8113, 2022.

[8] X. Yan, Y. Jiang, G. Wu, C. Chen, G. Huang, and X. Li, "Multi-modal interaction control of ultrasound scanning robots with safe human guidance and contact recovery," *arXiv preprint arXiv:2302.05685*, 2023.

[9] H. Gray, *Anatomy of the human body*. Lea & Febiger, 1878, vol. 8.

[10] H. Jiang, A. Zhao, Q. Yang, X. Yan, T. Wang, Y. Wang, N. Jia, J. Wang, G. Wu, Y. Yue *et al.*, "Towards expert-level autonomous carotid ultrasonography with large-scale learning-based robotic system," *Nature Communications*, vol. 16, no. 1, p. 7893, 2025.

[11] X. Deng, Y. Chen, F. Chen, and M. Li, "Learning robotic ultrasound scanning skills via human demonstrations and guided explorations," in *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2021, pp. 372–378.

[12] R. Droste, L. Drukker, A. T. Papageorghiou, and J. A. Noble, "Automatic probe movement guidance for freehand obstetric ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 583–592.

[13] Y. Bi, Y. Su, N. Navab, and Z. Jiang, "Gaze-guided robotic vascular ultrasound leveraging human intention estimation," *IEEE Robotics and Automation Letters*, 2025.

[14] H. Jiang, Z. Sun, N. Jia, M. Li, Y. Sun, S. Luo, S. Song, and G. Huang, "Cardiac copilot: Automatic probe guidance for echocardiography with world model," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 190–199.

[15] Y. Bi, Z. Jiang, Y. Gao, T. Wendler, A. Karlas, and N. Navab, "Vesnet-rl: Simulation-based reinforcement learning for real-world us probe navigation," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6638–6645, 2022.

[16] K. Li, J. Wang, Y. Xu, H. Qin, D. Liu, L. Liu, and M. Q.-H. Meng, "Autonomous navigation of an ultrasound probe towards standard scan planes with deep reinforcement learning," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 8302–8308.

[17] K. Li, A. Li, Y. Xu, H. Xiong, and M. Q.-H. Meng, "Rl-tee: Autonomous probe guidance for transesophageal echocardiography based on attention-augmented deep reinforcement learning," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 2, pp. 1526–1538, 2023.

[18] H. Jiang, Z. Sun, Y. Sun, N. Jia, M. Li, S. Luo, S. Song, and G. Huang, "Sequence-aware pre-training for echocardiography probe guidance," *arXiv preprint arXiv:2408.15026*, 2024.

[19] Y. Bi, C. Qian, Z. Zhang, N. Navab, and Z. Jiang, "Autonomous path planning for intercostal robotic ultrasound imaging using reinforcement learning," *arXiv preprint arXiv:2404.09927*, 2024.

[20] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," *Advances in neural information processing systems*, vol. 34, pp. 15 084–15 097, 2021.

[21] A. Albu-Schaffer and G. Hirzinger, "Cartesian impedance control techniques for torque controlled light-weight robots," in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, vol. 1. IEEE, 2002, pp. 657–663.