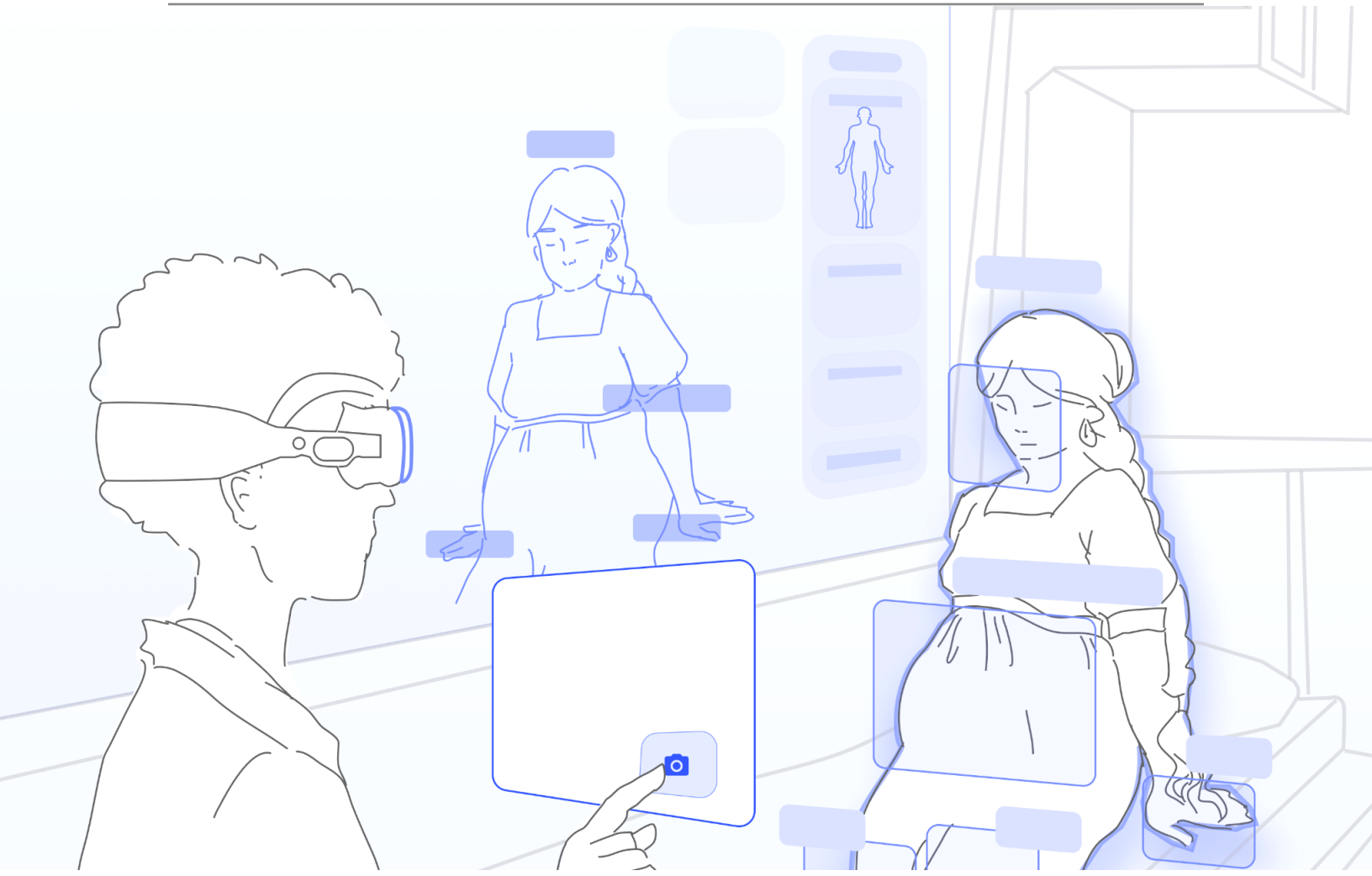
ARPA-H PARADIGM Program Solicitation
ARPA-H-SOL-24-02

VOLUME 1: TECHNICAL AND MANAGEMENT

FIELD GUIDER

CLINICAL GUIDANCE FOR MOBILE HEALTHCARE WORKERS

ARPAHPARADIGMDESIGN.COM



A goinvo COLLABORATION WITH: mediate THE FAMILY VAN BOSTON 1992 Boston Children's Hospital

Solicitation #	ARPA-H-SOL-24-02
Proposal Title	Field Guider: Real-time clinical guidance for mobile healthcare workers
Technical Area	TA5
Proposer Organization	We Create Goodness, LLC dba GoInvo
Type of Organization	Small Business
Proposer's Internal Reference Number, if any	
Technical Point of Contact (POC)	Name: Juhan Sonin Mailing Address: 661 Massachusetts Ave., Third Floor, Arlington, MA 02476 Telephone: 617-803-7043 Email: juhan@goinvo.com
Administrative POC	Name: Jonathan Follett Mailing Address: 661 Massachusetts Ave., Arlington, MA 02476 Telephone: 617-803-7043 Email: jon@goinvo.com
Award Instrument Requested	Other Transaction Agreement
Total Proposed Cost	Total: $12,564,722.50
Place(s) of Performance	We Create Goodness, LLC dba GoInvo office locations, Virtual Collaboration Research, Inc. (Mediate) office locations, Boston Children's Hospital, and Harvard Medical School
Other Team Members (subawardees and consultants) if any	Technical POC Name: Dr. Cagri Hakan Zaman Organization: Virtual Collaboration Research, Inc. (Mediate) Organization Type: Small Business Technical POC Name: John Brownstein Organization: Boston Children's Hospital Organization Type: Large Business Technical POC Name: Mollie Williams Organization: Harvard Medical School Organization Type: Educational
Date Proposal was Prepared	April 25, 2024
Proposal Validity Period (minimum 120 days)	120 days

Table of Contents

PROPOSAL SUMMARY

A. DISCUSSION

Enhancing Healthcare Accessibility through a Virtual Medical Guide

The aim of this project is to enhance healthcare accessibility for people in rural or remote regions by minimizing the skill gap among health workers, thereby empowering them to deliver high-quality care effectively. Field Guider is an Augmented Reality (AR) Guidance system that assists real-time communication, spatialized instruction, and expert decision making support for medical staff.

Specifically, the Field Guider system offers:

- Seamless communication between patients and the Care Delivery Platform (CDP) for scheduling and confirming with mobile health units.
- Patient check-in automation to alleviate the administrative workload on staff, onsite or virtual (extending to scheduling and pre-appointment interactions).
- Guidance for healthcare professionals during patient consultations to ensure more effective, safer examinations, diagnoses, and treatments.
- Secure health data collection for continuous service improvement and scaling of healthcare services.

Currently, healthcare delivery in remote areas often suffers from a lack of specialized medical staff and resources, leading to delayed diagnoses and treatment.

Traditional methods rely heavily on periodic visits by mobile health units or require patients to travel significant distances to access medical facilities. This approach is limited by logistical challenges, inconsistent patient follow-ups, and often inadequate data management that can affect the quality of care.

The innovation in this project lies in integrating a sophisticated virtual medical agent that supports seamless communication and automated patient management tasks.

This agent is equipped with advanced language processing capabilities for intuitive dialogues and can perform real-time multimodal data processing, including medical imaging and patient monitoring through sensors. This system not only improves the efficiency of healthcare delivery but also ensures continuous patient engagement and monitoring.

Success in the Field Guider project can lead to broader adoption of digital health solutions in underserved areas, significantly improving health outcomes and healthcare efficiency. Residents of rural and remote areas will be the primary beneficiaries, as they will gain improved access to quality healthcare. Healthcare providers will also benefit from enhanced decision-support tools and reduced administrative burdens.

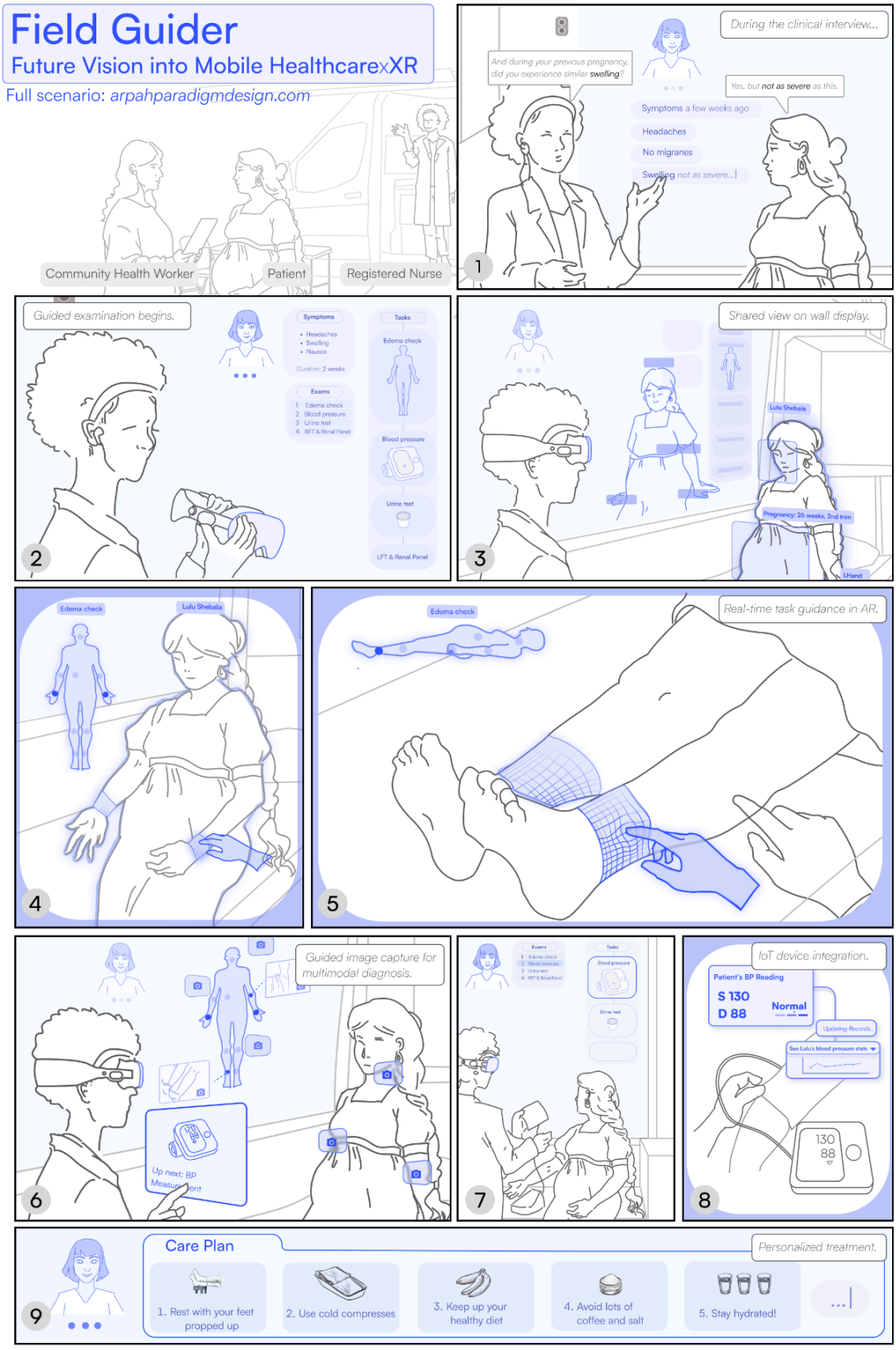


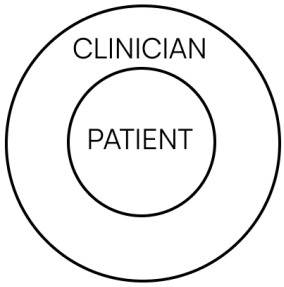
Figure 1. Spatialized Guidance System Design and UX.
For complete experience scenario and system visit: arpahparadigmdesign.com

Technical Challenges

The key technical challenges for Field Guider include:

- **Real-time multimodal data integration:** Ensuring accurate and timely processing of diverse data types, from audio inputs to medical imaging. To address this challenge, we will develop a Vector/Augmentation Database to store and access real-time data efficiently.
- **Contextual awareness and task execution:** The system must understand and adapt to dynamic clinical environments and patient conditions. To address this, for 3D object detection using RGB images, we plan to implement the Multi-View CNN (MVCNN) approach, which enhances the detection and recognition of three-dimensional shapes by integrating features extracted from multiple viewpoints of an object. We will also employ state-of-the-art action recognition models for real-time action recognition.
- **Data security and privacy:** Maintaining the confidentiality and integrity of sensitive health data. To address this, we will employ robust encryption methods for data security.
- **Bandwidth limitations:** The Field Guider system will employ edge computing for almost all tasks — including an LLM inference engine that can be activated when cloud is not accessible. It will also utilize an end-to-end system which will allow it to gather all necessary information about the patient before the session, so any online data fetching can be completed upfront.

Our Design Strategy



The use case in Figure 1 illustrates an initial concept for the Field Guider spatialized guidance system design and UX. Designing a system that is an integral part of a broader vision for healthcare is a central element of our approach to Field Guider. We designed the concept for the guidance system following a Clinician-front-and-center (Patient-in-the-epicenter) strategy. The design of the UIs, UX workflows, and systematic testing with medical professionals — the system's end users — the evaluation of the design and technological elements based on realistic metrics, and the incorporation of feedback for revising and optimizing the designs are at the core of our strategy.

Task Guidance

With Field Guider, each medical experience becomes a hands-on educational opportunity for the trainee medical service provider through the visual and auditory, high-resolution guidance provided by the system.

Cardiac Ultrasound

We've identified cardiac ultrasound as the initial clinical procedure for task guidance for the Field Guider project. Table 1 provides an example of skills we have identified for guidance and

how they correlate with different tasks involved in a cardiac ultrasound.

Defining Clinical Services and Tasks

A clinical service may encompass various medical procedures each requiring specific operational steps. Tasks within a clinical service are defined as unique skilled actions necessary for the execution of the service. For instance, in an ultrasound examination, tasks may include adjusting the patient's position, applying gel to the ultrasound probe, placing the probe on the patient, optimizing probe positioning for more precise imaging, and capturing the final image. Performing an ultrasound, including a cardiac ultrasound, requires various specialized skills. These skills involve the technical ability to operate ultrasound equipment and understand anatomy, patient care, and image interpretation. The manner in which these tasks are executed varies significantly based on the specific requirements of each clinical service. The techniques for positioning a patient and handling the probe during a cardiac ultrasound are specifically tailored to obtain necessary cardiac views and images, which differ markedly from the tasks used in a knee ultrasound.

Table 1. Example tasks in Cardiac Ultrasound

Prototype	**Tasks/Prototype**	**Proposed Skill**
Y1, Experience with Spatial Guidance	1. Position Patient	Patient Preparation
	2. Ultrasound Setup (settings for Cardiac View 1)	Equipment Preparation
	3. Gel Application	Equipment Operation
	4. Probe Placement on the Patient’s Body (Cardiac View 1)	Anatomical Knowledge
Y2, User-driven prototype integrated in mock-CDP	5. Ultrasound Setup (settings for View 2)	Equipment Preparation
	6. Probe Placement on the Patient’s Body (Cardiac View 2)	Anatomical Knowledge
Y2.5, User-driven prototype, conversational agent implemented	7. Probe Adjustment for Image Acquisition (Cardiac View 1)	Sonographic Techniques Knowledge
	8. Probe Adjustment for Image Acquisition (Cardiac View 2)	Sonographic Techniques Knowledge
Y3, Device-driven prototype integrated in mock-CDP, diagnostic module implemented	9. Diagnostic Assessment	Image Capture, Image Evaluation
	10. Explain Procedure(s) to Patient	Patient Communication
Y3.5, Full CDP integration of Experience	1-10 for other Ultrasound, or other Clinical Service	New Equipment, Anatomical, and Technical Knowledge
Y4, Open-source APIs, Databases	(same as prev.)	(same as prev.)

B. Innovative Claims Table

Our proposed work is unique in its comprehensive approach to integrating advanced AI with healthcare delivery, specifically tailored for remote settings.

Announcement pg. para(s)	Technical Challenge	Innovation and evidence	proposal section(s), pg., para(s).
ARPA-H-SOL-24-02 P. 25, PARA 2	Bridging the skills gap in real-time while training the next generation of medical staff.	**Health workers just-in-time upskilling** Deploying AI-driven tutorials and decision support tools that adapt to the clinician's current task and knowledge level.	3.C.1.2. Develop Decision Making and Action Recommendation System p.20
ARPA-H-SOL-24-02 P. 26, PARA 3	Capturing and replicating medical knowledge and human expertise.	**Human-AI collaboration for dynamic clinical decision support (Dx & Tx)** Capture and analyze real-time data of the patient for diagnostics and treatment, enhancing human decision-making with AI-driven oversight and guidance.	3.G. Clinical Decision Support (Dx) Module p.26
ARPA-H-SOL-24-02 P. 26, PARA 1	Coordinating and synthesizing diverse data inputs (audio-voice, video, and other environment and medical sensors) to design guided healthcare experiences.	**Multimodal, spatialized Guidance: natural language instructions combined with visual, spatialized task representations** Integration of Augmented Reality visual cues with natural language instructions to deliver precise, real-time, and context-sensitive guidance to clinical staff of varying training levels and experience during medical procedures.	3.A. Environment Model (EV), Van's Digital Twin p.12
ARPA-H-SOL-24-02 P. 27, PARA 1	Experience design integrating patient and medical staff experience to provide efficient point-of-care guidance and continued care.	**End-to-end Clinical Guidance Service Design with secure data integration and real-time patient monitoring** Virtual and Augmented Reality interface design that simulate real-world scenarios for training and real-time guidance during clinician-patient interactions.	2.A. Guidance Experience Design p.9

Goals and Impact

Health Workers Upskilling with Real-time Guidance

Field Guider aims to enhance healthcare accessibility in rural and remote regions by bridging the skill gap among health workers and empowering them to deliver high-quality care effectively.

The key innovative aspects of this project include:

1. Integration of a sophisticated virtual medical agent that supports seamless communication and automated patient management tasks. This agent is equipped with advanced language processing capabilities for intuitive dialogues and can perform real-time multimodal data processing, including medical imaging and patient monitoring through sensors.
2. Facilitating real-time communication and decision-making support for patients and medical staff, encompassing everything from initial consultation to ongoing treatment and health management.
3. Improving the efficiency of healthcare delivery while ensuring continuous patient engagement and monitoring.

The specific outcomes of the proposed research include:

- Assisting and empowering community healthcare workers, EMTs, and RNs to perform specialized medical procedures that they may not have been able to do previously. This expands access to care in underserved areas.
- Reducing the learning curve for performing examinations, which can lead to more uniform and reliable diagnostic results across a variety of clinical operators. This can improve patient outcomes.
- The use of augmented reality, AI, and real-time guidance has great potential to increase the efficiency, accuracy, and safety of medical interventions, especially in emergency situations. This can be a valuable tool for healthcare providers.

If successful, the project will make a significant difference in the lives of residents in rural and remote areas by providing them with improved access to quality healthcare. Healthcare providers will also benefit from enhanced decision-support tools and reduced administrative burdens. The Field Guider intelligent task guidance system can improve access to specialized medical care, enhance patient outcomes, and transform the delivery of healthcare, especially in underserved communities.

An Iterative Approach to Design and Development for Field Guider

The development process for the Field Guider project encompasses a series of research, design and testing cycles across five iterative prototypes, specifying the skills associated with each clinical task, identifying experiential pain points, and exploring guidance possibilities both in task models and in situ.

Each iteration is aimed at refining system characteristics and improving user experience. Usability and proficiency evaluations will be conducted through Likert scale assessments and competency targets, ensuring that the final product effectively enhances the precision and effectiveness of clinical task execution in the real-world CDP van setting.

1. Observational Studies and Data Collection

Research data on how clinicians perform medical procedures and tasks are essential to innovate with conversational and spatialized (AR) guides in clinical service guidance.

We will conduct up to five rounds of data collection with different focuses depending on the phase and needs of the system development. A series of studies with clinicians and medical staff in hospitals or mock-up environments will allow a deeper understanding of the problems and the pain points in medical procedures that can be supported with intelligent virtual guidance, to enhance the workflows without altering the medical routines.

1.A Ethnographic Research

We will systematically research and incorporate evidence-based practices and clinical guidelines into the design and functionality of our system to make sure it is grounded in real-world clinical insights and patient needs.

1.A.1 Clinical requirements for clinical service(s) (CS1+ 2)

The initial phase of defining clinical requirements for two clinical services involves an exploratory workplace study within the hospital setting. The goal is to observe clinicians performing the procedures under scrutiny, and identify their sequences of actions, the tools and technologies they use, and their interactions with patients during specific tasks. This step will document our observations of clinical workflows and interactions through synthetic encounters, by taking photos and videos. This preliminary research is essential for establishing the project's foundational concepts and our technical and design objectives. It will also allow us to scale the system’s capabilities in the future. By accounting for general processes and requirements in the design principles, the system will be adaptable to multiple clinical procedures and tasks.

As a next step, the design and medical (BCH) teams will assess the observations to identify specific clinical skills, knowledge gaps, and scenarios where technology can provide significant

enhancements. These assessments will cover a broad spectrum, including user needs, data usage, operational sites, and device interactions. We will thus make sure the clinical services developed are practically applicable and relevant to the needs of both clinicians and patients.

1.A.2. Stakeholder Interviews

Consulting stakeholders who know the challenges of healthcare in underserved areas, including clinicians in rural settings and mobile clinics (clinics-on-wheels) is essential for understanding the specific needs of these environments. This task aims to capture the challenges outside conventional urban hospitals and clinics through a structured series of remote and on-location interviews and targeted surveys. The interviews will focus on specific circumstances of rural and isolated areas, such as staffing constraints, training necessities, access to medical resources, frequency of common medical incidents, and other issues that affect healthcare delivery in these regions.

1.B. Clinical Studies

A series of studies, each linked with the development of each deliverable software prototype, will be held to collect data for analysis and use in the design and engineering of the medical guidance system.

1.B.1. Environment Requirements for R&D Mock-up

This task involves a detailed recording of the spatial and functional relationships within the clinical space to facilitate the construction of a full-scale R&D mock-up inside GoInvo's studio space to be used in the system development in Phase II. The design and engineering teams will map out the area and document interactions related to the placement, usage, and storage of furniture, objects, materials, and devices associated with the clinical services and specific tasks under study. Based on these specifications, a design plan will be formulated to build the in-house environment mock-up. This simulated environment will provide a foundation for further development and testing of the prototypes.

1.B.2. Clinical Procedure Modeling

Targeted **documentation** of synthetic encounters using camera and audio inputs within the equipment and R&D setup is crucial for capturing detailed visual and auditory data of clinical tasks. (Camera equipment is described in Technical Objectives, A.1). The process involves precisely defining the start and end points of actions within various types of clinical encounters, identifying key movements or gestures made by healthcare professionals during procedures, and noting any relevant objects or tools that are manipulated. The engineering team will create detailed **3D annotations and representations** of clinical actions, visual and textual. Procedure modeling will be repeated multiple times throughout the development process as required, with each iteration becoming increasingly focused and targeted in the guidance process. This **iterative refinement** approach will help develop refined training modules and simulation systems that accurately capture and replicate real-world clinical scenarios in 3D.

1.B.3 Spatial Task Guidance Database

Following the comprehensive procedure modeling, the Spatial Tasks Guidance Database will be designed to store the refined data from the modeling phase, including:

- **Narratives**: descriptions and step-by-step guides for each task providing context and procedural insights,
- **Visualizations:** static and dynamic 3D visualizations that reflect the precise actions and techniques involved in each task, enhancing task understanding and training,
- **Task Constraints:** as specific requirements, limitations, and standards associated with each task, ensuring that the database supports a wide range of clinical scenarios and medical procedures. See also Technical Objectives, 3.F.1.

2. System Design & Evaluation

Design and evaluation cycles are closely linked with the development of prototype iterations. It is anticipated that each prototype undergoes a design and testing phase, which is repeated once to further develop the prototype.

2.A. Guidance Experience Design

2.A.1. Design Concepts

We will explore and develop initial design concepts for the guidance system's augmented reality (AR) components, including:

2.A.1.1. AR Visualizations

To enhance healthcare workers' understanding of clinical tasks and their seemless interaction with the system, we will create visually engaging and informative 3D models and virtual objects that can be overlaid on the real-world environment. We plan to ensure that the AR visualizations are intuitive, easy to interpret, and provide valuable information to healthcare workers. Additionally, we intend to experiment with different rendering styles, lighting, and animation to make the visualizations visually appealing and impactful.

2.A.1.2. Floating Windows

We will design interactive 2D windows or panels that will be positioned within the AR virtual 3D space. These floating windows should be easy to access, resize, and manipulate, allowing healthcare workers to customize the layout and information displayed. Our intention is to explore different ways to seamlessly integrate the floating windows with the surrounding AR elements, creating a smooth immersive experience.

2.A.1.3. Shared Display

In our UX scenario and proposal, we introduce and will further develop the concept for shared display where multiple users (clinician and patient) can view and interact with the same virtual

content simultaneously. We will consider ways to facilitate collaboration and communication between the clinician — wearing or not the AR headset — the patient, and the virtual AI agent. Will achieve this through shared visualizations, annotations, pointers, and other custom visual assets. Additionally, we will explore methods to ensure a real-time synchronization of the displays for a consistent user experience across multiple devices or platforms. We plan to utilize this shared visualization for real-time auditing and fact checking of the AI algorithm outputs and decision-making by the clinician.

After developing the initial design concepts, we'll refine and iterate on them. This may involve exploring alternative approaches, layouts, and interactions to find the most effective and engaging solutions.

2.A.2. Preliminary UX Workflows / Interaction Design

We'll develop the user experience (UX) workflows and interaction design for the AR components. This involves mapping out the healthcare worker's journey through various tasks and scenarios they may encounter within the hybrid physical/van and virtual environment. We plan to define essential interactions and gestures that users will use to navigate, manipulate, and interact both with AR and physical elements, such as medical devices and other task-relevant objects.

As part of this task, we will design intuitive interaction patterns, such as how users can select, move, or resize floating windows and AR visualizations. We will also consider the ergonomics and physical constraints of the hybrid environment to ensure our interaction design is user-friendly. Finally, we aim to explore methods for providing clear feedback and guidance to users, ensuring a seamless and intuitive experience throughout their interaction with the AR components.

2.A.3. UI Mockups for 2D and 3D AR Components

For this task, based on the initial design concepts and interaction design, we'll create detailed UI mockups for both the 2D and 3D AR components.

2.A.3.1. 2D UI Mockups

We will produce detailed visual representations of all UI elements, including the precise placement of controls, information displays, and interactive elements, designed according to medical workflow standards. These mockups also define the visual style, typography, and color scheme to ensure the user interface is visually appealing and easy to integrate, to navigate without prior knowledge, and to follow instructions in high-intensity clinical settings. The mockups will also show how the 2D UI elements will visually integrate with the surroundings and augmented reality (AR) environment.

2.A.3.2. 3D AR Mockups

We will design 3D mockups for AR that are visually compelling and informative. We will determine the appropriate scale, positioning, and orientation of the 3D AR elements within the hybrid environment. This is a particularly demanding task with multiple usability and efficiency constraints related to the clinician's experience, the nature of each clinical task, and the

limitations of the van space. We will employ various design methods to make the AR components responsive to diverse user scenarios to provide engaging interactions and guidance.

2.B. Prototype Testing and Evaluation

2.B.1. Usability Testing (Prototypes 1-3)

For each of the four initial Prototypes, will conduct usability testing specifically tailored to validate the design concepts of the clinical task guidance system, to ensure it meets the task performance requirements in the mobile clinic setting. We will initially recruit groups of 5-10 participants from clinical and non-clinical backgrounds to provide a broad range of insights on both the system's usability as well as the task descriptions. We will design test scenarios that enable these participants to interact with the AR components and the medical devices and provide detailed feedback. Throughout the testing process, we will observe and document the users' interactions, noting any pain points and gathering their overall impressions of the AR experience. We will collect quantitative data, such as task completion duration rates and user satisfaction scores, and qualitative feedback to understand the user experience. The prototype will be tested for metrics as full encounter duration, individual task duration, and user response time per instruction to compare between clinicians vs. non-clinicians. We will also test different versions of the designs and workflows, and gather feedback in the form of semi-structured interviews or surveys. The results will show areas for improvement and informing the subsequent iterations of the design to enhance its effectiveness in clinical applications and for the purposes of medical skill learning. Below is a sample interaction scenario to test the second, MVP Prototype in the cardiac ultrasound case.

Clinician and Patient Interaction: Clinician follows instructions to starts Guidance system (Task 1), puts on AR headset (Task 2), and to apply gel on probe (Task 3), place it on Patient (approximate positioning) (Task 4) and views image on monitor and AR UI. Spatial Procedure Guidance Database v0.1.

2.B.2. Clinical Review and Validation

Clinical validation will ensure the safety, efficacy, and reliability of the UX for the guidance system. We will collaborate with the BCH clinical team to define the clinical validation requirements and protocols, ensuring all aspects meet necessary standards. By designing and conducting these studies, we aim to evaluate the performance and accuracy of the visualizations and interactions for their medical and scientific accuracy, as well as for their potential to integrate into existing clinical workflows. We will collect and analyze the outcomes of the guided tasks for any potential adverse effects or safety concerns. The clinical validation process will be iterative, involving multiple rounds of testing and refinement of the representations of each new task and clinical service.

2.B.3. Usability Studies (Prototypes 4-6)

To optimize the user experience of the intelligent task guidance system for healthcare staff inside the van, we will develop a comprehensive usability testing protocol to run in the moch-CDP environment. Similar to the previous studies, we will recruit a diverse group of users with varying levels of technological proficiency and clinical experience to participate in the usability

testing. We will observe and collect data on the participants' interactions, errors, and feedback as they perform different tasks. To assess usability we will employ quantitative and qualitative methods and metrics, such as task completion rates, error rates, time-on-task, and user satisfaction surveys. We will taylor each study to the new features incorporated in each prototype. For example, the study to assess Prototype 4 will focus on verbal communication of instructions alongside spatialized guides, whereas the Prototype 5 study will focus primarily on system's decision-making, correctness, and communication of medical diagnoses.

2.B.4. Effectiveness Study

We will develop a study protocol to evaluate the impact of our system on the quality of ultrasound examinations conducted by registered nurses and community healthcare workers. This involves recruiting a diverse group of participants, representing a variety of experiences, technological proficiencies, and geographical locations, to participate in the study. As a first step, we will run a pilot study with groups of approx. 5 participants per group that will allow us to refine our testing experiences and study objectives, and limitations. Participants will use the intelligent task guidance system to train on different tasks and we will monitor their learning curve and their proficiency in performing ultrasounds with and without the system's assistance. Field trials will be conducted in real-world settings, such as rural healthcare facilities, to gauge the system's effectiveness in supporting healthcare workers during clinical service performance - eg. ultrasound examinations. For the ultrasound scenario, we will gather data on the accuracy, consistency, and efficiency of the ultrasounds performed under the system's guidance, alongside patient outcomes and satisfaction. This data will be analyzed to determine the system's efficacy in healthworker upskilling, enhancing patient care, minimizing errors, which will help pinpoint areas for future improvement.

3. Technical Objectives

Our system comprises several interconnected systems that work together to provide clinical guidance inside the care delivery platform (CDP) vehicle, that we refer to as the 'van'. Below we describe in detail the components of the system.

3.A. Environment Model (EV), Van's Digital Twin

Enhancing the performance of clinical routines by understanding real-time interactions and providing spatialized, real-time guidance is central to our approach. The Environment Model is necessary for real-time 3D localization and identification of individuals, healthcare workers and patients, in the van by capturing their interactions with objects and devices in the environment and precisely annotating anatomical parts on their bodies. To provide high precision and throughput from a range of sensors and devices, we introduce a 3D digital twin, a densely annotated 3D environment model of the van to synchronize sensory information from cameras, microphones, and environmental sensors with objects and medical devices of interest. The model serves as a reference for identifying events and interactions among people, devices and other objects. The model will also work as a reference for placing precomputed guidance information precisely located over medical devices and other relevant objects within the van to provide training and real-time guidance to the van staff. Identifying patients within the spatial model will

also enable 3D localization of just-in-time information over the patient's body, providing precise visual and aural cues for clinical task guidance.

3.A.1 Setup Hardware Components

3.A.1.1.Setup and Calibrate Cameras

First we will setup a camera array. We will employ 8 high-frequency cameras, calibrated and time-synchronized, to capture the van interior. We will experiment with several camera alternatives, each with different advantages and disadvantages. Motion capture cameras (Optitrack Prime Color), have very high frequency (250fps) and decent spatial resolution (1080p) and provide automatic temporal synchronization. The downside of this system is that it only captures RGB channels and the localization system will rely on Multiview Stereo (MVS) algorithms. The temporal resolution will help denoising the potential errors caused by these algorithms. Alternatively, we will employ depth cameras (i.e. Intel Realsense D455) to provide measured depth (RGB-D) for better 3D localization. However, they have significantly less temporal resolution (~30fps), and less decent color resolution than motion capture cameras. Some alternatives, which use stereo cameras, provide hardware solutions for automatic calibration and stereo fusion and are capable of industrial-level applications. We will evaluate the best-performing setup and integrate the camera array into our system.

We will also use focused RGB-D Cameras. The additional set of 2 RGB-D cameras will be strategically placed within the van, enabling us to obtain a higher-resolution image of the patient's body for accurate surface reconstruction.

3.A.1.2. Setup and Calibrate Microphone and Speaker Array

An array of microphones (standalone or integrated into the camera array) as well as focused-beam speakers, will be placed to capture sound and produce spatialized audio.

3.A.2. Software Components

3.A.2.1 3D Digital Twin Model

A precise 3D CAD model of the van will be obtained (in-house R&D space, then CDP mockup environment) with geometric annotations of medical devices, tools, and medical devices. This model will be calibrated and aligned with a Simultaneous Localization and Mapping (SLAM) model, obtained from a 3D scan of the van. By integrating the annotated model with a SLAM, any future localization within the system will share the same spatial reference, and additional sensors such as the AR devices and mobile cameras (for future use cases) will be localized within the van. As the camera array produces localized representations of people and objects, they will be cross referenced with the annotated model for precise guidance and spatial reasoning algorithms.

3.A.2.2 Human Pose Estimation with Annotated Mesh Reconstruction

The camera array will be used to obtain an accurate representation of people within the van, through 3D Pose estimation, which is a structural model of joints projected into 3D space, as well as a parametric reconstruction of the body in the form of a surface mesh. Obtaining localized 3D human posture and mesh is well documented in the literature, such as [1,2,3,4,5,6].

The performance of a system with pre-calibrated, high-frequency cameras is significantly better than in-the-wild implementations. Therefore, the controlled environment of the van will allow obtaining satisfactory results for this task.

As an improvement over the state-of-the-art, we introduce a parametric anatomical annotation of the human body, specifically for clinical guidance. While the current approaches in human pose and mesh reconstruction are sufficient for recognizing a person's overall orientation and attention, clinical guidance often requires precise localization of anatomical parts, e.g., when performing cardiac ultrasound procedures (ref to description in proposal?). Mapping the medical guidance information on the body without appropriate annotation thus posits challenges. We will overcome this challenge by parameterizing the mesh reconstruction process, and reconstructing an anatomical reference model with semantic part annotations, similar to [7]. This approach reconstructs a hand model with bone and muscular structures. While annotation at the muscular level is desirable, it might prove difficult to fit a similar model at the scale of the full human body. The alternative will be a surface level segmentation, which will be sufficient for clinical guidance for most of the procedures. See also [8, 9] and dataset in [10, 11].

3.A.2.3 Person Identification with one-shot learning

To disambiguate localized 3D human poses, we need to uniquely identify each individual within the van. The most robust approach for this task is to use facial recognition, which will allow us to use only a single image of the patient and the CDC members, and does not require storing any private data after the session. While advanced face recognition algorithms - such as in [12,13,14, 15] - can identify thousands of individuals in crowded images, our application does not require such a complex system. The simplicity of our use case also allows us to sidestep potential privacy and security concerns.

3.A.2.4 Object Detection

3.A.2.4.1 Implement Pre-Trained Object Detection Models

To achieve real-time object detection in a clinical environment using a pre-calibrated multi-camera RGB setup, we will employ a streamlined approach focused on leveraging advanced object detection algorithms. Utilizing frameworks such as YOLO (You Only Look Once) [13] or SSD (Single Shot MultiBox Detector), which are designed for high-speed and accurate object detection, the system will process images from multiple cameras simultaneously. These algorithms are capable of detecting objects in real-time by evaluating each frame rapidly to locate and classify various medical tools and devices visible in the camera feeds.

3.A.2.4.2 Develop 3D Object Detection

In our proposal for 3D object detection using RGB images, we plan to implement the Multi-View CNN (MVCNN) approach, which enhances the detection and recognition of three-dimensional shapes by integrating features extracted from multiple viewpoints of an object [4, 5]. This method effectively aggregates 2D information from various angles to construct a comprehensive 3D representation, enabling precise object localization and identification within the clinical environment. By utilizing MVCNN, we aim to leverage the spatial consistency and additional context provided by multiple images, thus improving the overall accuracy and reliability of our object detection system in real-time applications.

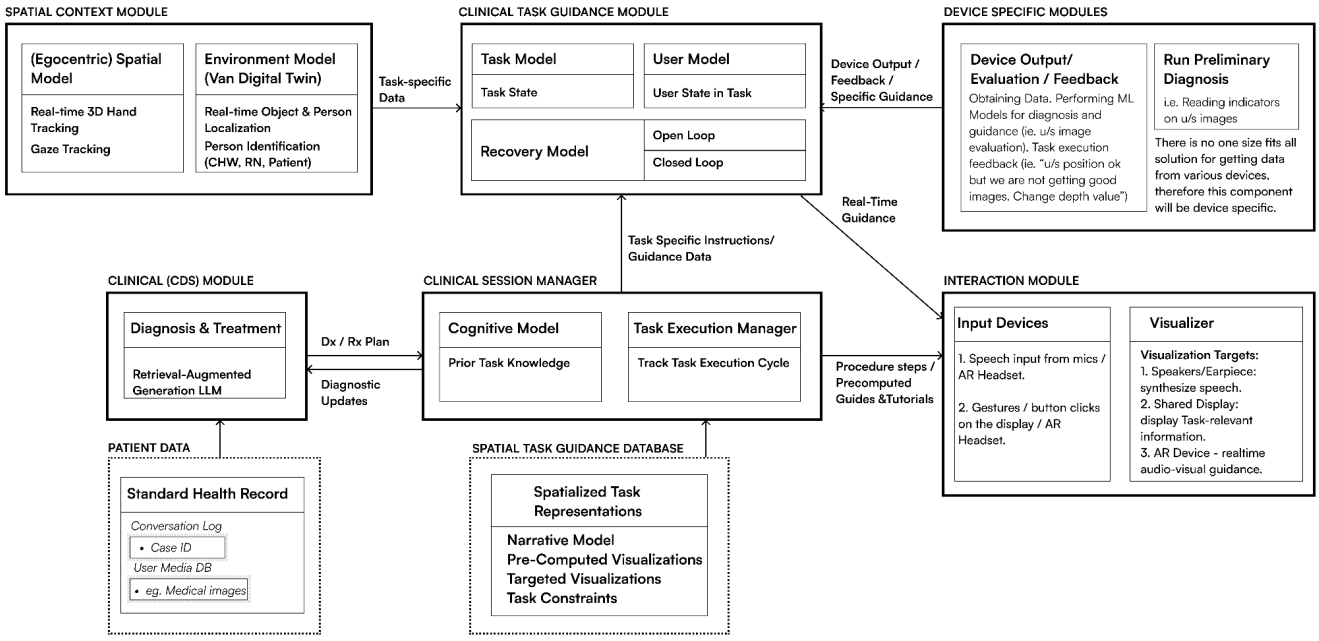


Figure 2. Field Guider system diagram.

3.A.2.5.Action Recognition

Another critical function of the environment model will be recognizing a set of actions from the live camera stream. State-of-the-art action recognition models outlined in [16] provide satisfactory results for real-time action recognition for most common actions, such as reading, writing, picking up, etc. We will employ a skeleton based action recognition model because the environment model already performs posture recognition, which is used in these methods such as [17, 18]. However, in the context of clinical guidance, “action” has a broader meaning based on a particular context. For example, a cardiac ultrasound involves complex actions such as positioning the patient, picking up the probe, placing the probe in contact with the skin, moving the probe over the skin, and so on. These are complemented with actions recognized from devices such as capturing an ultrasound image. Therefore, in our approach, we develop a procedural action recognition system, fine tuning and expanding on the state-of-the-art action recognition models. A mesh based model will complement the skeleton based model for performing surface-contact recognition. We will employ strategically placed RGB-D cameras (different from the camera array), for close up mesh reconstruction of the body parts and objects (see also section 3.A.1.2.). Action recognition will be task-driven and contextualized within the scope of an ongoing task. We will filter out unnecessary information to be processed by other parts of the guidance system, which will rely on the real-time data from the environment model.

3.A.2.5.1 Task-Specific Action Descriptors

3.A.2.5.1.1 Comprehensive Task Analysis

The initial phase of our action recognition system development begins with a comprehensive analysis of the clinical tasks that need to be accurately recognized and monitored. For this

project, we have selected two clinical case studies to demonstrate guidance: **Cardiac Ultrasound (Cardiac U/S)** and **Electrocardiogram (EKG)**.

In Table 2, we show how each procedure is broken down into distinct, manageable components, and the required type of action recognition. This table serves as a foundation for developing and fine-tuning our action recognition algorithms specific to each task within these procedures. The distinctions between skeletal action recognition, surface contact, and device/tool interaction are crucial for accurately interpreting actions and assisting in clinical environments.

Table 2. Components and Action Recognition Type

Procedure	Action	Description	Recognition Type
Cardiac U/S	Positioning the Patient	Adjusting the patient's position to access the area of interest.	Skeletal Action Recognition
Cardiac U/S	Handling the Ultrasound Probe	Grasping, orienting, and maneuvering the probe.	Skeletal Action Recognition
Cardiac U/S	Applying Gel	Dispensing and spreading gel on probe or patient's skin directly.	Surface Contact
Cardiac U/S	Skin Contact	Placing the probe in contact with the skin and adjusting pressure.	Surface Contact
Cardiac U/S	Probe Manipulation	Moving the probe along different axes to obtain the best imaging results.	Surface Contact
Cardiac U/S	Adjusting Ultrasound Settings*	Adjusting settings on the ultrasound machine such as depth, brightness, and contrast.	Device/Tool Interaction
EKG	Positioning the Patient	Adjusting the patient's position to properly place EKG leads.	Skeletal Action Recognition
EKG	Attaching EKG Leads	Placing and securing the leads on the patient's body.	Surface Contact
EKG	Adjusting EKG Settings*	Modifying settings on the EKG device to ensure accurate readings.	Device/Tool Interaction
EKG	Monitoring EKG Output*	Observing and interpreting the EKG output on the device monitor.	Device/Tool Interaction

**Actions directly recognized by Device Specific Modules.*

3.A.2.5.1.2 Development of Action Descriptors

Based on the segmentation, a set of detailed action descriptors is created. These descriptors will be used for task-driven action recognition and are crucial for training the recognition models. The descriptors will include movement trajectories, application points, and device interaction times.

3.A.2.5.2. Skeletal Action Recognition Systems

3.A.2.5.2.1 Implement Pre-trained Models (Phase I)

In our proposed project, we aim to refine skeletal action recognition systems tailored for clinical settings. Using state-of-the-art models as a base, we will apply transfer learning to fine-tune them with a specific set of clinically relevant data. This step is crucial for ensuring our models accurately recognize and differentiate essential medical actions from other movements. Our objective is to enhance the model's ability to identify actions critical to patient care, such as manipulating medical instruments or adjusting patient positioning. Fine-tuning our models on specific clinical actions will improve their accuracy and reliability in distinguishing these vital activities.

3.A.2.5.2.2 Fine Tune

1. **Public Healthcare Datasets:** We will use existing healthcare datasets that provide annotated skeletal movements, such as UTKinect-Action3D or NTU RGB+D HealthCare [19]. These datasets will serve as an initial training base, helping our model adapt to general healthcare-related actions.
2. **Custom Clinical Data Collection:** To meet the unique demands of our targeted clinical tasks, including actions involved in cardiac ultrasound and EKG procedures, we will gather custom data. This involves recording specific procedure movements with RGB-D cameras in a clinical setting. The scope of the data collection will be determined by specific requirements of tasks identified by TA1.
3. **Annotation and Enhancement:** Clinical experts will annotate the collected data to accurately represent each clinical action. This precision is vital for training the model to recognize the exact movements associated with each task.

3.A.2.5.3. Surface Contact Action Recognition System

The anatomically annotated mesh data integrates with skeletal data, providing a comprehensive dataset that records both structural movements and surface contacts. This helps to accurately map where and how instruments like probes interact with the body, which is crucial for procedures like cardiac ultrasounds or EKGs.

3.A.2.5.3.1. Algorithmic Contact Recognition

We will develop a contact detection algorithm to recognize specific interactions such as 'probe in contact with skin'. This algorithm analyzes the collision and proximity of objects and body parts with annotated regions to detect how and when they interact. We also aim to identify the orientation and pressure applied by action based on the deformation of the mesh at the contact points. While not crucial, having this information would greatly benefit the guidance system.

3.A.2.5.3.2. Instrument Handling Recognition

For actions like 'holding the probe', the system assesses the configuration of the hand and fingers' skeletal data in relation to the probe’s position and orientation. This involves analyzing the spatial relationship between the skeletal joints of the hand and the instrument to confirm a grip posture.

3.A.2.6.Integration with Clinical Task Guidance System (Environment API)

Integrating our Environment Model and the Clinical Task Guidance System focuses on communicating detailed environmental data to support clinical tasks. This integration involves specific system components designed to facilitate seamless data transfer and accessibility.

3.A.2.6.1. Develop Data Aggregation Hub:

This component is the central repository where all data from the Environment Model's subsystems—such as 3D posture recognition, action recognition, face recognition, and object detection—are collected and standardized. The hub integrates data using a standard format and protocol, ensuring that information from various sensors and cameras is synchronized and easily accessible. This setup enables the Task Guidance System to pull consolidated and coherent environmental data as needed.

3.A.2.6.2. Develop Query Interface:

The query interface serves as the communication bridge between the Task Guidance System and the Environment Model. It allows the guidance system to request specific data relevant to ongoing clinical tasks. The interface supports requests for data like a patient's current posture, the specific location of an object, or a clinician's orientation, which are needed for tasks that require precise spatial and situational awareness.

3.B. Augmented Reality Hardware For Hand-Tracking & Egocentric Pose Data

The Apple Vision Pro Headset will track the medical staff’s head position, orientation, hand movements, and gaze. This headset provides essential tracking data that we integrate into our system to comprehensively monitor the clinician’s interactions when they wear it to provide spatial guidance. It is central to guidance tasks that involve precise localization and 3D maneuvering of devices (such as the ultrasound probe) and will be used in surface contact and object-holding recognition processes in the environment model.

3.B.1. Data Collection

The headset collects tracking data, including head position, orientation, hand movements, and gaze direction. Hand-tracking data is critical as it provides detailed information on the clinician's manual interactions with the environment and equipment, which we are not tracking with other devices. The device API provides access to 3D hand joints, which will be streamed to the central system for further processing.

3.B.2. Localization with Integrated World Map

Utilizing the headset's integrated world map feature, we accurately localize the clinician’s movements, particularly the hand movements, within the van’s Digital Twin model. This precise localization is crucial for monitoring detailed interactions and ensuring they are accurately represented in the virtual environment. This approach ensures that the clinician’s manual and physical interactions are accurately monitored and represented within our system.

3.C. Clinical Task Guidance System

System Overview. The Clinical Task Guidance System is a critical component of our integrated clinical environment, designed to facilitate precise and context-sensitive guidance to clinicians during medical procedures. This system gathers real-time data from the Environment Model, Session Manager, and various clinical device modules to provide dynamic direction through both interactive displays, wall-mount speakers and augmented reality (AR) devices. The system effectively synthesizes data from multiple sources to ensure comprehensive support and guidance:

Environment Model: This component provides continuous insight into the clinical environment. It captures and updates the positions and movements of clinicians and patients, along with the location and status of medical tools. By doing so, it helps maintain a real-time overview of the physical and dynamic aspects of the setting, crucial for accurate monitoring and interaction.

Session Manager: Acts as the procedural backbone, offering detailed task descriptions and structured, step-by-step guidance for clinical procedures. This manager ensures that the instructions delivered to clinicians are precise and tailored to meet specific procedural objectives. Session manager fetches visualizations and other relevant data from Spatial Procedure Guidance Database, including pre-computed and targeted 3D visualization such as indicators, 3D hand and body avatars, textual descriptions, video recordings associated with each task. The guidance system uses this information during the procedure and relays them to the interaction module.

Device Modules: These modules are integrated directly with various medical devices, such as ultrasound machines, to gather and analyze output data like imaging results. The evaluations provide additional information for assessing the progress and success of a task. For example, when a particular image is required from the ultrasound system, an ultrasound device module assesses the quality of the image and relays this information to the task guidance system.

The system's guidance and feedback are delivered to clinicians through the **Interaction Module,** which displays 3D guidance on AR headsets and provides visual and textual information on shared screens.

3.C.1. Task Model

The Task Model is designed to effectively guide clinicians through medical procedures by evaluating real-time observations and guiding actions towards predefined procedural goals. This model leverages provided task descriptions, which have annotated features such as positional information, and integrates observed data from various sources to ensure precision and compliance in clinical task execution.

3.C.1.1. Develop Task Evaluation System

The system continuously assesses the current state of the procedure using data inputs from the Environment Model and Device Modules. This includes tracking the positions and actions of clinicians, and the status and evaluations of medical tools. It incorporates task descriptions to

define the expected states and sequences for each task within a procedure. This integration ensures that the guidance provided aligns with the task description.

3.C.1.2. Develop Decision Making and Action Recommendation System:

The Task Model employs a logic-based decision-making framework to compare the current state against the expected state. This framework identifies discrepancies or deviations from the task description, using the specified constraints on posture, contact, and spatial location in the task description .Based on the analysis, the model generates specific recommendations for the next steps.Recommendations are directly tied to observations from the Environment Model, AR Hardware and the Device Modules and direct at matching them to the desired output. These recommendations direct the clinician on how to correct any deviations and proceed with the procedure according to the task description.

3.C.1.3 Integrate with system components

Data from the Environment Model, which includes dynamic updates on clinician movements, tool usage, and patient interactions, feeds into the Task Model. This data provides the necessary context for accurate state assessment. We also develop a communication protocol and utilities to communicate evaluations and recommendations through the Clinical Task Guidance System.

3.C.2.Develop User Model

The User Model is designed to personalize the clinical guidance system, tailoring it to the unique characteristics and needs of each clinician. This model enhances the interaction between the clinical system and the user by adapting its guidance to match the clinician's skills, preferences, and performance history.

3.C.2.1. Clinician Profiling

The model collects data on clinician interactions with the system during procedures. This includes tracking which guidance messages were followed, the outcomes of those actions, any deviations from suggested protocols, and the clinician's response time to prompts. Based on this data, the system creates a detailed profile for each clinician that includes their proficiency levels, preferred learning modalities (e.g., visual, textual, auditory), and historical performance metrics.

3.C.2.2. Adaptive Guidance System:

Develop specific algorithms and models to analyze the collected data to identify patterns in each clinician’s behavior and responses to system interactions. Techniques such as clustering for behavior grouping and regression analysis for performance trends are used.The system uses this analysis to customize the guidance provided. For example, it may offer more detailed instructions and visual aids to clinicians who are less experienced with certain procedures, or it may choose to provide succinct prompts to those who have demonstrated proficiency.

3.C.2.3. Continuous Learning and Updating:

The User Model incorporates a feedback loop that continually updates the clinician profiles based on new data from each interaction. This dynamic updating ensures that the guidance remains relevant and effectively supports the clinicians as their skills and preferences evolve.

Regular assessments of clinician performance, guided by the personalized feedback and instructions, help in measuring the effectiveness of the model. Adjustments to the guidance algorithms are made based on these assessments to enhance overall system performance.

3.C.3. Develop Recovery Model

There are two strategies we will use to recover from errors that occur during the guidance and to provide appropriate feedback. The first one, a closed-loop recovery, will perform two step error detection: (1) cross referencing Environment Model, AR Hardware Tracking, and Device Modules (2) automatic adjustment of guidance based on the deviation of the observed data from the expected observation. The second one, open-loop recovery, is designed to override the system guidance by the clinician if they think our guidance is not appropriate or the observation is incorrect. The open loop recovery is performed through the Interaction Module, which will allow skipping / disabling guidance information for each step of a procedure.

3.C3.1. Closed-Loop Recovery

Closed-loop recovery focuses on real-time error detection and correction within the system. This approach uses continuous feedback from the system's monitoring to adjust the guidance based on the clinician's actions and the procedure's progression. The system continuously monitors the execution of tasks against the expected protocols using data from the Environment Model, AR Hardware, and Device Modules. The first recovery is applied when a discrepancy is identified between the data obtained from these sources. For example, the AR Hardware might position the clinician's hand in contact with a surface, while the Environment Model positions it elsewhere. In these situations, an automatic resetting will align sensory information.

If all sensory information is aligned and discrepancies or deviations are detected between the expected observation and the guidance description, such as a clinician applying an incorrect technique or using the wrong tool, the system recognizes these errors. Upon detecting an error, the system automatically generates corrective feedback. This specific and actionable feedback guides the clinician to adjust their actions or revert to the correct procedural path. For instance, if the ultrasound probe is positioned incorrectly, the system will provide precise instructions to adjust its placement.

3.C3.2. Open-loop Recovery:

Open-loop recovery involves predefined interventions that do not rely on immediate feedback from the system's monitoring but are triggered based on certain conditions being met during the procedure. The system has a set of predefined responses or actions ready to be deployed when specific, non-critical errors occur or when there is a failure in the closed-loop system. These are based on common errors or typical procedural deviations known from clinical practice. In cases where the automated system may not adequately address an issue, or where clinician judgment is crucial, the system allows for manual override. Clinicians can choose to follow alternative paths or implement different strategies the system recommends, and skip or disable a guidance step through the Interaction Module.

3.D. Interaction Module

3.D.1. AR Visualizations

The clinician's AR headset has two main visualization components that enhance clinical examinations and procedures.

3.D.1.1. 3D Guidance System

This feature utilizes advanced AR technology to project visual assets directly into the clinician's field of view. It includes 'ghost' hands demonstrating clinical gestures either in first-person or third-person perspectives, allowing the clinician to follow precise movements in real space. Additionally, spatial indicators and labels are superimposed on the patient's body, clearly marking anatomical parts, areas of interest, or examination points, facilitating accurate clinical assessments.

3.D.1.2. Floating Windows

These are interactive UI elements within the AR display that serve multiple functions. They provide the clinician with exam checklists, lists of necessary equipment or tasks, and other essential information accessible within the visual field without interrupting the workflow. These panels are designed to float within the clinician's view, offering easy interaction and information access, streamlining the examination process and ensuring all necessary steps are followed.

3.D.2. Shared Visualization Display

The van should be equipped with an advanced Shared Visualization Display system in the main examination area to support real-time, augmented interactions between the clinician and the patient. The display is an essential tool for concurrent visual communication and clinical decision support. The display ensures that both the clinician and patient can view and understand the clinical guidance provided by CASEY. It enhances the diagnostic and treatment process by providing a view into the clinician’s actions and documenting the recommended steps and procedures of the clinical session. The display ensures that all actions are visible to the clinician and the patient, overcoming potential patient alienation due to the presence of the AR headset. Also, it allows for immediate auditing of the algorithm’s accuracy and relevance by the clinician, thereby maintaining a high level of precision and trust in the provided healthcare services.

3.D.2.1. Shared UI Framework

The development of the Shared UI Framework for the Shared Visualization Display involves utilizing JavaScript along with relevant libraries to create an intuitive and responsive user interface. To enhance realism and functionality, integration with web-based frameworks such as React.js or AngularJS can be considered. These frameworks offer robust solutions for building single-page applications (SPAs) with reusable components, state management, and efficient data binding capabilities, which are essential for creating a responsive and scalable UI in a healthcare setting. Furthermore, the framework should be designed to seamlessly integrate with the AR system and support the display of real-time augmented interactions between the clinician and the patient. Key considerations include optimizing the framework for performance, ensuring compatibility with the AR headset's capabilities, and facilitating easy interaction and information access for both the clinician and the patient. Moreover, incorporating AR-specific libraries and

SDKs such as ARCore for Android or ARKit for iOS will facilitate the integration of AR capabilities into the Shared UI Framework, ensuring compatibility with AR-enabled devices like smart glasses or mobile phones. These libraries provide APIs for features like plane detection, object tracking, and occlusion, allowing for the accurate placement and interaction of virtual content within the physical environment. Finally, the framework should support dynamic content updates and provide a flexible architecture for future enhancements and customization according to specific clinical needs and workflow requirements.

3.D.3. Conversational Interaction & Ambient Listening

The van is equipped to enhance auditory awareness during clinical sessions. It involves sophisticated microphone arrays strategically placed around the examination area to capture and analyze ambient sounds. This system allows the clinical staff to monitor the environment effectively, picking up on subtle cues such as changes in the patient's breathing or voice tone, which might indicate discomfort or stress. The listening system is refined to filter out background noise, focusing on relevant clinical sounds, thereby aiding in a more responsive and attuned patient care experience.

3.D.3.1. Develop Context-Aware Conversation Agent

A context-aware conversation agent that can interact with the patient and the clinician before and during the examination, depicted as CASEY, is a central component of our system. The conversation manager is responsible for listening to an ongoing conversation, identifying speakers, analyzing utterances, and extracting relevant information that needs to be relayed to the session manager. The conversation agent also determines if there is a need for simultaneous translation, and synthesizes speech to communicate with the clinician and/or the patient. To create the system, we implement the following [20]:

3.D.3.1.1. Speaker Identification

Our initial step involves implementing wake word detection, which serves as the trigger for activating the voice assistant (similar to “Hey Siri”). This detection task is managed by a compact on-device audio classification model (for example, Audio Classification Transformer model) designed to recognize specific wake words amidst ongoing audio input.

3.D.3.1.2. Analyze Transcript

Once the wake word is identified, the system transitions to speech transcription, where the spoken query is swiftly converted into text. To ensure efficiency, we opt for on-device automatic speech recognition (ASR) models, bypassing the need for transferring large audio files to the Cloud. Such models allow for near real-time transcription, enabling seamless interaction.

3.D.3.1.3 Perform Real-Time Processing using Domain Specific Large Language Model (LLM)

With the user's query transformed into text, the system proceeds to language model query processing, LLMs (see Section 3.H) to understand the semantics and generate a suitable response.

3.D.3.1.4. Synthesize Speech

Finally, the synthesized speech stage employs on-device text-to-speech (TTS, for example the Microsoft SpeechT5 TTS) models to convert the generated response into spoken words, providing a seamless user experience. The final system will be able to carry on long conversations on real-time based on up-to-date data provided by the session manager, including the patient's medical history, real-time data, and conversation history.

3.F. Session Manager

Session Manager is the main entry point to the clinical guidance system and is responsible for deciding on the course of actions, onboarding the clinician, providing introductory examination plan and task list, as well as controlling the guidance system for each medical procedure and its subtasks. It works in collaboration with DX for fetching diagnostic data and updating it with outcomes of procedures including medical imagery and evaluations, test results, and clinician's diagnosis. The Spatial Task Guidance Database stores necessary guidance data spatialized into the van's environment, which is fetched by the session manager when needed.

3.F.1. Spatial Task Guidance Database

The Spatial Task Guidance Database is designed to deliver structured and relevant data supporting the Clinical Task Guidance System by organizing a wide array of information for various medical procedures. This includes narratives, both static and dynamic visualizations, and specific task-related constraints and criteria.

3.F.1.1. Data Structures (Expand with each clinical service)

3.F.1.1.1 Narrative Model:

Each medical procedure is mapped in the database with a comprehensive narrative model. This encompasses procedural steps, relevant information, and the sequence required to guide clinicians effectively. It ensures clinicians follow the procedure correctly and understand each step's purpose and requirements.

3.F.1.1.2 Precomputed 3D Guidance Data:

This category stores extensive precomputed data like 3D models, visual indicators, text descriptions, and multimedia content. These assets are used to visually guide clinicians in a training context or as a reference during procedures, and as static instructional content.

3.F.1.1.3 Targeted Visualizations

Includes data that supports real-time rendering of 3D models and visual indicators during procedures. These visualizations are designed to be spatially and contextually accurate, dynamically aligning with the physical environment to guide the clinician's actions precisely. Targeted visualizations include a specific target such as anatomical region of the body, an instrument, or relative location around the clinician / patient, so they are mapped to the environment dynamically.

3.F.1.1.4 Task Constraints and Transition Criteria:

The database defines explicit constraints and success criteria for each task, such as the exact positioning required for patient setups or the orientation of medical tools. These parameters are crucial for verifying task completion and ensuring procedural accuracy. Table 3, below, provides an example of data types stored in the database:

Table 3. Data Storage.

Category	Task	Data Type	Description	Examples
Narrative Model		Structured Text	Detailed sequential procedural steps and relevant sub-tasks.	Steps include: Patient Preparation, Probe Placement, Image Capture, Image Analysis
Precomputed 3D Guidance Data	Patient Positioning	3D Models & Visual Indicators	Static models used for training or reference during patient setup.	3D model showing correct patient positioning on the bed.
	Probe Placement	3D Models & Visual Indicators	Static models and indicators for probe handling and placement.	3D model with highlighted zones on the torso for probe placement; instructional videos.
Targeted Visualizations	Probe Handling	Dynamic Templates	Real-time overlays showing correct probe handling and placement.	AR overlay showing probe placement on the patient's body in real-time.
	Probe Adjustment	Dynamic Templates	Real-time visual cues for adjusting probe angle and pressure.	Visual cues for adjusting probe angle and applying correct pressure; feedback on image quality.
Task Constraints and Transition Criteria	Setup Verification	Rule Sets	Constraints defining required setup conditions and verification criteria.	Rules defining required patient posture and probe positioning before proceeding.
	Image Capture Conditions	Rule Sets	Criteria specifying conditions under which images should be captured.	Constraints on probe position and patient response required to capture optimal images.

3.G. Clinical Decision Support (Dx) Module

The Dx Module will support the generation of necessary examinations lists and treatment plans.

3.G.1. Large Language Model with Retrieval Augmented Generation

The primary component of the conversation agent is an LLM that can base its reasoning on an available data source. The conversation agent should use up-to-date information regarding the examination process and be able to carry on a long conversation naturally.

3.G.1.1. Implement Pre-Trained LLM for Medical Diagnosis

To achieve real-time, high-quality medical diagnosis using an LLM, we will employ a streamlined approach that focuses on leveraging state-of-the-art LLMs for medical diagnosis. Utilizing LLMs such as Med-PaLM 2 by Google Research [22] - the first LLM to pass the U.S. Medical Licensing Examination - or the open-source Medical-Sft-Llama-3 by John Snow Labs [23, 24], which are designed for high-speed and accurate medical diagnosis, the system will process the individual information of each patient to generate accurate diagnostic predictions. Additionally, the conversation agent will be equipped with natural language processing capabilities to engage in meaningful dialogues with patients, eliciting pertinent information about their symptoms, medical history, and concerns. By seamlessly integrating these pre-trained LLMs with conversational abilities, our system aims to not only provide accurate medical diagnoses but also foster a human-like interaction experience, promoting patient trust and engagement throughout the diagnostic process.

3.G.1.2. Develop a Vector / Augmentation Database

To enhance the LLM's capabilities, we will develop a Vector/Augmentation Database to store and access real-time data efficiently. This database will employ advanced indexing techniques, such as vector embeddings and hashing, to represent and organize diverse types of information, including patient records, medical literature, and updates from clinical trials. By utilizing vector representations, we can capture the semantic similarities between different patients, enabling the LLM to retrieve relevant information effectively during the diagnostic process. Additionally, the database will support augmentation techniques, allowing the LLM to enrich its understanding by incorporating new knowledge and insights as they become available.

Through this approach, the LLM will have access to a dynamic repository of up-to-date information, enabling it to adapt and evolve alongside advancements in medical research and practice. This Vector/Augmentation Database will serve as a foundational component in empowering the LLM to provide accurate and informed medical diagnoses in real-time.

3.G.1.3. Fine-tune to Clinical Environment

We will use the aforementioned vector database to fine-tune open-source models, like the Medical-Sft-Llama-3. This fine-tuning process will optimize the models' performance by aligning them with the nuances of the clinical environment, including patient demographics, prevalent conditions, treatment protocols, and regional variations in medical practices. By tailoring the models to the specific characteristics of the proposed setting, we ensure their accuracy and effectiveness in providing timely and contextually relevant medical insights. To fine-tune the models we will utilize existing frameworks/libraries like the Transformers and the Autotrain libraries [25].

3.G.1.4 Retrieval-Augmented Generation (RAG)

Incorporating Retrieval-Augmented Generation (RAG) techniques into our system will further enhance its capabilities. By leveraging pre-existing knowledge from the Vector/Augmentation Database, RAG enables the model to generate responses by combining retrieved information with its own understanding. This approach not only ensures that generated content is contextually relevant but also facilitates the incorporation of the latest medical insights into

diagnostic recommendations. Through RAG, our system can dynamically adapt its responses based on real-time data and refine its diagnostic accuracy over time.

3.H. System Integration

For comprehensive system integration, we will connect the virtual medical agent application with other CDP healthcare systems and databases to enable seamless data exchange. We will ensure that data is normalized, transformed, and cleansed to maintain consistency and integrity across systems. We intend to establish protocols and standards for application interoperability to facilitate smooth communication. Additionally, we will utilize middleware components to manage interactions between the virtual agent and other applications. We will test the integrated system to identify and resolve any bugs or errors. We will also implement routine maintenance procedures to ensure the smooth operation of the integrated system and monitor system performance, making necessary adjustments to optimize functionality over time.

4. Open Source Release

The Field Guider intelligent task guidance system can be successfully transitioned to an open source project that encourages community participation, ensures code quality and stability, and promotes the widespread adoption and improvement of the system.

4.A. Licensing and Copyright

We will choose an open source license that allows for free use, modification, and distribution, such as the Apache License 2.0 or the MIT License. This will encourage adoption and contributions from the community. We will clearly state the copyright and license information in the project's README file and in the source code files.

4.B. Software Sharing and Documentation

4.B.1 Software Releases and Version Control

The project's source code will be hosted in a public code repository like GitHub, GitLab, or Bitbucket, making it easy for others to access, fork, and contribute to the project. We will use a version control system like Git to manage changes to the codebase, following best practices for commit messages, branching, and merging.

We will provide release notes for each version that highlight the new features, bug fixes, and breaking changes to make it easy for users to upgrade to newer versions.

4.B.2. Documentation

We will create comprehensive documentation that explains the project's purpose, features, architecture, and installation/setup instructions. Include a README file in the root of the repository. We will document the code using comments and docstrings and explain the purpose

and usage of each module, class, and function. We will provide usage examples and tutorials to help new users get started quickly.

Documentation will also include process and implementation details related to the medical parts of the project to facilitate the extension of clinical task databases to new tasks, and related research.

4.B.3 Continuous Integration and Testing

To automate integration processes ensuring the system's high quality we plan to:

1. Set up continuous integration (CI) to automatically build, test, and deploy the project whenever changes are made to the codebase, which will help ensure code quality and stability.
2. Write comprehensive unit tests and integration tests to verify the functionality of the system. Use test-driven development (TDD) practices to guide the development process.
3. Generate code coverage reports to track the percentage of the codebase that is covered by tests. We aim for high coverage to catch regressions early in the development cycle and ensure the stability of the system.

4.B.4 Security and Maintenance

To ensure robust security and maintenance, we will establish a triage and feedback loop for promptly addressing issues reported in the field. We plan to regularly review and update dependencies to mitigate security vulnerabilities and bugs. Additionally, we intend to respond swiftly to security issues raised by the community, maintaining a transparent process for addressing and disclosing these concerns. To guarantee the ongoing functionality and accessibility of our infrastructure, we will maintain the project's code repository, documentation, and communication channels.

4.C. Community Engagement

We will encourage community participation by providing clear guidelines for contributing, including outlining the process for submitting bug reports, feature requests, and pull requests. We will seek to set up communication channels like a mailing list, forum, or chat room where users and contributors can ask questions, discuss ideas, and collaborate. We will designate oversight and governance maintainers, and respond promptly to issues and pull requests submitted by the community.

To create an active community around the product, we will seek to present and publish it in relevant — medical, design, engineering — venues, encouraging interested individuals to use the service, and provide their ideas, experiences and feedback.

4.D. Commercialization Strategy

We will develop a commercialization strategy so that the Field Guider intelligent task guidance system can be successfully transitioned from a research project to a commercially viable product that improves patient care. Elements of the commercialization strategy will include:

1. **Intellectual Property Protection**: Securing intellectual property rights, such as patents and trademarks, to protect the unique features and innovations of the system.
2. **Partnerships and Licensing**: Identifying potential partnerships with healthcare technology companies and healthcare providers to facilitate the commercialization and distribution of the system.
3. **Marketing**: Create a comprehensive marketing strategy to promote the intelligent task guidance system to potential customers, including healthcare facilities, medical schools, and professional organizations.
4. **Training and Support**: Developing a comprehensive training program and providing ongoing support to ensure that healthcare workers can effectively use the system and maximize its benefits.
5. **Scalability and Expansion**: Optimizing the system with scalability in mind, allowing for easy integration with existing healthcare systems and the potential for expansion into other clinical applications.

CAPABILITIES/MANAGEMENT PLAN

Summary of Team Capabilities and Expertise

Juhan Sonin, GoInvo's Director will serve as the Principal Investigator (PI) for the Field Guider project. We Create Goodness, LLC dba GoInvo is a human-centered design firm focused on healthcare. For over 15 years, GoInvo's design approach has incorporated a deep understanding of clinicians, patients, and administrators with technical knowledge of health IT, integrated with UX design best practice and software engineering. GoInvo has worked with organizations as far-reaching as AstraZeneca, Becton Dickinson, Johnson & Johnson, 3M Health Information Services, and the U.S. Department of Health and Human Services.

Dr. Cagri Hakan Zaman, Founding Director of MIT Virtual Experience Design Lab and Co-Founder and President of Virtual Collaboration Research, Inc. (Mediate) will serve as the project's Technical Lead. He leverages his expertise in AI, computer vision, and spatialized immersion to develop novel neural networks that robustly parse 3D spaces and optimize them for edge devices.

John Brownstein, the Chief Innovation Officer at Boston Children's Hospital, will serve as the project's Clinical Innovation Lead. Boston Children's Hospital is ranked among the best pediatric hospitals in the US.

Mollie Williams, the Executive Director of Mobile Health Map and The Family Van, will serve as the project's Clinical User Lead. Mobile Health Map and The Family Van, programs of Harvard Medical School, are designed to expand access to healthcare services across the diverse communities in the Greater Boston area and help other mobile clinics across the U.S. measure and communicate their impact.

Team Organization and Responsibilities

The key roles and responsibilities in the project organization are as follows:

Project Management and Design

Juhan Sonin, GoInvo's Director, will serve as the Principal Investigator for the project. He will be responsible for overall project planning, design coordination, and delivery of the final integrated hardware and software AR task guidance system. He and the design team will:

- Oversee the project timeline, budget, and resource allocation across the different phases.
- Design the Field Guider system and user experience, from initial concepts to finished AR interfaces.
- Conduct user testing and evaluation of the system's impact on task proficiency and user trust.
- Ensure effective collaboration between the project's design, technical and clinical teams, as well as the teams working on other technical areas, TA1-4.
- Plan for open source commercialization and shipping of the final AR task guidance system.
- Carry out compliance checks to verify adherence to healthcare regulations and safety and privacy standards.

Technical Team

Dr. Cagri Hakan Zaman, Mediate's Director, will serve as the Technical Lead for the project. The engineering team will be responsible for the development and integration of the AR software and hardware components. He and the engineering team will:

- Conduct the needs assessment, system architecture design, and hardware selection.
- Create the initial working prototypes for the AR interface and task guidance.
- Implement the AR software, including the user interface, task models, feedback mechanisms, and content management system.
- Carry out comprehensive testing (unit, integration, system, stress, and load) to ensure seamless hardware-software interaction and system stability.
- Plan for system deployment including site preparation, phased rollouts, and scalability considerations.
- Ensure the seamless integration of the AR system into the PARADIGM care delivery platform.
- Prepare the documentation and compliance reports for the final system.

Clinical Team

John Brownstein, Chief Innovation Officer at Boston Children's Hospital serves as the project's Clinical Innovation Lead. He and the clinical innovation team will:

- Assist in outlining the clinical requirements and user needs for the AR task guidance system.
- Provide expert input on medical workflows and task models.

Mollie Williams, Executive Director of Mobile Health Map and The Family Van serves as the project's Clinical User Lead. She and the clinical user team will:

- Participate in the usability testing and iterative evaluation of the system to provide feedback for continuous improvements.
- Help integrate clinical expertise into the software to enable handling of a broader range of tasks.

Bibliography

[1] El Kaid A, Baïna K. A Systematic Review of Recent Deep Learning Approaches for 3D Human Pose Estimation. Journal of Imaging. 2023 Dec 12;9(12):275. www.mdpi.com/2313-433X/9/12/275

[2] Zhou L, Meng X, Liu Z, Wu M, Gao Z, Wang P. Human Pose-based Estimation, Tracking and Action Recognition with Deep Learning: A Survey. arXiv preprint arXiv:2310.13039. 2023 Oct 19. ar5iv.labs.arxiv.org/html/2310.13039

[3] Kocabas M, Huang CH, Tesch J, Müller L, Hilliges O, Black MJ. SPEC: Seeing people in the wild with an estimated camera. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021 (pp. 11035-11045). arxiv.org/pdf/2110.00620.pdf.

[4] Elmi A, Mazzini D, Tortella P. Light3DPose: real-time multi-person 3D pose estimation from multiple views. In2020 25th International Conference on Pattern Recognition (ICPR) 2021 Jan 10 (pp. 2755-2762). IEEE. arxiv.org/pdf/2004.02688.pdf

[5] Ye H, Zhu W, Wang C, Wu R, Wang Y. Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In European Conference on Computer Vision 2022 Oct 23 (pp. 142-159). Cham: Springer Nature Switzerland. arxiv.org/pdf/2207.10955.pdf.

[6] Hong Y, Zhang J, Jiang B, Guo Y, Liu L, Bao H. Stereopifu: Depth aware clothed human digitization via stereo vision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021 (pp. 535-545). arxiv.org/pdf/2104.05289.pdf

[7] Li Y, Zhang L, Qiu Z, Jiang Y, Li N, Ma Y, Zhang Y, Xu L, Yu J. Nimble: a non-rigid hand model with bones and muscles. ACM Transactions on Graphics (TOG). 2022;41(4):1-6. arxiv.org/pdf/2202.04533.pdf

[8] Zhang H, Tian Y, Zhou X, Ouyang W, Liu Y, Wang L, Sun Z. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021 (pp. 11446-11456). Code: github.com/HongwenZhang/PyMAF

[9] TensorFlow. Mesh Segmentation using Feature Steered Graph Convolutions.Colab: colab.research.google.com/github/tensorflow/graphics/blob/master/tensorflow_graphics/notebooks/mesh_segmentation_demo.ipynb. Accessed July 2020

[10] Joo H, Liu H, Tan L, Gui L, Nabbe B, Matthews I, Kanade T, Nobuhara S, Sheikh Y. Panoptic studio: A massively multiview system for social motion capture. In Proceedings of the IEEE International Conference on Computer Vision 2015 (pp. 3334-3342).
[11] ibid. Dataset: Panoptic domedb.perception.cs.cmu.edu/

[12] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015 (pp. 815-823). arxiv.org/pdf/1503.03832.pdf

[13] Qi D, Tan W, Yao Q, Liu J. YOLO5Face: Why reinventing a face detector. European Conference on Computer Vision 2022 (pp. 228-244). Springer Nature. arxiv.org/pdf/2105.12931.pdf

[14] Chanda S, GV AC, Brun A, Hast A, Pal U, Doermann D. Face recognition-a one-shot learning perspective. In2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS) 2019 (pp. 113-119). IEEE. ieeexplore.ieee.org/abstract/document/9067938

[15] Bae G, de La Gorce M, Baltrušaitis T, Hewitt C, Chen D, Valentin J, Cipolla R, Shen J. Digiface-1m: 1 million digital face images for face recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2023 (pp. 3526-3535). github.com/microsoft/DigiFace1M

[16] Shaikh MB, Chai D. RGB-D data-based action recognition: a review. Sensors. 2021;21(12):4246. www.mdpi.com/1424-8220/21/12/4246

[17] Chang H, Chen J, Li Y, Chen J, Zhang X. Wavelet-Decoupling Contrastive Enhancement Network for Fine-Grained Skeleton-Based Action Recognition. arXiv preprint arXiv:2402.02210. 2024 Feb 3. ar5iv.labs.arxiv.org/html/2402.02210

[18] Duan H, Zhao Y, Chen K, Lin D, Dai B. Revisiting skeleton-based action recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022 (pp. 2969-2978). arxiv.org/pdf/2104.13586v2.pdf

[19] Liu J, Shahroudy A, Perez M, Wang G, Duan LY, Kot AC. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence. 2019;42(10):2684-701. paperswithcode.com/dataset/ntu-rgb-d

[20] HuggingFace. Creating a voice assistant. Available on: huggingface.co/learn/audio-course/en/chapter7/voice-assistant. Accessed February 2024.

[21] Gao X, Li W, Loomes M, Wang L. A fused deep learning architecture for viewpoint classification of echocardiography. Information Fusion. 2017;36:103-13. doi.org/10.1016/j.inffus.2016.11.007

[22] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80. www.nature.com/articles/s41586-023-06291-2

[23] HuggingFace. John Snow Labs. JSL-Med-Sft-Llama-3-8B. Available on: huggingface.co/johnsnowlabs/JSL-Med-Sft-Llama-3-8B. Accessed April 2024.

[24] HuggingFace. The Open Medical-LLM Leaderboard: Benchmarking Large Language Models in Healthcare. Available on: huggingface.co/blog/leaderboard-medicalllm. Accessed April 2024.

[25] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771. 2019 Oct 9. huggingface.co/docs/transformers/en/index

Resumes

Resumes for the project's key personnel follow:

Juhan Sonin, Director, GoInvo

Dr. Cagri Hakan Zaman, Director, Mediate

John Brownstein, Chief Innovation Officer, Boston Children's Hospital

Mollie Williams, Executive Director of Mobile Health Map and The Family Van