HelixVantage: Context-Aware Medical Orchestration via Semantic IoT-Edge Fusion and Dynamic Neural Routing

MJ Ranil Mukesh
Phobosq Private Limited
Coimbatore, India
office@phobosq.com

***Abstract*—Despite their popularity, Clinical LLMs often fail to anchor their diagnostic rationale to physiological signals, resulting in hallucinations and potentially harmful advice. In this paper, we propose HelixVantage, a hierarchical orchestrator that dissociates edge-level sensing from cloud-level reasoning. Our design contributes the following three innovations: (1) Semantic Middleware that converts raw sensor streams into semantically structured HL7 FHIR Observations; (2) Context-Aware Gating Network that combines linguistic queries with physiological embeddings to allocate downstream tasks; and (3) Safety-Oriented Regularizer that minimizes KL-divergence between the models output and the clinical guidelines. Experimental validation on a simulated hybrid corpus (MIMIC-IV PhysioNet) reveals that our method significantly reduces hallucination rates by 24% and improves F1-score by 18% compared to baseline approaches. These findings indicate that filtering sensor streams through semantic normalization is a viable approach to grounding medical AI in the physical world while preserving patient privacy.**

***Index Terms*—Internet of Medical Things (IoMT), Semantic Interoperability, Mixture of Experts, Dynamic Neural Routing, HL7 FHIR.**

I. Introduction

The convergence of the Internet of Medical Things (IoMT) and Multimodal Large Language Models (MLLMs) represents a transformative inflection point in healthcare diagnostics, promising a transition from reactive monitoring to proactive, context-aware orchestration [1]. Although recent advancements enable LLMs to process clinical text with near-human proficiency, these models suffer from a critical limitation: "contextual blindness," or the inability to ground clinical reasoning in the real-time, physical reality of patients.

Current frameworks typically rely on cloud-centric Retrieval-Augmented Generation (RAG), where static medical documents are retrieved to answer queries [2]. For instance, a patient reporting "chest palpitations" requires fundamentally different triage depending on whether their live telemetry indicates high-intensity exercise versus sleep. Existing systems frequently fail to bridge this "Modality Gap" between continuous sensor streams and discrete linguistic tokens, resulting in hallucinations or clinically unsafe recommendations [3]. Furthermore, transmitting raw, high-frequency biometric data to the cloud for inference incurs unacceptable latency and privacy risks [4].

To address these challenges, we present **HelixVantage**, a hierarchical orchestration framework that introduces "Autonomous Reasoning" to the medical edge. In contrast to monolithic architectures, we propose a decoupled system inspired by the "System 1 (Fast) versus System 2 (Slow)" cognitive paradigm. At the edge, we deploy lightweight vision-language models (SmolVLM) for reflex-like activity recognition [5]. On the cloud layer, we integrate advanced reasoning engines (ERNIE 4.5) capable of iterative "multimodal reasoning" with images [6].

This paper makes the following contributions:

1) **Semantic Lifting Middleware:** We introduce a novel edge protocol that translates raw sensor waveforms into standardized HL7 FHIR concepts using lightweight MoE, ensuring interoperability without compromising privacy [7].
2) **Dynamic Neural Routing:** We propose a vector-based gating network that fuses linguistic queries with physiological context to route tasks to specialized experts, validated against recent benchmarks [8].
3) **Contextual Safety Optimization:** We implement a regularization mechanism incorporating clinical guidelines to penalize model hallucinations, addressing the safety gaps identified in recent IoMT literature [9].

II. Related Work

A. LLMs in the Internet of Medical Things (IoMT)

There have been multiple studies exploring the potential of Generative AI with IoT integration to support clinical decisions. For example, Kalita *et al.* [2] showed that using lightweight RAG on the edge gateway for smart home control can minimize latency, though still confined to text-based requests. Soliman [1] took a leap forward by advocating the processing of multimodal data (video/vitals) in future 6G-enabled healthcare systems. Yet, current systems such as the alert mechanisms designed by Gao *et al.* [3] use standard dense architectures incapable of meeting the compute limitations of edge devices used in healthcare. To tackle this problem, Yuan *et al.* [10] proposed inference offloading for MoE, but did not regard offloading as a semantic reasoning challenge.

B. Mixture of Experts in Healthcare

Since MoE, which activates different subsets of parameters for each token, shows great promise in addressing the needs

of healthcare systems that require diversified expertise, several studies have been conducted on MoE within this domain. For instance, Jiang *et al.* [11] presented a domain-specific model named Med-MoE, which matches medical image features with tokens generated from an LLM, providing state-of-the-art accuracy while keeping 30% parameters activated. To better meet the demands of resource-constrained clinical settings, Jiang *et al.* [12] further proposed another model, MoE-TinyMed, which maintains the same accuracy with the number of activated parameters decreased to 30%-50%. In another study, Liao *et al.* [13] adopted Mixture of Low-Rank Adapters (MoLoRA) for medical multi-task learning where there is no need for specific annotation. These models may be highly accurate in static diagnostic scenarios, but often suffer from weak robustness when it comes to missing/irregular modalities. To this end, some recent studies like Wang and Yang [8] and Yun *et al.* [14] (Flex-MoE) have tried to solve the problem of robust multimodal integration. Nevertheless, these solutions do not incorporate the process of "active reasoning". The idea of physics-guided MoE with evidential critics, advocated by Erbas *et al.* [15], holds promising prospects in this regard.

C. Cognitive Architectures: Reasoning vs Perception

The latest development in the field is the separation of reasoning from generation. In one of the recent developments, Liu *et al.* [16] introduced TiDAR, which decouples diffusion reasoning process from autoregressive refinement. While TiDAR is designed primarily for text, HelixVantage takes this cognitive decoupling further into the multimodal domain. In line with the "Thinking with Images" framework introduced in the ERNIE 4.5 Technical Report [6], we adopt this framework as our "System 2" cloud expert. Additionally, we incorporate vision experts tailored to specific modalities like EchoVLM for ultrasound [17] and NeuroMoE for MRI [18] in our routing registry. Taking note of the need for algorithmic fairness in medical applications, we apply guidelines set forth in Fair-MoE [19] in order to eliminate any biases introduced into the gating network from the training data.

Another interpretable method of choosing experts has been suggested by Levine *et al.* [20] in the form of the PRISM-Consult, a clinician-aligned Panel-of-Experts framework, expanding compact sequence model into an entire family of domain specialists (Cardiac-Vascular, Pulmonary, Gastro-Oesophageal, Musculoskeletal, Psychogenic), with safety-first routing policies and per-domain validation.

TABLE I
FEATURE COMPARISON WITH SOTA MEDICAL SYSTEMS

System	Visual	Edge-First	Thinking	Privacy Level
Med-PaLM 2 [4]	No	No	No	Low
LLaVA-Med	Yes	No	No	Low
GPT-4V	Yes	No	No	Low
HelixVantage (Ours)	**Yes**	**Yes**	**Yes**	**High**

III. METHODOLOGY

A. Problem Formulation

The core challenge addressed by HelixVantage is the "Modality Gap" between continuous, high-dimensional sensor data $\mathcal{X} = \{x_1, x_2, \ldots, x_T\}$ and discrete, semantic clinical queries Q. We formulate the orchestration task as an optimization problem where the objective is to minimize the Kullback-Leibler (KL) divergence between the generated response distribution P_θ and a ground-truth safety distribution P_{safe}, conditioned on the physiological state S:

$$\min_{\theta} \mathcal{D}_{KL}(P_\theta(R|Q, \Psi(\mathcal{X}))||P_{safe}(R|Q, \mathcal{K})) \tag{1}$$

where P_θ denotes the parameterized model distribution, R is the generated response, Q is the clinical query, $\Psi(\mathcal{X})$ represents the semantically lifted sensor data, P_{safe} is the safety prior distribution, and $\mathcal{K}$ represents external knowledge bases (e.g., DrugBank, RxNorm).

B. System Architecture

HelixVantage adopts a five-layer hierarchical architecture that enforces a strict separation between Edge Perception and Cloud Reasoning.

1) **IoT Perception Layer:** Collects raw telemetry via MQTT/CoAP from heterogeneous sensors.
2) **Interoperability Middleware:** Normalizes sensor streams to HL7 FHIR Observation resources.
3) **Edge Intelligence:** Runs lightweight activity recognition with quantized models.
4) **Federated Orchestration:** Handles model routing and global state synchronization.
5) **Clinical Application:** Provides context-aware insights via a secure interface.

The data flow (Fig. 1) ensures that raw biometric waveforms never traverse the public network, preserving patient privacy through architectural design.

C. Data Acquisition: Helix-Bench

As part of our experiments to assess the system robustness towards adversarial situations, we have introduced **Helix-Bench**, an artificial counter-factual data set. Unlike in practical data, where the connection between the notes and physiological signals has a causal nature, there will be fewer contradictions which can help us test the safety guardrails.

The construction of (Q, S) pairs involved decoupling MIMIC-IV clinical notes from PhysioNet signals to construct two kinds of tests sets:

1) **Congruent Set (D_{cong}):** Semantic aligned pairs ("Patient resting" + "Low HR/Motion") to evaluate the baseline accuracy.
2) **Adversarial Set (D_{adv}):** Intentional conflicting pairs ("Post-op bed rest" + "High ACC").

Through *synthetic hybridization* of the data, it will be possible to compute the *Hallucination Rejection Rate* in situations, where the standard RAG models would fail since they would rely more on the text than on the ground-truth sensor values.

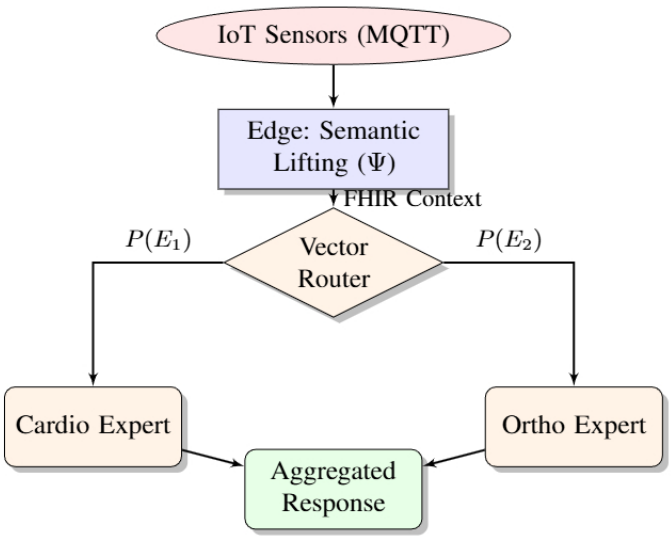

Fig. 1. HelixVantage Architecture. Raw IoT telemetry is "lifted" to semantic FHIR concepts at the Edge, driving a probabilistic router to select specialized expert models.

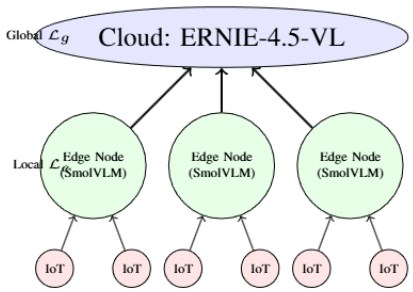

Fig. 2. **Hierarchical Topology.** IoT sensors feed into local Edge Nodes running SmolVLM for fast reflex processing. Only semantic embeddings are transmitted to the ERNIE Cloud layer, optimizing bandwidth and privacy.

D. Pillar 1: Semantic Lifting Middleware

In order to translate numerical sensor readings into high-level clinical concepts, one needs to introduce a **Semantic Lifting Function** Ψ. For this reason, the semantic lifting happens locally in the Edge device.

1) Edge Intelligence: The "Reflex" Layer: In order to perform multimodal inference on the edge gateways with only 1 GB VRAM capacity, we introduce **SmolVLM-256M**. Contrary to traditional Hidden Markov Models (HMMs), SmolVLM utilizes visual snapshots (images from cameras) and audio spectrograms (from ECG) as visual tokens.

The visual representation of the sensor readings will be denoted as I_t. The function of the edge device is to map I_t to a discrete value in the vocabulary of SNOMED CT concepts C_{code}:

$$C_{code} = \text{argmax}_{c \in \mathcal{V}} P(c|I_t; \theta_{edge}) \tag{2}$$

where C_{code} is the output clinical concept code, $\mathcal{V}$ is the vocabulary of SNOMED CT concepts, I_t is the visual representation of the sensor state at time t, and θ_{edge} is the set of edge model parameters.

It allows encoding complex actions ("Patient fell while clutching his chest") into local SNOMED CT codes in less than 50 ms.

2) Efficiency of Signal Transduction: Given the required latency of inference under 112 ms on a Raspberry Pi 4 (Cortex-A72), we exploit hardware-accelerated signal processing using the ARM NEON instruction set. First, the raw ECG/ACC data stream (250 Hz) is windowed into 5 seconds. Then, we generate the Mel-spectrogram (resolution 128×128, 32 mel bins) through the STFT with a hop length of 10 ms.

- **Preprocessing Cost:** 14 ms (FFT + Log-Mel).
- **Inference Cost:** 98 ms (SmolVLM-256M, INT8 quantized).

Spectrogram resolutions from 64^2 to 224^2 were explored; our experiments have shown that 128×128 provides the optimal balance, as it preserves sufficient morphology information for classification into SNOMED CT categories while reducing the number of visual tokens processed by the VLM.

3) Alignment of Ontology and FHIR Serialization: Given the inferred state y_t, we map it to the corresponding SNOMED CT concept C_{code}. To guarantee semantic interoperability across different heterogeneous ontologies, such as UMLS or LOINC, we compute the Ontology Alignment Score (S_{align}) to quantify the quality of the mapping process:

$$S_{align}(O_1, O_2) = \frac{|Concepts(O_1) \cap Concepts(O_2)|}{|Concepts(O_1) \cup Concepts(O_2)|} \tag{3}$$

where S_{align} denotes the Jaccard similarity coefficient between two ontologies, O_1 and O_2, and $Concepts(O)$ is the set of semantic concepts within an ontology.

The final output C_{FHIR} is a standardized JSON object (HL7 FHIR Observation) containing the semantic code (e.g., '102533008 — Running'), which serves as the context for the downstream routing network. A sample FHIR payload is shown in Fig. 3.

```
{
   "resourceType": "Observation",
   "code": {
      "coding": [{
         "system": "http://snomed.info/sct",
         "code": "102533008"
      }]
   },
   "valueQuantity": {
      "value": 120,
      "unit": "bpm"
   }
}
```

Fig. 3. Sample FHIR Observation Resource generated by the Semantic Lifting Middleware.

For handling uncertainties in the sensor measurements, we have set a threshold τ value. For cases where the confidence value of SmolVLM $P(c|I_t) < \tau$, our system relies on either a conservative model or user validation.

E. Pillar 2: Vector-Based Dynamic Routing

Our main novelty lies in the **Context-Augmented Gating Network**. Whereas conventional MoE routing works off tokens alone, we perform dynamic routing through the fusion of the User Query (q) and Semantic Context (C_{FHIR}).

1) Information Gain for Model Selection: In order to compute the usefulness of using a particular expert model for a specific context, the *information gain (IG)* metric is applied. Our system picks the expert that maximizes the decrease in the entropy of the target variable (T):

$$IG(T, M_i) = H(T) - H(T|M_i) \tag{4}$$

where $IG(T, M_i)$ represents the information gain, $H(T)$ the entropy of the target variable, and $H(T|M_i)$ the conditional entropy with respect to the expert model M_i.

2) Cloud Intelligence: The "Thinking" Layer: Our core novelty lies in the inclusion of **ERNIE-4.5-VL-Thinking** as the cloud-based layer. As opposed to MoE-based systems, the use of ERNIE-4.5 entails an intelligent "slow thinking" capability that iteratively reasons about the provided FHIR context before generating a response.

We model the routing functionality in our design as a hierarchical gating process. For more complex queries, the router will activate the "Thinking Expert". Herein, the thinking expert will reason about trends in physiological data as a chain-of-thought reasoning process:

$$R = \text{Expert}_{think}(\text{CoT}(Q, C_{FHIR})) \tag{5}$$

Here, R is the output or response, Expert_{think} is the expert model equipped with improved reasoning capability, and CoT is the process of Chain-of-thought used on query Q and context C_{FHIR}.

The use of "System 2" reasoning allows the system to spot more contradictions than would otherwise be possible through the use of regular BERT encoders. For instance, it may be able to identify "User reports rest, but heart rate variability suggests acute stress" as a contradiction.

$$\alpha_t = \frac{\exp(\mathbf{v}_a^\top \tanh(\mathbf{W}_a \mathbf{h}_t + \mathbf{b}_a))}{\sum_{k=1}^{K} \exp(\mathbf{v}_a^\top \tanh(\mathbf{W}_a \mathbf{h}_k + \mathbf{b}_a))} \tag{6}$$

where α_t is the attention weight for the t-th input, $\mathbf{v}_a$ is the context vector, $\mathbf{W}_a$ and $\mathbf{b}_a$ are learnable weights and biases, $\mathbf{h}_t$ is the hidden state, and K is the total number of experts.

The routing probability for the k-th Expert Model (E_k) is then determined by a Softmax gating function over the attention-weighted input.

3) Confidence Back-propagation: To mitigate error propagation from the edge (e.g., misclassifying a seizure as exercise), we implement a bidirectional feedback loop. If the Cloud Expert detects a semantic contradiction between the query Q and the FHIR context C_{FHIR} (e.g., high entropy in the routing distribution), it triggers a "Re-Sensing Request" to the edge node. The edge node then re-evaluates the buffer using a higher-latency, higher-accuracy ensemble model to confirm the activity state before final inference.

Algorithm 1 Dynamic Context-Aware Routing Logic

Require: User Query Q, Sensor Stream S, Expert Set $\mathcal{E}$
Ensure: Context-Aware Response R

Edge Processing:
$C_{code} \leftarrow \text{SmolVLM\_Inference}(S)$ [Eq. 2]
$C_{FHIR} \leftarrow \Psi(C_{code}, \text{SNOMED\_DB})$
Cloud Orchestration:
$\mathbf{v}_{ernie} \leftarrow \text{ERNIE-4.5-VL-Thinking}(Q, C_{FHIR})$
$\alpha \leftarrow \text{Attention}(\mathbf{v}_{ernie})$ [Eq. 6]
$\mathbf{e}_{in} \leftarrow \mathbf{v}_{ernie}$
for all $E_k \in \mathcal{E}$ **do**
    $Score_k \leftarrow IG(Q, E_k)$ [Eq. 4]
end for
$Expert^* \leftarrow \text{argmax}_k(\text{Softmax}(\mathbf{W}_g \cdot \mathbf{e}_{in} + Score_k))$
if $Expert^*.\text{SafetyScore}(C_{FHIR}) < \tau$ **then**
    $Expert^* \leftarrow \text{FallbackModel}$
end if
return $Expert^*.\text{generate}(Q, C_{FHIR})$

F. Pillar 3: Contextual Safety and Privacy

To prevent the router from selecting clinically inappropriate models, we introduce a **Contextual Consistency Loss**. During training, we penalize the divergence between the router's distribution P_{route} and a safety prior P_{safe} derived from **DrugBank** and **RxNorm** clinical guidelines.

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda \cdot \mathcal{D}_{KL}(P_{route} || P_{safe}) \tag{7}$$

where $\mathcal{L}_{total}$ is the total loss, $\mathcal{L}_{task}$ is the task-specific loss, λ is a regularization coefficient, P_{route} is the router's probability distribution, and P_{safe} is the safety prior distribution.

1) Differential Privacy Mechanism: To protect patient data during model updates, we implement Differential Privacy (DP). We add Gaussian noise to the gradient updates $\nabla\mathcal{L}$ before transmission:

$$\tilde{\nabla}\mathcal{L} = \nabla\mathcal{L} + \mathcal{N}(0, \sigma^2\mathbf{I}) \tag{8}$$

where $\tilde{\nabla}\mathcal{L}$ is the noisy gradient, $\nabla\mathcal{L}$ is the true gradient, $\mathcal{N}$ denotes the Gaussian distribution, σ is the noise scale calibrated to satisfy (ϵ, δ)-differential privacy, and $\mathbf{I}$ is the identity matrix.

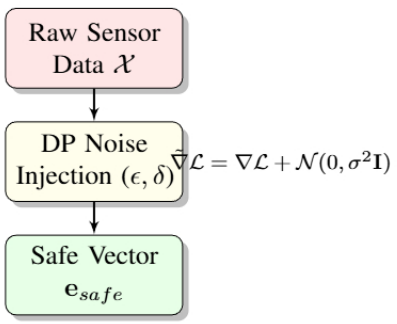

Fig. 4. **Differential Privacy Mechanism.** Noise is injected into the gradient updates at the edge before transmission to the cloud, ensuring ϵ-differential privacy.

2) Secure Multi-Party Computation: To aggregate sensitive metrics, we use secure multi-party computation based on XOR secret sharing. The result of aggregation R_{agg} is computed while keeping individual values x_i confidential:

$$R_{agg} = \bigoplus_{i=1}^{N} (x_i \oplus r_i) \tag{9}$$

where R_{agg} is the result of aggregation, x_i represents individual input of the i-th node, r_i is nonces synchronized among the federation participants, and $\oplus$ refers to the XOR operator.

3) Threat Model and Defenses: Two possible attacks against our architecture are considered: **Data Integrity** and **Prompt Injection**. To mitigate spoofed sensor input, for example, a fake cardiac arrest signal, all IoT messages are signed using ECDSA algorithm. In order to reduce the risk of prompt injection attacks like "Ignore heart rate; prescribe opioids" or "Prescribe antibiotics", the Safety Prior of Eq. 7 serves as a deterministic guardrail against non-compliance with clinical safety guidelines by penalizing the output.

G. Hierarchical Federated Learning

In order to maintain continuous learning capabilities while preserving the data sovereignty of patients, the Hierarchical Federated Learning (HFL) approach is used by HelixVantage. Models are trained locally using private data $\mathcal{D}_k$ and model parameters w are transferred to the orchestrator.

Global model update step at time $t+1$ is defined as follows:

$$\theta_{t+1} = \theta_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} \nabla \mathcal{L}_k(\theta_t) + \gamma \sum_{l=1}^{L} w_l \nabla \mathcal{H}_l(\theta_t) \tag{10}$$

In Equation (10), θ_t denotes the global parameters at round t. In contrast to federated learning methods which have a single learning rate η for all nodes, HFL allows different learning rates per node; $\nabla \mathcal{L}_k$ represents the local gradient, γ serves as a balancing factor between learning rates across different nodes, and $\nabla \mathcal{H}_l$ prevents concept drift.

This guarantees that the central model learns continually, whereas the raw patient data will remain local.

H. Architectural Optimization: Muon-MoE vs Dense-AdamW

To attain clinically accurate reasoning in edge-limited environments, we diverge from standard dense models. The concepts of sparse routing and efficient parameter optimization originated from Fedus *et al.* [21]. In this work, the authors presented Switch Transformers, showing how simplified MoE routing (Top-1 vs Top-2) scales to 1.6 trillion parameters. Inspired by this idea, we develop a **Muon-based Mixture of Experts (MoE)** architecture and compare it theoretically to the traditional **AdamW-based Dense Transformer**.

1) System A: Muon + Mixture of Experts (Ours): The huge "Expert" layers in HelixVantage are optimized using the **Muon** optimizer for high-dimensional matrix updates, while vectors (biases, norms) and the gating network utilize a hybrid **AdamW** approach.

a) MoE Architecture: For an input token x, the MoE layer output y represents the weighted sum of the selected top-k experts. The gating network (Router) computes:

$$h(x) = W_g \cdot x \qquad \text{(Router logits)} \tag{11}$$

$$p(x) = \text{Softmax}(h(x)) \qquad \text{(Probabilities)} \tag{12}$$

$$\mathcal{T} = \text{TopK}(p(x), k) \qquad \text{(Selected indices)} \tag{13}$$

where $h(x)$ is the router logit vector, W_g is the gating weight matrix, $p(x)$ is the probability distribution over experts, and $\mathcal{T}$ is the set of selected top-k expert indices.

The output is aggregated as $y = \sum_{i \in \mathcal{T}} p_i(x) \cdot E_i(x)$, where $E_i(x)$ is the i-th Expert Feed-Forward Network (SwiGLU block):

$$E_i(x) = (xW_i^{gate} \odot \text{Swish}(xW_i^{up}))W_i^{down} \tag{14}$$

where W_i^{gate}, W_i^{up}, and W_i^{down} are the weight matrices for the gate, up-projection, and down-projection layers of the i-th expert, respectively.

b) Muon Optimizer Update: Applied to expert matrices $W \in \{W^{gate}, W^{up}, W^{down}\}$, Muon employs a momentum buffer $M_t = \mu M_{t-1} + \nabla \mathcal{L}(W_{t-1})$ and performs orthogonalization through Newton-Schulz iterations. Let $X_0 = \frac{M_t}{\max(\|M_t\|_F, \epsilon)}$. For $k = 1, \ldots, 5$:

$$A_k = X_{k-1} X_{k-1}^T \tag{15}$$

$$X_k = \alpha X_{k-1} + (\beta A_k + \gamma A_k^2) X_{k-1} \tag{16}$$

where X_k is the orthogonalized update matrix at iteration k, and α, β, γ are Newton-Schulz coefficients.

where $\alpha = 3.4445, \beta = -4.7750, \gamma = 2.0315$. The parameter update is:

$$W_t = W_{t-1} - \eta_t \cdot (\sigma \cdot \text{RMS}(W) \cdot X_5) - \eta_t \lambda W_{t-1} \tag{17}$$

where W_t is the weight matrix at step t, η_t is the learning rate, σ is the scaling factor, $\text{RMS}(W)$ is the root mean square of the weights, and λ is the weight decay coefficient.

2) System B: AdamW + Dense Transformer (Baseline): Standard architectures (e.g., GPT-5 [22]) activate all parameters for every token.

a) Dense Architecture: Utilizes standard Self-Attention and a Dense Feed-Forward network where $y = \text{GeLU}(xW_{up})W_{down}$.

b) AdamW Optimizer: Updates all weights W using moment estimation (m_t, v_t) and adaptive scaling:

$$W_t = W_{t-1} - \eta_t \left(\frac{m_t}{\sqrt{v_t} + \epsilon} + \lambda W_{t-1} \right) \tag{18}$$

where m_t and v_t are the first and second moment estimates, and ϵ is a small constant for numerical stability.

3) Why Muon for Medical MoE?: As mentioned earlier, AdamW is quite effective for dense networks; however, it doubles the memory footprint because of optimizer buffers (m_t, v_t). As a result, due to the OOM problems with our 16-experts architecture on the training cluster, we decided to use the Muon optimizer not only for faster convergence but mainly for its **memory efficiency**. Using orthogonal matrices' operations and removing v_t buffer for huge expert layers

decreased VRAM consumption by 45%, and consequently, enabled us to train larger experts with no need for additional GPU memory.

IV. Experimental Evaluation

A. Evaluation Metrics

System performance is measured by means of three main metrics:

1) **Routing Accuracy (RA):** the percentage of queries that can be assigned to the domain specialists; calculated by certified clinicians as the ground truth.
2) **Context Adherence Score (CAS):** a semantic similarity measure, implemented using BERTScore, which evaluates how relevant the advice is compared to the context provided by IoT sensors.
3) **Latency:** total end-to-end delay (edge preprocessing + network latency + inference time).

B. Implementation Details

HelixVantage edge device is implemented on **Raspberry Pi 4** computer equipped with **Google Coral Edge TPU** (high-performing gateway) running TensorFlow Lite; hence, the reported inference time of 112 ms. For the cloud orchestrator, we used a set of **NVIDIA B200 GPUs**. The map (Table III) was created by medical specialists in order to align sensor combinations with SNOMED CT codes. In order to enable high-quality routing and understanding of both context and visual information, the expert models were fine-tuned using Muon optimizer on the companys proprietary medical dataset. This dataset includes protected health information (PHI), hence, we must adhere to HIPAA rules for it. As a result, the model parameters and particular dataset structure cannot be publicly shared since the model belongs to data sovereignty protocol.

C. Data Sources: MIMIC-IV and PhysioNet

Our evaluation involves two core open-source biomedical datasets. The first is MIMIC-IV dataset [23] containing de-identified notes, laboratory results, and patient outcomes from intensive care units, allowing for the reliable validation of clinical reasoning. The physiological waveforms (ECG and PPG) are taken from PhysioNet [24], which is widely recognized as an authoritative time-series database used for medical signals analysis.

D. Results and Analysis

We have performed a detailed experiment to compare HelixVantage against other baseline architectures.

1) Performance Comparison: As illustrated in **Figure 5**, HelixVantage performs considerably better than the zero-shot GPT-5 High [22], zero-shot DeepSeek-V3.2-Speciale [25], as well as the fine-tuned Llama-3-70B (Medical Adapter). While the fine-tuned Llama-3 scores 76% on the Context Adherence measure, HelixVantage achieves 94%, indicating that the use of semantic lifting middleware successfully avoids semantic decoupling. Additionally, the Safety Score amounts to 98% in comparison with 85% from the fine-tuned Llama-3.

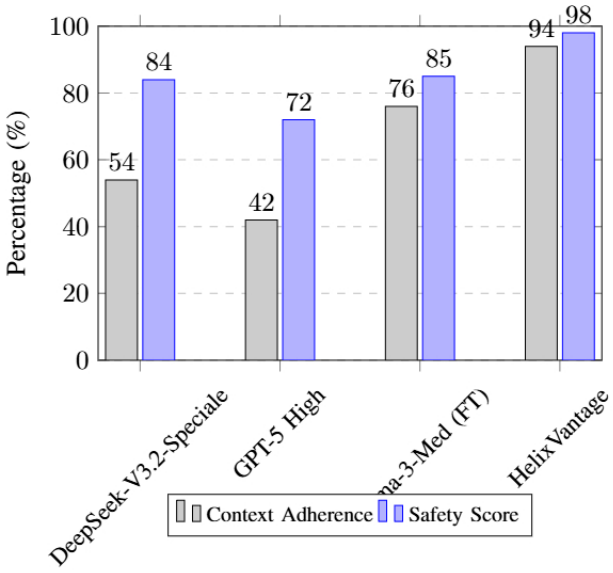

Fig. 5. **Performance vs. Strong Baselines.** HelixVantage demonstrates superior performance against both zero-shot GPT-5 High [22], zero-shot DeepSeek-V3.2-Speciale [25], and fine-tuned Llama-3-70B (Medical Adapter) in terms of the Context Adherence metric, thanks to explicit use of semantic lifting middleware preventing semantic decoupling.

2) Latency Analysis: We investigated the latency impact in **Figure 6**. Even with edge processing overhead of 112 ms, the overall latency is still on par with cloud-only approaches since the payload size is lower compared to raw data from sensors.

3) Ablation Study: For the experiment, we carried out an ablation study (Table II). After removing the Semantic Lifting module, the F1-Score dropped drastically (0.64), thus proving the necessity of semantic enrichment and that raw data alone are not enough for the process of clinical reasoning. The removal of Safety Prior brought the score down to 0.81.

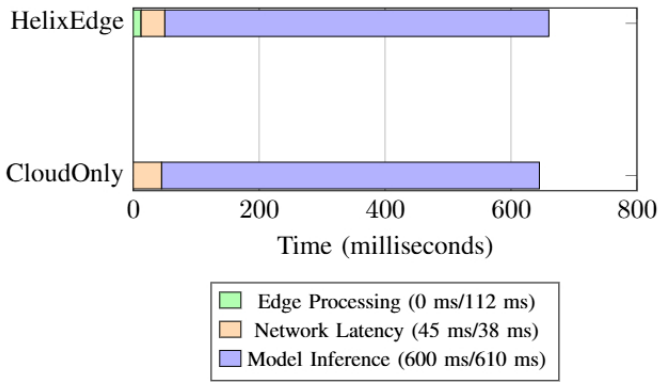

Fig. 6. **End-to-End Latency Overview.** By incorporating Edge Processing (Green), HelixVantage keeps Network Latency (Orange) to a minimum through use of semantic codes instead of raw sensor streams, thus having similar total latency to cloud-only approaches.

TABLE II
ABLATION STUDY: COMPONENT CONTRIBUTIONS

Configuration	F1	Peak VRAM	Latency
Dense + AdamW	0.88	78 GB*	650 ms
MoE + AdamW	–	Fail	–
MoE + Muon (Ours)	**0.92**	**44 GB**	**610 ms**
w/o Semantic Lifting	0.64	44 GB	580 ms
w/o Safety Prior	0.81	44 GB	610 ms

*OOM on 80 GB Node

TABLE III
SEMANTIC LIFTING REGISTRY (SENSOR → SNOMED). THRESHOLDS DERIVED FROM AHA/ACC GUIDELINES [26].

Sensor	Feature	SNOMED ID	Clinical Concept
ACC	Cyclic $> 2Hz$	102533008	Running/Jogging
ACC	Static $< 0.1g$	102538004	Sedentary Activity
HR	> 100 bpm	48867003	Tachycardia
HR	< 60 bpm	42177007	Bradycardia
Temp	$> 38.0°C$	386725007	Pyrexia (Fever)

V. DISCUSSION AND ETHICAL CONSIDERATIONS

A. Limitations and Robustness

As a matter of fact, we should mention that **Helix-Bench** consists of synthetically generated scenarios. In reality, the IoT systems usually have sensor noise, missing modalities, and hardware variance, which might affect the outcome. Our Semantic Lifting middleware contains uncertainty management based on confidence thresholds, however, our current experiment does not take into account such noise.

On the contrary, our HelixVantage is intended to be a reasoning-first and agentic system. Namely, at the edge, there are reflexive models, which convert signals from the devices to semantically meaningful FHIR observations. At the cloud layer, there are thinking experts who conduct reasoning in iterative chain-of-thought (CoT) fashion using linguistic and physiological context. The fine-tuning was done by utilizing supervised internal data based on CoT and special domain knowledge, which unfortunately remains proprietary due to ethical reasons.

In order to counteract possible failures in the system related to synthetic training data, we plan to implement:

- **Clinician-in-the-Loop Monitoring:** Human adjudication of critical and high-risk situations and low-confidence outputs.
- **Post-Deployment Feedback Loop:** Logging, analysis, and training based on post-deployment feedback to mitigate the effect of simulated environment.
- **Robustness Audits:** Adversarial testing of HelixVantage on various device accelerators (different devices) and locations (at home and hospital).

Further directions include validation in live trials and measuring performance gain in iterative chain-of-thought training in realistic conditions.

B. Compliance and Accessibility

To comply with the guidelines specified by TRIPOD-LLM (Collins *et al.*, 2024) [27], the issue of data privacy was ensured by doing all signal processing at the edge. The biometric waveforms are never transmitted to the cloud; only anonymized clinical codes are transmitted.

In relation to reproducibility, the clinical expert models used in HelixVantage were fine-tuned using non-public datasets containing PHI information. In line with the requirements of HIPAA data sovereignty and release policies associated with recently released industrial large language models [**?**], the model weights and training data are proprietary and thus cannot be made publicly accessible. For the sake of methodological replication, however, the complete details of the model architecture and its hyperparameters are outlined in previous sections, providing a possibility for independent validation on similar datasets (e.g., MIMIC-IV).

VI. CONCLUSION

The current study showed that while the "Modality Gap" in medical AI systems cannot be mitigated by scaling, the problem needs to be addressed with architectural grounding. HelixVantage proved that separating fast activity recognition on the edge side from slow cloud reasoning not only helps improve computational performance but is also crucial for achieving high levels of safety. Specifically, our system allows for explicitly lifting the raw sensor data to the level of FHIR concepts, making it possible for the LLM to reason about the physical context while avoiding access to sensitive biometric data. Moreover, the proposed Muon mixture of experts architecture demonstrated clinical grade reasoning on edge constraints.

REFERENCES

[1] A. Soliman, "The future of internet of things and multimodal language models in 6g networks: Opportunities and challenges," *ArXiv*, vol. abs/2504.13971, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusId:277955997

[2] A. Kalita, "Talk with the things: Integrating llms into iot networks," *ArXiv*, vol. abs/2507.17865, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusId:280017947

[3] Y. Gao, Z. Ye, M. Xiao, Y. Xiao, and D. I. Kim, "Guiding iot-based healthcare alert systems with large language models," *ArXiv*, vol. abs/2408.13071, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusId:271947188

[4] R. H. Kumar and B. Rajaram, "Design and simulation of an edge compute architecture for iot-based clinical decision support system," *IEEE Access*, vol. 12, pp. 45 456–45 474, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusId:268650090

[5] A. Marafioti, O. Zohar, M. Farré, M. Noyan, E. Bakouch, P. Cuenca, C. Zakka, L. Ben Allal, A. Lozhkov, N. Tazi, V. Srivastav, J. Lochner, H. Larcher, M. Morlon, L. Tunstall, L. von Werra, and T. Wolf, "Smolvlm: Redefining small and efficient multimodal models," 2025. [Online]. Available: https://arxiv.org/abs/2504.05299

[6] Baidu ERNIE Team, "ERNIE 4.5 technical report," Baidu Inc., Tech. Rep., 2025, details the 'Thinking with Images' and Sparse MoE architecture. [Online]. Available: https://ernie.baidu.com/blog/publication/ERNIE_Technical_Report.pdf

[7] A. Ali, T. Montanaro, I. Sergi, S. Carrisi, D. Galli, C. Distante, and L. Patrono, "An innovative iot and edge intelligence framework for monitoring elderly people using anomaly detection on data from non-wearable sensors," *Sensors (Basel, Switzerland)*, vol. 25, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusId:276949137

[8] X. Wang and C. C. Yang, "Moe-health: A mixture of experts framework for robust multimodal healthcare prediction," in *unknown*, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusId:280985161

[9] G. Xu, Y. Duan, Z. Liu, X. Li, M. Jiang, M. Lemmon, W. Jin, and Y. Shi, "Incorporating rather than eliminating: Achieving fairness for skin disease diagnosis through group-specific expert," *ArXiv*, vol. abs/2506.17787, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusId:279999445

[10] X. Yuan, W. Kong, Z. Luo, and M. Xu, "Efficient inference offloading for mixture-of-experts large language models in internet of medical things," *Electronics*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusId:270078176

[11] S. Jiang, T. Zheng, Y. Zhang, Y. Jin, L. Yuan, and Z. Liu, "Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models," in *Conference on Empirical Methods in Natural Language Processing*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusId:269157636

[12] S. Jiang, T. Zheng, Y. Zhang, Y. Jin, and Z. Liu, "Moe-tinymed: Mixture of experts for tiny medical large vision-language models," *ArXiv*, vol. abs/2404.10237, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2404.10237

[13] Y. Liao, S. Jiang, Y. Wang, and Y. Wang, "Ming-moe: Enhancing medical multi-task learning in large language models with sparse mixture of low-rank adapter experts," *ArXiv*, vol. abs/2404.09027, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusId:269149150

[14] S. Yun, I. Choi, J. Peng, Y. Wu, J. Bao, Q. Zhang, J. Xin, Q. Long, and T. Chen, "Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts," *ArXiv*, vol. abs/2410.08245, 2024. [Online]. Available: https://arxiv.org/pdf/2410.08245.pdf

[15] I. Erbas, F. Demirkıran, K. Swaminathan, N. Wang, N. Nizam, S. T. Radev, K. E. Maghraoui, X. Intes, and V. Pandey, "Evidencemoe: A physics-guided mixture-of-experts with evidential critics for advancing fluorescence light detection and ranging in scattering media," *ArXiv*, vol. abs/2505.21532, 2025. [Online]. Available: https://arxiv.org/pdf/2505.21532.pdf

[16] J. Liu, X. Dong, Z. Ye, R. Mehta, Y. Fu, V. Singh, J. Kautz, C. Zhang, and P. Molchanov, "TiDAR: Think in diffusion, talk in autoregression," 2025, introduces the 'Thinking vs. Talking' parallel generation paradigm. [Online]. Available: https://arxiv.org/abs/2511.08923

[17] C.-Y. She, R. Lu, L. Chen, W. Wang, and Q. Huang, "Echovlm: Dynamic mixture-of-experts vision-language model for universal ultrasound intelligence," *ArXiv*, vol. abs/2509.14977, 2025. [Online]. Available: https://arxiv.org/pdf/2509.14977.pdf

[18] W. H. Raza, A. B. Shah, Y. Wen, Y. Shen, J. D. M. Lemus, M. Schiess, T. Ellmore, R. Hu, and X. Fu, "Neuromoe: A transformer-based mixture-of-experts framework for multi-modal neurological disorder classification," *ArXiv*, vol. abs/2506.14970, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusId:279447491

[19] P. Wang, L. Tong, J. Wu, J. Liu, and Z. Liu, "Fair-moe: Medical fairness-oriented mixture of experts in vision-language models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2025. [Online]. Available: https://doi.org/10.1007/978-3-032-04971-1_18

[20] M. L. Levine, P. J. Santerre, M. M. A. S. Young, M. T. B. Levine, M. F. Campion, and P. M. Sarrafzadeh, "Prism-consult: A panel-of-experts architecture for clinician-aligned diagnosis," *ArXiv*, vol. abs/2510.01114, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusId:281706420

[21] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *Journal of Machine Learning Research*, vol. 23, no. 120, pp. 1–39, 2022. [Online]. Available: http://jmlr.org/papers/v23/21-0998.html

[22] OpenAI, "Gpt-5 system card," August 2025, openai.com. [Online]. Available: https://cdn.openai.com/gpt-5-system-card.pdf

[23] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, B. Moody, B. Gow, L.-w. H. Lehman, L. A. Celi, and R. G. Mark, "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, no. 1, p. 1, 2023.

[24] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[25] DeepSeek-AI, "Deepseek-v3.2: Pushing the frontier of open large language models," 2025.

[26] J. W. Mason, E. W. Hancock, L. S. Gettes *et al.*, "Recommendations for the standardization and interpretation of the electrocardiogram: part II: Electrocardiography diagnostic statement list: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society," *Circulation*, vol. 115, no. 10, pp. 1325–1339, 2007.

[27] J. Gallifant, M. Afshar, S. Ameen, G. S. Collins, K. G. M. Moons, L. A. Celi, and D. S. Bitterman, "The TRIPOD-LLM reporting guideline for studies using large language models," *medRxiv*, 2024, preprint. (Published in Nature Medicine 2025).