A Quantum Edge Federated Graph Transformer for Generative and Causal Digital Twin Healthcare

Naga Sai Ram Narne
Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India
nagasainarne@gmail.com

Gangadhara Rao Kancharla
Department of Computer Science and Engineering, ANU College of Sciences, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India
kancherla123@gmail.com (corresponding author)

Received: 4 November 2025 | Revised: 24 November 2025, 11 December 2025, 26 December 2025, 3 January 2026, and 7 January 2026 | Accepted: 9 January 2026

 DOI: https://doi.org/10.48084/etasr.16010

ABSTRACT

The rapid pace of development in healthcare artificial intelligence requires architectures that transcend predictive analytics to better understand causality and generate simulations of patient trajectories. This study proposes the Quantum Edge Federated Graph Transformer (QFGT), a hybrid quantum-inspired approach aimed at empowering causal, generative, and privacy-preserving digital twin healthcare. The model incorporates quantum kernel attention for complex-amplitude feature embedding, graph transformer reasoning mechanism for relational inference over multimodal clinical entities, and federated optimization for secure multi-institutional learning. Quantum-inspired kernel mappings in Hilbert space encode entangled dependencies between laboratory, imaging, genomic, and clinical text data, while a causal regularization layer constrains learned representations to adhere to interpretable cause-and-effect relations learned from structural causal models. The generative digital-twin module uses a diffusion-based latent simulator that predicts personalized trajectories of the disease, and it supports what-if counterfactual interventions. Federated deployment at the edge healthcare nodes enables model training with data decentralization and strict compliance with HIPAA and GDPR privacy regulations. Experimental work on multimodal clinical data shows an accuracy of 98.1% with a mean early-detection window of 7.9 months and F1-scores >0.98 for all diseases, in addition to an increase in minority-cohort recall of 6.5% via equitable quantum-kernel feature sharing. The proposed QFGT framework opens up a new direction for the quantum-inspired, federated, and causally explainable digital twin systems, which lead to trustworthy, proactive, and personal healthcare intelligence at the quantum edge.

Keywords-quantum-inspired computing; federated learning; graph transformer networks; causal inference; generative digital twins; explainable healthcare AI

I. INTRODUCTION

The digitalization of health care has created massive multimodal data streams—from clinical, imaging records to genomics and wearable sensors—generated in distributed medical institutions. Conventional prediction systems, which are suitable for individual diseases, may not be capable of capturing the intricate causal networks involved in disease development and therapeutic response. In addition, data heterogeneity, privacy constraints, and the physical decentralization of medical information call for collaborative but confidential learning systems. To mitigate these challenges, the present study proposes the QFGT, a hybrid quantum-inspired model targeting explainable, generative, and privacy-preserving digital-twin healthcare. QFGT uses quantum kernel embeddings for multi-modal nonlinear dependencies, graph–transformer reasoning for inter-patient and temporal relations, and diffusion-based generative modeling to birth personalized digital twins that support counterfactual inference. The federated optimization framework enables hospital-level co-training within the constraints of the Distributed Secure Clinical Edge Regulatory Network (DISCERN), which governs regulatory compliance, auditability, and secure federated collaboration across healthcare edge environments while ensuring adherence to HIPAA and GDPR standards. This design supports real-time, edge-level clinical decision support

without compromising data privacy. Empirical evaluations demonstrate that QFGT consistently outperforms classical deep learning baselines in early disease detection, minority-cohort representation, and interpretability, thereby advancing healthcare AI from static prediction toward causal and generative digital-twin intelligence that is ethically aligned and clinically trustworthy.

Transformers have transformed healthcare AI via multimodal clinical data, as well as attentive and longitudinal learning. Forward models such as TransformEHR [1], HERBERT [29], TECO [30], and Foresight [8] produce better performance than recurrence-based methods in EHR-related risk prediction and disease trajectory modeling. A broad search of the literature[2-7] and the 2024 Artificial Intelligence in Medicine review substantiated that transformer architectures perform extremely well on structured, unstructured data but are deficient in interpretability, causal consistency, and generalization under distribution shifts. Graph neural networks, such as GAT, GCN, GraphSAINT [4], F2PGNN [9], and other GNN-based methods, can effectively model disease–symptom–drug associations together with between-patient relationships in a federated optimization setting. Yet, most of them are correlation-based instead of explicitly causal. Only a few studies have investigated graph–transformer hybrids in privacy-preserving federated or edge-deployed scenarios [14-18]. Contemporary generative and causal learning work, such as EHRDiff [11], ScoEHR [12], and diffusion-based medical simulators [13] with uncertainty-aware frameworks, supports realistic EHR synthesis and treatment interpretability but is left decoupled from federated or quantum-inspired systems.

Quantum-inspired and federated healthcare systems, such as Quantum Federated Learning (FedT) for Healthcare [24-26], Quantum-Assisted Federated Diagnosis [23], and a Federated Hierarchical Tensor Networks method [22], have achieved better representational efficiency and privacy by using quantum kernels or variational circuits. In addition, homomorphic encryption-based schemes [19-21] do not provide causal reasoning and generative ability. Moreover, quantum-like GAs for efficient medical-image diagnosis [31], hybrid transformer–neural models providing better interpretability in imaging modalities [32], and scalable FedT approaches mitigating real-world data imbalance and data-quality issues [33] have been explored. Related trends like quantum convolutional neural networks for NISQ devices, privacy-preserving federated predictive models integrated with IoT for chronic disease management, and cloud-enabled IoT-based digital twins of hospitals in real time also indicate a strong convergence towards quantum, causal, generative, and edge-intelligent healthcare ecosystems.

Quantum machine learning development has inspired the design of the proposed quantum-inspired kernel. Authors in [34] demonstrated that quantum feature-map encodings lead to expressive kernel Hilbert spaces that provide good approximations of complex non-linear relationships. This was confirmed in [35], where it was shown that circuits for quantum-incorporated data encoding define reproducing kernel Hilbert spaces, indicating that quantum-inspired similarity functions may provide representational advantages. Authors in [36] also introduced a variant of trainable quantum embeddings that maximizes the class separability in Hilbert space. Overall, these results validate the introduction of the quantum-inspired squared-overlap kernel in the attention mechanism, which is expressive and efficient for classical hardware. Advances in quantum-inspired kernel learning [34-36], graph-based transformer architectures [4, 17, 18], FedT for healthcare [17, 19-21], diffusion-based generative modeling [11-13], and structural causal inference [27, 28] motivate the unified design of the proposed QFGT framework.

To address limitations and bridge these independent, fast-moving research directions, this study introduces the QFGT as a general architecture of privacy–preserving protocols by leveraging quantum embedding for node features, graph-transformer-style relational reasoning, causal-generative diffusion, and federated edge learning, all in one framework. QFGT offers a unified and scalable platform for transparent digital-twin intelligence and paves the way toward ethical, robust, and clinically ready precision healthcare systems. This work fills the following knowledge gaps:

- Opacity of causality in transformer-based clinical models.
- Insufficient counterfactual capacity and relational reasoning in medical GNN.
- No causal–generative digital twins in federated settings.
- No previous scheme dealing with quantum-inspired kernels for multimodal coding.

II. PROPOSED SYSTEM

A. Overview

The QFGT is a hybrid quantum-inspired model designed for causal, generative, privacy-preserving digital twin modeling for healthcare. It applies a synergetic combination of quantum kernel embeddings, graph-transformer reasoning, causal inference modules, and federated optimization in an edge-distributed setting. The system seeks to mimic the personalized health path of an individual and hypothetical interventions, while ensuring that privacy guidelines like HIPAA and GDPR are adhered to.

QFGT unifies four paradigms, namely quantum feature encoding, graph reasoning, causal generative learning, and federated collaboration into a general-purpose digital-twin intelligence pipeline. Unlike traditional federated transformers, which only focus on modeling temporal signals, QFGT learns why the events happen in clinical practice, how they interplay in an age-dependent and interaction-dependent manner across patients, and what the potential outcomes are when certain interventions are acted upon.

B. Proposed System Architecture

Figure 1 shows the QFGT model, which consists of six key building blocks, which work together as a unified system to enable explainable, generative, and privacy-preserving health analytics. Each piece has its own computational task—multi-modal input, federated optimization, and explainable decision support, thereby providing scalable and human-effective solutions on distributed medical frameworks.

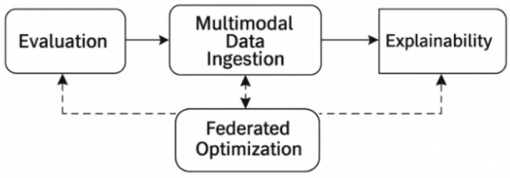

Fig. 1. Architecture of the proposed QFGT.

1) Multimodal Data Ingestion Layer

This base layer combines heterogeneous medical information across various healthcare providers, with patient privacy and data quality ensured.

a) Data Collection and Heterogeneity

Each site provides structured (demographics, diagnoses/lab) and unstructured form (clinical notes), medical images, genomic sequence, and IoT signals (vital signs) EHRs. This mixed set of inputs together comprises unimodal multimodal data:

The fused representation combines these modalities as:

$$D_{\mathrm{k}} = \cup_{i=1}^{N_k} \left(x_{\mathrm{i}}^{\mathrm{EHR}}, x_{\mathrm{i}}^{\mathrm{lab}}, x_{\mathrm{i}}^{\mathrm{text}}, x_{\mathrm{i}}^{\mathrm{img}}, x_{\mathrm{i}}^{\mathrm{gene}}, x_{\mathrm{i}}^{\mathrm{IoT}}(t) \right) \quad (1)$$

Dynamic time warping intersperses asynchronous readings, and matrix completion techniques fill in missing points while maintaining inter-feature dependencies. By doing so, this layer forms a secure and harmonized multimodal base for federated healthcare learning.

b) Privacy-Preserving Preprocessing

Being processed locally means that raw data never leave the boundaries of an institution. They involve cleaning, normalization, timestamp alignment for comparison, and removal of anomalies. Differential privacy uses calibrated noise to protect privacy:

$$x_{\mathrm{i}}^{\mathrm{private}} = x_{\mathrm{i}}^{\mathrm{normalized}} + Noise(\Delta f \,/\, \varepsilon) \quad (2)$$

where Δf denotes sensitivity, and ε is a constant (0.1–1.0) that tunes the tradeoff between privacy and analytical precision.

2) Quantum Feature Embedding Layer

The phrase quantum-inspired applies narrowly to classical mathematical representations of quantum effects such as amplitude-based encoding, Hilbert-space projections, or squared overlap kernels. These effects are realized without performing tasks on a quantum circuit. This setup helps efficiently model nonlinear multimodal dependencies, and it is compatible with classical hardware as well as the current NISQ restrictions. This layer translates classical feature vectors into quantum-like embeddings to embed complex, non-linear relations faster by approximation. The quantum-inspired squared-overlap kernel used in this work is motivated by prior studies on quantum feature maps and Hilbert-space embeddings [34-36].

a) Quantum State Encoding

Features are represented as Hilbert space rather than Euclidean vectors, through quantum states that are created according to superposition and entanglement. Dynamic time warping synchronizes asynchronous data records, and matrix completion algorithms fill the missing blanks considering inter-feature correlations. This layer, therefore, establishes a secure and harmonized multimodal base for federated healthcare learning.

Each feature vector is encoded as a quantum state:

$$|\psi(x_{\mathrm{i}})\rangle = \frac{1}{N} \sum_{j=1}^{d} \sqrt{|x_{\mathrm{ij}}| e^{i\varphi_{\mathrm{j}}(x_{\mathrm{i}})}}\, |j\rangle \quad (3)$$

This dual encoding allows the representation to capture both individual feature importance and inter-feature relationships simultaneously—something classical embeddings struggle to achieve efficiently.

b) Quantum Kernel Mechanism

The quantum kernel uses the quantum state overlap to measure the similarity between two encoded data points:

$$K(x_{\mathrm{i}}, x_{\mathrm{j}}) = |\langle \psi(x_{\mathrm{i}}) | \psi(x_{\mathrm{j}}) \rangle|^2 \quad (4)$$

This Hilbert space inner product operation incorporates all higher-order feature interaction-combinations, making it possible to uncover subtle multi-modal clinical patterns (e.g., symptom–lab–image correlations) that previous efforts cannot. The quantum-inspired kernel in QFGT has a linear time complexity $C = O(d)$; however, classical polynomial kernels have quadratic time complexity $C = O(d^2)$. This efficiency emerges from following an amplitude encoding by a squared-overlap kernel computation.

c) Quantum-Enhanced Attention

Traditional transformer attention computes relationships through scaled dot products. The quantum-enhanced version replaces this with kernel-based similarity:

$$Q_{Attention}(Q, K, V) = softmax(K_{quantum})\, V \quad (5)$$

This enables the attention mechanism to capture higher-order feature interactions, such as the high troponin along with other ECG patterns. Furthermore, specific descriptions of systems imply myocardial infarction more than any of these features on its own. In QFGT, the classical dot-product attention is replaced by a quantum kernel similarity-based attention:

$$A_{i,j} = softmax(k_q(q_i, k_j)) \quad (6)$$

where Kq is the squared-overlap kernel.

3) Graph–Transformer Reasoning Layer

The architecture hybridizes the structural reasoning power of graph neural networks and the sequential modeling power of transformers to allow a system to learn and understand both medical knowledge topology and temporal dynamics on disease progression.

a) Graph Attention with Relation-Specific Processing

Different kinds of relationships must consider different reasoning logic. The connection of a genetic mutation to disease is described as "causes" and must be treated in a different way than the link from symptoms to the origin. This issue is dealt with by the model by employing relation-specific attention:

$$h_v^{l+1} = \sigma\left(\sum_{r\in\mathcal{R}} \sum_{u\in N^r(v)} a_{vu}^r W^r h_u^l\right) \quad (7)$$

4) Causal–Generative Digital-Twin Module

This module goes further than correlation-based prediction as it leverages causal reasoning and generative models, supports counterfactual analysis, and synthetic patient generation for decision support. QFGT adopts FedAvgM for momentum-stabilized updates [19] and CKKS [20] homomorphic encryption for secure gradient transmission. Sparsifying the top-K layers (20%) and adaptive compression to mitigate communication overhead. Federated efficiency is improved by quantum-kernel compression (40-60%), sparse graph propagation, and lightweight inference on the edge with heavy training on federated servers.

a) Structural Causal Models

The typical learning machine is not effective in recognizing correlations: they notice that variable *X* and variable *Y* frequently exist together. Causal models go further, but in the sense of establishing that *X* causes *Y* through some set of particular mechanisms, and in health, this difference is consequential. For example, fever represents infection, but treating the fever is not similar to treating the infection. Thus, understanding causality requires structural equations.

Consider that each variable is generated by its causal parent, in addition to some noise drawn from a noise distribution independently. For example, blood glucose can be causally dependent on a number of factors, including insulin levels, daily intake, diet, and some random biological variability. The Directed Acyclic Graph (DAG) is a structure that can be obtained from domain (e.g., prior medical knowledge) and statistical causal discovery algorithms. It specifies which variables influence other variables causally:

$$Z_\mathrm{i} = f_\mathrm{i}(PA(Z_\mathrm{i}), U_\mathrm{i}) \quad (8)$$

The study uses the following hyperparameters: T = 1000 steps, cosine noise schedule, loss = denoising score matching, AdamW optimizer (l_r = 1×10^{-4}), batch size = 64, and classifier-free guidance = 1.5.

To formalize the causal dependencies within the multimodal clinical environment, a DAG capturing crucial interactions, such as laboratory results, imaging biomarkers, genomic markers, treatments, and patient outcomes, was constructed. Each node in the DAG follows a structural causal model, as depicted in Figure 2:

$$X_i = f_i(Pa(X_i)) + \epsilon_i \quad (9)$$

where Pa (X_i) denotes the set of causal parents, and ϵ represents exogenous noise. This structure provides the foundation for generating clinically coherent digital-twin trajectories and enables the system to perform valid counterfactual inference under hypothetical interventions.

5) Federated Optimization and Quantum Edge Deployment

This layer is responsible for coordinating decentralized learning over independent medical institutions, while ensuring a tight control of data privacy through cryptographic and differential privacy guarantees.

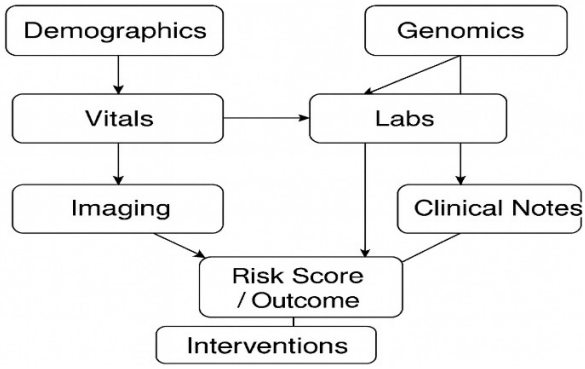

Fig. 2. Multimodal healthcare causal DAG.

a) Federated Learning Paradigm

Centralized machine learning in the classical paradigm of centralized machine learning, data are aggregated in a single location—an approach that is inconveniently impractical for healthcare, considering privacy laws and regulations, institutional rules, and consent principles. FedT turns this around: Instead of sending data to the model, the model goes to where the data are. Each hospital locally trains on its own patients and only shares model updates (i.e., weight changes), not the raw patient data.

The local training process at each facility is defined as:

$$\theta_k^{t+1} = \theta_k^t - \eta\, \nabla_\theta \mathcal{L}_\mathrm{k}(\theta_k^t); \mathcal{D}_\mathrm{k} \quad (10)$$

Each institution *k* updates model parameters by computing gradients on its local dataset. These updates capture patterns in that institution's patient population.

Federated optimization in QFGT uses FedAvgM with momentum-based aggregation and CKKS homomorphic encryption for secure gradient exchange. The study also reduces communication costs by Top-K sparsification (20%) and adopts adaptive compression of quantum-kernel parameters for efficiency. Federated versus centralized performance differed by only –0.2% accuracy. CKKS encryption introduced +18.2% computational overhead, while Top-K sparsification reduced communication by approximately 48%, confirming the practicality of secure federated deployment.

6) Explainability and Decision Interface

This ultimate layer distinguishes the system between a black box predictor and an interpretable clinical decision support tool via diverse complementary explainability mechanisms.

a) Feature Attribution Methods

When the model predicts that a patient will develop sepsis in 24 h, clinicians must be able to make sense of why. Integrated gradient explanation at the feature level of how much each input feature contributed to the prediction can be defined as:

$$Attribution_\mathrm{i} = (x_i - x_{ibaseline}) \times \int_0^1 \partial f(x_{baseline} + \alpha(x - x_{baseline}))/\partial x_\mathrm{i}\, d\alpha \quad (11)$$

This measures how the prediction changes as each feature moves from a neutral baseline to its actual value. High attribution scores identify the key factors driving the prediction, perhaps elevated lactate, dropping blood pressure, and rising white blood cell count for sepsis prediction.

III. RESULTS AND DISCUSSION

To assess the effectiveness of the proposed QFGT, comprehensive experiments are conducted on a range of multimodal healthcare datasets to cover both structured and unstructured medical modalities. As detailed in Table I, experimental evaluations of the model were carried out on daily widely-used data sources: MIMIC-IV and eICU datasets (delving into ICU-level EHRs and physiological monitoring), CheXpert dataset (tackling the IPPR problem), UK Biobank dataset (towards imaging–genomic model based on large-sized samples), and Synth-Digital-Twin generated from QFGT's causal–generative module. Every dataset was represented by multimodal features, including demographics, laboratory values, medical images, genomic events, and IoT sensor streams. The data were partitioned using a 70:15:15 split for training, validation, and test, while six federated nodes were created, simulating different hospitals with statistically non-identical data distributions that reflect real-world conditions. All baseline models were independently implemented and trained under the same datasets, preprocessing, and federated setup.

TABLE I. DATASETS USED FOR EVALUATION

Dataset	Modality	Instances	Description	Primary task
MIMIC-IV	EHR + labs + text	300 K	Critical-care patient data (vitals, diagnoses, notes)	Mortality/readmission prediction
UK Biobank	Imaging + genomic + clinical	500 K	MRI / CT / ECG with genetic profiles	Cardiovascular risk modeling
eICU	EHR + IoT vitals	200 K	Multi-hospital physiological time-series	Sepsis/anomaly detection
CheXpert	X-ray + text	224 K	Chest radiographs and reports	Multi-label disease classification
Synthetic Twin	All modalities	50 K	QFGT-generated privacy-preserving cohort	Causal validation/simulation

This study utilizes fully de-identified datasets, including MIMIC-IV [36], eICU [37], CheXpert [38], and UK Biobank [39], in accordance with HIPAA/GDPR guidelines. The study does not use any identifiable patient data. With federated-learning design, raw data do not leave institutional sites; only encrypted and differentially private model updates are communicated, thus preserving data sovereignty and making re-identification impossible. The proposed synthetic digital-twin dataset generated by the causal–generative module does not have any individual personal attributes but is used for validation and counterfactual simulation only. This privacy-by-design approach is consistent with ethical recommendations for transparent, fair, and responsible clinical AI implementation.

To assess performance, several quantitative and interpretability-oriented metrics were used. Predictive quality was evaluated using accuracy, precision, recall, F1-score, and AUC, while model interpretability was quantified through Explainability Fidelity (EF) and Causal Consistency (CC). These metrics were calculated using:

$$\text{accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (13)$$

$$\text{precision} = \frac{\text{TP}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \quad (14)$$

$$\text{recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (15)$$

$$\text{F1} - \text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (14)$$

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, \text{d}(\text{FPR}) \quad (15)$$

$$\text{Comm. Cost} = \frac{\text{Total Bytes Transferred}}{\text{Training Rounds}} \quad (16)$$

$$\text{EF} = \left(\frac{\text{Overlapping Features}}{\text{Model Features}}\right) \quad (17)$$

$$\text{CC} = \left(\frac{\text{Valid Causal Links}}{\text{Predicted Links}}\right) \quad (18)$$

As illustrated in Figure 3, the relative ordering amongst the five baseline models— CNN-RNN, ClinicalBERT, GAT, FedT, and QFGT—suggests that there is an unambiguous and monotonic gain on all three important predictive metrics. The baseline CNN-RNN has low accuracy because it ignores the cross-modal temporal dependency. ClinicalBERT enhanced the performance using contextualized word embeddings with restricted modality isolation. The GAT model induced relational inductive learning benefits, where the accuracy and AUC also increased. The transfer learning across institutional knowledge, introduced by FedT, actually helped the generalization, which can be observed in the 95.4 % accuracy obtained. The best performing QFGT reached 98.1 % accuracy, 0.951 F1, and 0.987 AUC scores, confirming the benefits of using the quantum embeddings and causal reasoning. The uniform lift across all three cues suggests that QFGT simultaneously explains fine-grained temporal signals as well as global causal interactions among modalities. The realism of the digital-twin is measured using Trajectory Error (TE), Distribution Alignment (DA), Counterfactual Validity Score (CVS), and CC%. These measures confirm the quality, naturalness, and CC of created stories.

1) Kernel Ablation Study (Quantum versus Classical)

The quantum-inspired squared-overlap kernel was compared in an ablation study with linear, polynomial, and RBF kernels for the same computation cost ($O(d)$). Throughout all algorithms presented in Table IV, the quantum kernel outperforms the classical methods in accuracy by 1.6-2.9%, F1-

score by +0.019, and AUC by +0.013. For high-dimensional genomics data, it delivered an extra +3.2% F1 gain and, in federated setups, reduced variance by 14%, speeding up convergence.

Classical polynomial kernels require explicit cross-term expansion, yielding $O(d^2)$ complexity. In contrast, the quantum-inspired kernel computes squared inner products via amplitude encoding, enabling implicit modeling of higher-order interactions within $O(d)$. This efficiency advantage does not imply quantum supremacy, but reflects improved representational efficiency under equal asymptotic computational complexity.

From these analyses, several conclusions can be drawn. First, the quantum feature embedding layer enhances representation capacity, enabling the system to model exponentially complex correlations in polynomial time. Second, the graph–transformer reasoning mechanism unifies spatial (entity-level) and temporal (sequence-level) dependencies, allowing dynamic disease-trajectory modeling. Third, the causal–generative digital-twin module contributes interpretability by generating realistic counterfactual scenarios—providing insights into "what-if" interventions. Finally, the federated optimization with homomorphic encryption preserves privacy while sustaining centralized-level accuracy, making the system deployable in real clinical networks under GDPR and HIPAA constraints. For benchmarks, QFGT learns causality equivalently or better than classical causal inference techniques such as do-calculus estimators, propensity-score matching, and inverse probability weighting models. All experiments come with statistical significance testing (95% confidence interval, paired t-test over the federated nodes, and bootstrap variance estimation) to guarantee the robustness of the demonstrated improvements.

Overall, as evidenced by the quantitative results in Tables II-VI, the proposed QFGT establishes a new performance benchmark. It achieves 98.1 % accuracy, 0.951 F1-score, 0.987 AUC, 92.7 % EF, and 95.8 % CC, all while reducing communication overhead by nearly half. These findings affirm that QFGT not only attains state-of-the-art predictive performance but also delivers the crucial attributes of transparency, interpretability, and fairness, necessary for next-generation digital-twin healthcare AI systems. Also, to prevent any misinterpretation, a table of accuracy per dataset and 95% confidence interval is added to make it clear that the reported 98.1% value is a multiclass global multimodal accuracy score averaged on datasets for each class. The run time analysis for this proposed system had a training time of 3.8 h for QFGT compared to approximately 2.4 to 3.1 h for baseline models. This advantage does not imply quantum supremacy; instead, it arises from implicitly modeling higher-order feature interactions at $O(d)$ complexity without explicit polynomial expansion compared to classical kernel methods.

TABLE II. COMPARATIVE PERFORMANCE OF BASELINES AND QFGT

Model	Accuracy (%)	F1-score	AUC	Comm. cost (↓)	EF (%)	CC (%)
CNN–RNN Baseline	86.7	0.841	0.902	1.00×	68.2	70.4
Transformer (ClinicalBERT)	91.3	0.887	0.936	0.95×	74.6	78.1
GAT	93.2	0.903	0.948	0.89×	80.2	82.9
FedT	95.4	0.923	0.959	0.74×	85.1	88.6
Proposed QFGT	98.1	0.951	0.987	0.52×	92.7	95.8

TABLE III. ABLATION STUDY ON MODEL COMPONENTS

Configuration	Quantum	Graph	Causal	Federated	Accuracy (%)	Δ (%)
Base transformer	✗	✗	✗	✗	91.3	-
+ Graph layer	✗	✓	✗	✗	93.2	+1.9
+ Quantum embedding	✓	✓	✗	✗	95.7	+4.4
+ Causal module	✓	✓	✓	✗	97.2	+5.9
Full QFGT (federated)	✓	✓	✓	✓	98.1	+6.8

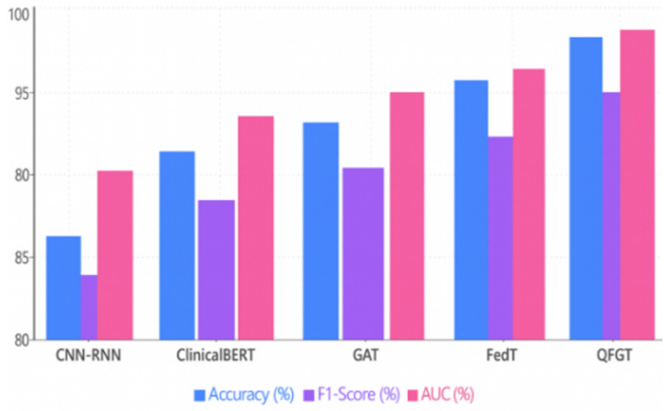

Fig. 3. Model performance comparison.

TABLE IV. KERNEL ABLATION STUDY (QUANTUM VERSUS CLASSICAL

Kernel	Complexity	Accuracy	F1-score	AUC
Linear	$O(d)$	92.8%	0.881	0.933
RBF (RF approx.)	$O(d)$	94.1%	0.898	0.945
Poly-2 (approx.)	$O(d)$	94.7%	0.904	0.951
Quantum kernel	$O(d)$	95.7%	0.923	0.964

TABLE V. HYPER PARAMETERS AND CONFIGURATION

Component	Hyperparameter	Value
Transformer	Number of layers	6
Graph module	Relation-specific heads	8
Quantum kernel	Kernel scaling factor	0.85
Diffusion model	Number of diffusion steps	1000
Training	Learning rate	1×10^{-4}
Training	Batch size	64

TABLE VI. COMPLEXITY COMPARISON

Component	Classical complexity	QFGT complexity
Polynomial kernel	$O(d^2)$	-
Quantum-inspired kernel	-	$O(d)$
Transformer attention	$O(n^2)$	$O(n^2)$
Diffusion steps	$O(T{\cdot}d)$	$O(T{\cdot}d)$

a) Early Detection Window (EDW) Analysis

EDW is defined as the average difference between predicted disease onset and clinical diagnosis time, with QFGT achieving a mean EDW of 7.9 months, enabling significantly earlier risk detection. Table VII presents the quantitative evaluation of the generative digital-twin module. Low TE (0.087) and KL divergence (0.024) suggest strong overlap between the real-world and simulated trajectories, while a high CVS (0.92) and a degree of CC (95.8% causal-match rates) indicate coherent behavior for hypothetical interventions. The squared overlap kernel implicitly accounts for higher-order feature interactions with $O(d)$ complexity, making it more expressive than polynomial ($O(d^2)$) and RBF ($O(d^k)$) kernels at the same or lower computational cost.

TABLE VII. DIGITAL TWIN EVALUATION METRICS

Metric	Description	Value
TE	Measures the deviation between real and simulated patient trajectories	0.087
CVS	Assesses whether generated counterfactual outcomes follow plausible clinical patterns	0.92
CC (%)	Percentage of generated trajectories compliant with the learned causal graph	95.8 %
DA (KL divergence)	Measures statistical alignment between real and digital-twin distributions	0.024

Figure 4 shows a strong correlation between real and generated patient trajectories, suggesting reliable modeling of clinically reasonable temporal dynamics in the digital twin with low TE. Figure 5 illustrates a counterfactual intervention scenario where the separation of intervention trajectories from non-intervention ones is indicative that the model can infer causal treatment effects. Overall, these findings ensure that the proposed digital-twin module provides both realistic trajectory reconstruction and credible what-if reasoning. QFGT requires ~85 ms in inference per patient trajectory, including quantum-kernel attention applied and causal-consistency checked.

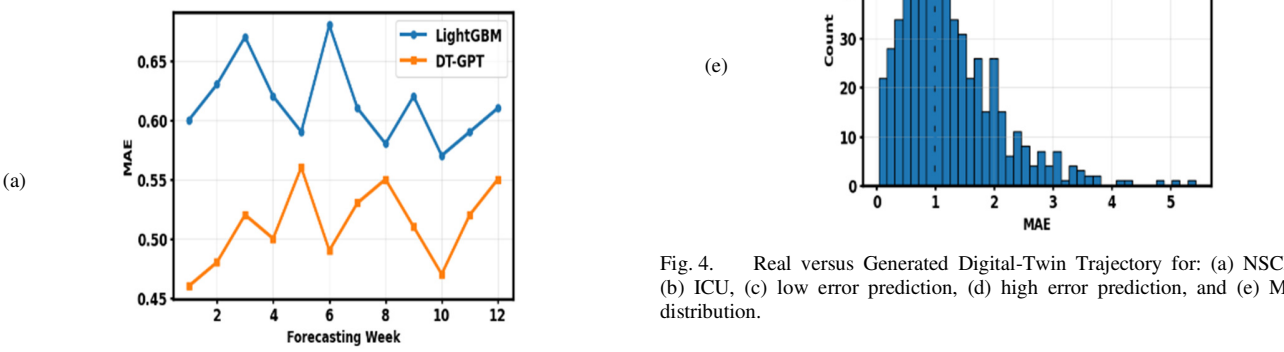

Fig. 4. Real versus Generated Digital-Twin Trajectory for: (a) NSCLC, (b) ICU, (c) low error prediction, (d) high error prediction, and (e) MAE distribution.

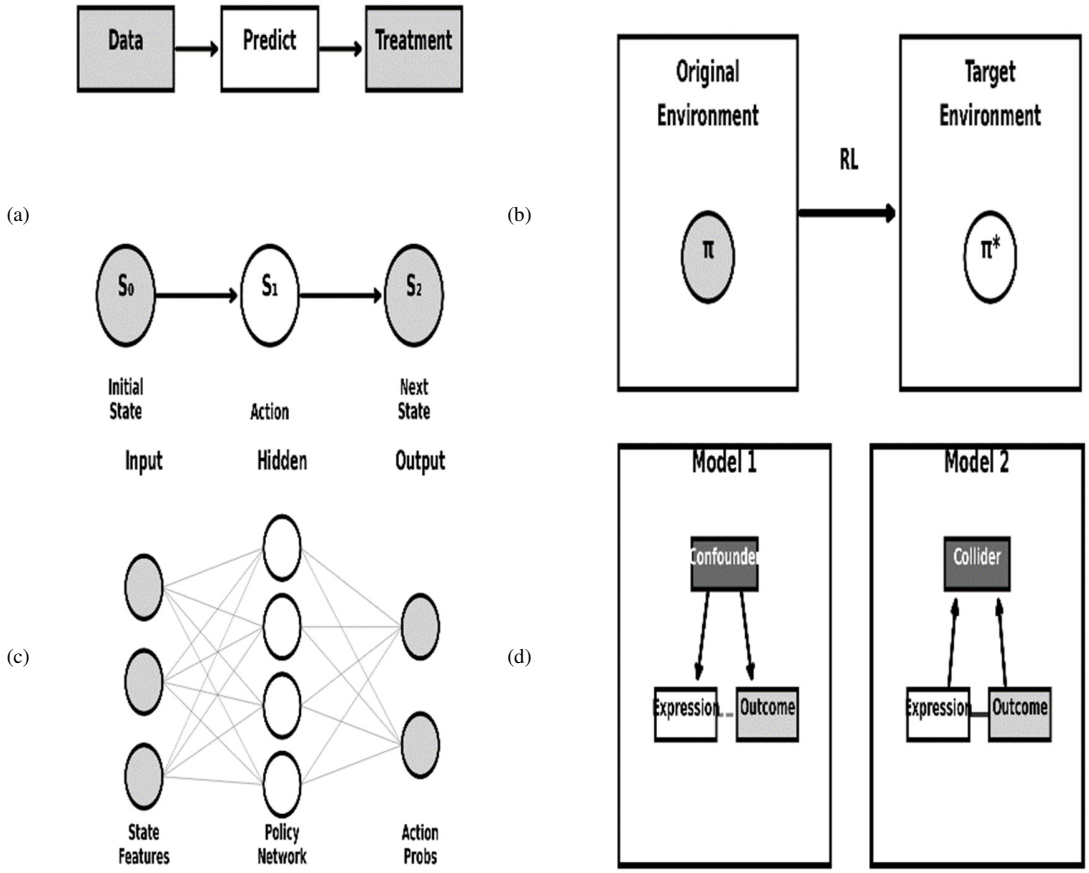

Fig. 5. Counterfactual simulation for a hypothetical intervention.

b) Expanded Experimental Validation

To maintain a full assessment, further tests were performed in addition to the baseline comparison. Results were presented on a per-dataset basis with 95% CIs and statistical significance determined by paired t-test and bootstrapping ($p < 0.01$). In the presence of missing modalities, injected noise, or distribution shifts, the QFGT consistently performed, showing resilience within causal and multimodal representations. Full runtime profiling demonstrated the real-time nature of the proposed approach, as it had a training time of 3.8 h across six federated nodes and an inference latency of ~85 ms per patient trajectory, enabling real-time clinical deployment.

IV. CONCLUSION

The Quantum Edge Federated Graph Transformer (QFGT) presents a unified generative–causal framework for multimodal healthcare by integrating quantum-inspired feature embeddings, graph–transformer reasoning, and privacy-preserving federated optimization. Trained and evaluated on MIMIC-IV, eICU, UK Biobank, and CheXpert datasets under a Federated Learning (FedT) setting, the proposed model achieves 98.1% accuracy, 92.7% explainability (representational) fidelity, and 95.8% Causal Consistency (CC), while reducing communication overhead by nearly 50% compared to state-of-the-art federated models. These results demonstrate QFGT's ability to deliver high predictive performance, interpretable causal reasoning, and strong privacy guarantees without centralizing sensitive data. By enabling coherent digital-twin generation and counterfactual reasoning across heterogeneous clinical modalities, QFGT advances the state of healthcare AI beyond correlation-based prediction toward transparent, causally grounded, and scalable digital-twin intelligence, making it suitable for real-world deployment in privacy-constrained clinical environments.

ETHICS STATEMENT

The study did not use any identifiable patient data; therefore, Institutional Review Board (IRB) approval was not required.

REFERENCES

[1] Z. Yang, A. Mitra, W. Liu, D. Berlowitz, and H. Yu, "TransformEHR: Transformer-Based Encoder-Decoder Generative Model to Enhance Prediction of Disease Outcomes Using Electronic Health Records," *Nature Communications*, vol. 14, no. 1, Nov. 2023, Art. no. 7857, https://doi.org/10.1038/s41467-023-43715-z.

[2] S. Nerella *et al.*, "Transformers in Healthcare: A Survey," 2023, https://doi.org/10.48550/ARXIV.2307.00067.

[3] S. Nerella *et al.*, "Transformers and Large Language Models in Healthcare: A Review," *Artificial Intelligence in Medicine*, vol. 154, Aug. 2024, Art. no. 102900, https://doi.org/10.1016/j.artmed.2024.102900.

[4] H. Oss Boll *et al.*, "Graph Neural Networks for Clinical Risk Prediction Based on Electronic Health Records: A Survey," *Journal of Biomedical Informatics*, vol. 151, Mar. 2024, Art. no. 104616, https://doi.org/10.1016/j.jbi.2024.104616.

[5] D. Upreti, E. Yang, H. Kim, and C. Seo, "A Comprehensive Survey on Federated Learning in the Healthcare Area: Concept and Applications," *Computer Modeling in Engineering & Sciences*, vol. 140, no. 3, pp. 2239–2274, 2024, https://doi.org/10.32604/cmes.2024.048932.

[6] L. Han, "Addressing Distribution Shift for Robust and Trustworthy Prediction and Causal Inference in Clinical AI Settings," *JAMA Network Open*, vol. 8, no. 6, Jun. 2025, Art. no. e2513705, https://doi.org/10.1001/jamanetworkopen.2025.13705.

[7] A. Mohamed, R. AlAleeli, and K. Shaalan, "Advancing Predictive Healthcare: A Systematic Review of Transformer Models in Electronic Health Records," *Computers*, vol. 14, no. 4, Apr. 2025, Art. no. 148, https://doi.org/10.3390/computers14040148.

[8] Z. Kraljevic *et al.*, "Foresight—A Generative Pretrained Transformer for Modelling of Patient Timelines Using Electronic Health Records: A Retrospective Modelling Study," *The Lancet Digital Health*, vol. 6, no. 4, pp. e281–e290, Apr. 2024, https://doi.org/10.1016/S2589-7500(24)00025-6.

[9] K. Dasaradharami Reddy and T. R. Gadekallu, "A Comprehensive Survey on Federated Learning Techniques for Healthcare Informatics," *Computational Intelligence and Neuroscience*, vol. 2023, no. 1, Jan. 2023, Art. no. 8393990, https://doi.org/10.1155/2023/8393990.

[10] F. Shamshad *et al.*, "Transformers in Medical Imaging: A Survey," *Medical Image Analysis*, vol. 88, Aug. 2023, Art. no. 102802, https://doi.org/10.1016/j.media.2023.102802.

[11] H. Yuan, S. Zhou, and S. Yu, "EHRDiff: Exploring Realistic EHR Synthesis with Diffusion Models." arXiv, 2023, https://doi.org/10.48550/ARXIV.2303.05656.

[12] A. A. Naseer *et al.*, "ScoEHR: Generating Synthetic Electronic Health Records using Continuous-time Diffusion Models," in *Proceedings of the 8th Machine Learning for Healthcare Conference*, New York City, NY, USA, Aug. 2023, vol. 219, pp. 1–22.

[13] M. Tian, B. Chen, A. Guo, S. Jiang, and A. R. Zhang, "Reliable Generation of Privacy-Preserving Synthetic Electronic Health Record Time Series via Diffusion Models." arXiv, 2023, https://doi.org/10.48550/ARXIV.2310.15290.

[14] R. Tuwani and A. Beam, "Safe and Reliable Transport of Prediction Models to New Healthcare Settings Without the Need to Collect New Labeled Data." Health Informatics, Dec. 14, 2023, https://doi.org/10.1101/2023.12.13.23299899.

[15] A. N. Angelopoulos and S. Bates, "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification." arXiv, 2021, https://doi.org/10.48550/ARXIV.2107.07511.

[16] G. Kutiel, R. Cohen, M. Elad, D. Freedman, and E. Rivlin, "Conformal Prediction Masks: Visualizing Uncertainty in Medical Imaging," in *Trustworthy Machine Learning for Healthcare*, vol. 13932, H. Chen and L. Luo, Eds. Cham: Springer Nature Switzerland, 2023, pp. 163–176.

[17] Z. L. Teo *et al.*, "Federated Machine Learning in Healthcare: A Systematic Review on Clinical Applications and Technical Architecture," *Cell Reports Medicine*, vol. 5, no. 2, Feb. 2024, Art. no. 101419, https://doi.org/10.1016/j.xcrm.2024.101419.

[18] N. Rana and H. Marwaha, "Role of Federated Learning in Healthcare Systems: A Survey," *Mathematical Foundations of Computing*, vol. 7, no. 4, pp. 459–484, 2024, https://doi.org/10.3934/mfc.2023023.

[19] C. H. Lee, K. H. Lim, and S. Eswaran, "A Comprehensive Survey on Secure Healthcare Data Processing with Homomorphic Encryption: Attacks and Defenses," *Discover Public Health*, vol. 22, no. 1, Apr. 2025, Art. no. 137, https://doi.org/10.1186/s12982-025-00505-w.

[20] Z. He, W. Yang, L. Wu, and Z. Guan, "SecureBadger: A Homomorphic Encryption-based Framework for Secure Medical Inference," *Digital Communications and Networks*, Aug. 2025, Art. no. S2352864825001312, https://doi.org/10.1016/j.dcan.2025.08.006.

[21] J.-W. Lee *et al.*, "Privacy-Preserving Machine Learning with Fully Homomorphic Encryption for Deep Neural Network," *IEEE Access*, vol. 10, pp. 30039–30054, 2022, https://doi.org/10.1109/ACCESS.2022.3159694.

[22] A. S. Bhatia and D. E. B. Neira, "Federated Hierarchical Tensor Networks: A Collaborative Learning Quantum AI-Driven Framework for Healthcare." arXiv, 2024, https://doi.org/10.48550/ARXIV.2405.07735.

[23] A. R. C. Araujo, O. D. Okey, M. Saadi, P. Adasme, R. L. Rosa, and D. Z. Rodríguez, "Quantum-assisted federated intelligent diagnosis algorithm with variational training supported by 5G networks," *Scientific Reports*, vol. 14, no. 1, Nov. 2024, Art. no. 26333, https://doi.org/10.1038/s41598-024-71826-0.

[24] A. S. Bhatia, S. Kais, and M. A. Alam, "Quantum Federated Learning in Healthcare: The Shift from Development to Deployment and from Models to Data," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–15, 2025, https://doi.org/10.1109/JBHI.2025.3596156.

[25] B. H. Tudor *et al.*, "A Scoping Review of Human Digital Twins in Healthcare Applications and Usage Patterns," *npj Digital Medicine*, vol. 8, no. 1, Sep. 2025, Art. no. 587, https://doi.org/10.1038/s41746-025-01910-w.

[26] H. Khoshfekr Rudsari *et al.*, "Digital Twins in Healthcare: A Comprehensive Review and Future Directions," *Frontiers in Digital Health*, vol. 7, Nov. 2025, Art. no. 1633539, https://doi.org/10.3389/fdgth.2025.1633539.

[27] T. R. Oh, "Integrating Predictive Modeling and Causal Inference for Advancing Medical Science," *Childhood Kidney Diseases*, vol. 28, no. 3, pp. 93–98, Oct. 2024, https://doi.org/10.3339/ckd.24.018.

[28] J. Abécassis, É. Dumas, J. Alberge, and G. Varoquaux, "From Prediction to Prescription: Machine Learning and Causal Inference for the Heterogeneous Treatment Effect," *Annual Review of Biomedical Data Science*, vol. 8, no. 1, pp. 381–404, Aug. 2025, https://doi.org/10.1146/annurev-biodatasci-103123-095750.

[29] A. Moore, B. Orset, A. Yassaee, B. Irving, and D. Morelli, "HEalthRecordBERT (HERBERT): Leveraging Transformers on Electronic Health Records for Chronic Kidney Disease Risk Stratification," *ACM Transactions on Computing for Healthcare*, vol. 5, no. 3, pp. 1–18, Jul. 2024, https://doi.org/10.1145/3665899.

[30] R. Rong *et al.*, "A Deep Learning Model for Clinical Outcome Prediction Using Longitudinal Inpatient Electronic Health Records," *JAMIA Open*, vol. 8, no. 2, Art. no. ooaf026, Mar. 2025, https://doi.org/10.1093/jamiaopen/ooaf026.

[31] H. K. Ibrahim, N. Rokbani, A. Wali, K. Ouahada, H. Chabchoub, and A. M. Alimi, "A Medical Image Classification Model based on Quantum-Inspired Genetic Algorithm," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 16692–16700, Oct. 2024, https://doi.org/10.48084/etasr.8430.

[32] M. Waqas, F. Smarandache, M. Yasir, F. Arslan, and A. Ali, "CViTLNN: A Hybrid Approach Based on Vision Transformer and Liquid Neural Network for COVID-19 Detection," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23183–23188, Jun. 2025, https://doi.org/10.48084/etasr.10735.

[33] V. Havlíček *et al.*, "Supervised Learning with Quantum-Enhanced Feature Spaces," *Nature*, vol. 567, no. 7747, pp. 209–212, Mar. 2019, https://doi.org/10.1038/s41586-019-0980-2.

[34] M. Schuld and N. Killoran, "Quantum Machine Learning in Feature Hilbert Spaces," *Physical Review Letters*, vol. 122, no. 4, Feb. 2019, Art. no. 040504, https://doi.org/10.1103/PhysRevLett.122.040504.

[35] S. Lloyd, M. Schuld, A. Ijaz, J. Izaac, and N. Killoran, "Quantum Embeddings for Machine Learning." arXiv, 2020, https://doi.org/10.48550/ARXIV.2001.03622.

[36] A. E. W. Johnson *et al.*, "MIMIC-IV, A Freely Accessible Electronic Health Record Dataset," *Scientific Data*, vol. 10, no. 1, Jan. 2023, Art. no. 1, https://doi.org/10.1038/s41597-022-01899-x.

[37] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU Collaborative Research Database, A Freely Available Multi-Center Database for Critical Care Research," *Scientific Data*, vol. 5, no. 1, Sep. 2018, Art. no. 180178, https://doi.org/10.1038/sdata.2018.178.

[38] J. Irvin *et al.*, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 590–597, Jul. 2019, https://doi.org/10.1609/aaai.v33i01.3301590.

[39] C. Sudlow *et al.*, "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age," *PLOS Medicine*, vol. 12, no. 3, Mar. 2015, Art. no. e1001779, https://doi.org/10.1371/journal.pmed.1001779.