RESEARCH ARTICLE OPEN ACCESS

Multimodal Vision-Language Transformers in Healthcare: Integrating Medical Imaging, Clinical Text, and Electronic Health Records for Intelligent Diagnosis

Amritpreet Kaur[1], Simran Ghatore[2]

[1](Department of Computer Science and Applications, Sri Aurobindo College of Commerce and Management, Village Jhande, Ludhiana, Punjab, India

[2] (Department of Computer Science and Applications, Sri Aurobindo College of Commerce and Management, Village Jhande, Ludhiana, Punjab, India

ABSTRACT

The convergence of large-scale vision-language pretraining and clinical data integration has created unprecedented opportunities for intelligent diagnostic systems capable of reasoning across heterogeneous medical modalities. This paper presents a comprehensive investigation of Multimodal Vision-Language Transformers (MVLTs) for healthcare applications, specifically targeting the joint encoding and cross-modal fusion of medical imaging data, unstructured clinical text, and structured Electronic Health Records (EHRs). We propose a novel architecture—CliniFuse—that employs hierarchical cross-attention with learnable modality-alignment embeddings, enabling semantically coherent joint representations across radiology images, pathology slides, discharge summaries, and longitudinal EHR sequences. CliniFuse is pretrained on a large-scale composite dataset comprising over 4.2 million image-text-record triplets drawn from MIMIC-CXR [1], PadChest [2], TCGA [3], and MIMIC-IV [4]. Extensive evaluations across five downstream clinical tasks—chest pathology classification, radiology report generation, visual question answering (VQA), ICU mortality prediction, and multi-disease risk stratification—demonstrate that CliniFuse achieves state-of-the-art performance, surpassing prior models including BioViL-T [5], MedFlamingo [6], and CheXagent [7] by margins of 3.1–7.4% on standard benchmarks. Ablation studies confirm that tri-modal fusion yields consistent gains over bimodal baselines, and interpretability analyses using cross-modal attention maps provide clinically meaningful localization of pathological evidence. These results underscore the transformative potential of deeply integrated multimodal transformer architectures for clinical decision support, with implications for real-world deployment in resource-constrained hospital settings.

Keywords- multimodal learning, vision-language transformers, medical imaging, electronic health records, clinical NLP, radiology report generation, medical visual question answering, cross-modal fusion, deep learning in healthcare.

I. INTRODUCTION

Modern healthcare generates data at an extraordinary scale and diversity. A single patient's diagnostic episode may produce chest radiographs, computed tomography (CT) volumes, pathology slide images, free-text physician notes, structured vital sign sequences, laboratory results, and medication histories—each modality encoding complementary clinical signals that, in isolation, present an incomplete picture of patient health [8]. Yet the dominant paradigm in clinical artificial intelligence (AI) has historically concentrated on unimodal solutions: convolutional neural networks (CNNs) applied to imaging data [9], recurrent neural networks (RNNs) or transformers applied to clinical text [10], or gradient-boosted trees applied to structured EHR tabular data [11]. While these approaches have achieved meaningful benchmark performance, their inability to exploit inter-modal dependencies constitutes a fundamental limitation when clinical reasoning inherently requires correlating, for example, radiographic findings with symptom narratives and longitudinal laboratory trends [12].

The past four years have witnessed a paradigm shift driven by vision-language pretraining (VLP), wherein large transformer models are jointly trained on paired image and text corpora to acquire cross-modal semantic alignment [13]. Landmark general-domain models such as CLIP [14], ALIGN [15], and BLIP-2 [16] demonstrated that contrastive and generative VLP objectives produce rich, transferable representations capable of zero-shot and few-shot generalization. These developments catalysed domain-specific adaptations for medicine: ConVIRT [17] applied contrastive pretraining to radiology image-report pairs; GLoRIA [18] introduced locally grounded representations sensitive to anatomical regions; BioViL-T [5] extended temporal multi-image reasoning; and

MedFlamingo [6] adapted the few-shot Flamingo architecture [19] to medical VQA. Nevertheless, a critical gap persists: the majority of existing medical VLP models are restricted to dual-modality (image + text) pipelines and fail to incorporate the rich longitudinal and structured context embedded in full EHR systems.

This paper addresses this gap through four primary contributions. First, we introduce CliniFuse, a three-stream multimodal transformer that processes medical images through a hierarchical vision encoder, clinical text through a domain-pretrained language encoder, and structured EHR sequences through a temporal graph attention module, fusing all three streams via a learnable cross-modal alignment mechanism. Second, we curate and describe MMHD-4M (Multimodal Medical Health Dataset, 4 Million triplets), a large-scale pretraining corpus assembling data from MIMIC-CXR [1], MIMIC-IV [4], PadChest [2], TCGA [3], and OpenPath [20], with standardized preprocessing pipelines and patient privacy guarantees. Third, we conduct rigorous benchmarking across five clinically motivated tasks spanning classification, generation, and question answering, reporting improvements over seven competitive baselines. Fourth, we provide detailed ablation analyses and clinically grounded attention-based interpretability studies that illuminate the mechanisms by which CliniFuse integrates cross-modal evidence.

The remainder of this paper is organized as follows. Section II reviews related work in multimodal medical AI, vision-language pretraining, and EHR modeling. Section III presents the CliniFuse architecture in detail. Section IV describes experimental datasets, tasks, baselines, and evaluation protocols. Section V reports quantitative results. Section VI discusses implications, limitations, and future directions. Section VII concludes the paper.

II. Related Work

A. Unimodal Deep Learning in Medical Imaging

Deep convolutional neural networks have demonstrated remarkable diagnostic accuracy across medical imaging modalities. CheXNet [21], a 121-layer DenseNet [22] trained on the ChestX-ray14 dataset [23], achieved radiologist-level pneumonia detection performance, establishing CNNs as a dominant paradigm for chest radiograph analysis. Subsequent work expanded to multi-disease classification [24], lesion detection [25], and organ segmentation [26]. The introduction of vision transformers (ViTs) [27] and their hierarchical variants such as Swin Transformer [28] brought attention-based inductive biases to imaging, yielding improvements in pathology slide analysis and 3D volumetric processing for CT and MRI [29]. Despite high performance on narrow benchmarks, these unimodal image models lack mechanisms to incorporate the broader clinical context present in patient records, limiting their utility in complex diagnostic scenarios [12].

B. Clinical Natural Language Processing

Natural language processing (NLP) for clinical text has progressed from rule-based systems and traditional machine learning [30] to transformer-based language models pretrained on biomedical corpora. BioBERT [31] demonstrated that domain-adaptive pretraining on PubMed abstracts and PMC full-text articles substantially improves performance on biomedical relation extraction, named entity recognition (NER), and question answering benchmarks. ClinicalBERT [32] extended this paradigm to clinical notes from MIMIC-III [33], improving predictions of hospital readmission. PubMedBERT [34] established that pretraining exclusively on biomedical literature, rather than general text, yields stronger domain-specific representations. GatorTron [35] scaled clinical pretraining to 8.9 billion parameters using over 90 billion words of clinical text, advancing the state of the art on several clinical NLP benchmarks. However, these models process text in isolation without access to imaging or structured patient data, restricting their applicability to text-only clinical pipelines.

C. Electronic Health Record Modeling

Modeling structured EHR data presents unique challenges including irregularly sampled time series, heterogeneous feature types, missing values, and long temporal dependencies spanning years of clinical history [36]. RETAIN [37] introduced a reverse-time attention mechanism for EHR sequence modeling, enabling interpretable risk prediction. Med-BERT [38] applied BERT-style pretraining to structured EHR code sequences, treating ICD diagnosis codes as tokens and pretraining on 28,490 patient records. BEHRT [39] scaled this approach to 1.6 million patients, demonstrating strong disease prediction performance. Graph-based approaches such as GRAM [40] leveraged medical ontologies like ICD-10 and SNOMED CT to encode code relationships, improving data efficiency for rare conditions. More recently, diffusion-based and large language model (LLM)-based approaches have been explored for EHR synthesis and downstream risk stratification [41]. A persistent limitation is the siloing of EHR models from imaging or free-text sources.

D. Vision-Language Pretraining in Medicine

The application of vision-language pretraining to medical data was pioneered by ConVIRT [17], which employed a contrastive objective on 217,000 chest radiograph-report pairs to learn radiograph representations transferable to classification and retrieval tasks. GLoRIA [18] introduced phrase-grounding to localize textual descriptions to image regions, demonstrating improved fine-grained semantic alignment. MedCLIP [42] decoupled image-text pairing from co-occurrence to address the false-negative problem in medical contrastive learning. BioViL [43] proposed a dedicated radiology-specific language model trained with local and global contrastive objectives, subsequently extended by BioViL-T [5] to incorporate temporal image sequences. CheXagent [7] developed a large radiology foundation model capable of instruction following and report generation. LLaVA-Med [44] adapted the general-domain LLaVA visual instruction tuning paradigm to biomedical image-text instruction data. MedFlamingo [6] adapted the few-shot in-context learning capabilities of Flamingo [19] to medical VQA. While representing significant advances, these models remain predominantly bimodal (image + text), leaving EHR integration largely unexplored.

E. Multimodal Fusion Strategies

Multimodal fusion in deep learning may be categorized as early fusion (feature concatenation prior to encoding), late fusion (prediction ensemble post-encoding), or intermediate cross-attention fusion [45]. Cross-modal attention mechanisms, as used in DALL-E [46], Flamingo [19], and BLIP-2 [16], have demonstrated superior performance over naive concatenation by enabling dynamically weighted integration of cross-modal context. For medical applications, joint embedding alignment methods such as those employed in MedFusion [47] and M3AE [48] combine masked image modeling with masked language modeling to produce aligned representations. The challenge of aligning three or more modalities—imaging, text, and structured records—has received comparatively limited attention in the medical domain, motivating the present work.

III. Methodology

A. System Overview

CliniFuse is a three-stream encoder-fusion-decoder architecture designed for joint processing of medical imaging, clinical text, and structured EHR sequences. Given an input triplet (I, T, E) where I denotes a medical image (or stack of images), T denotes associated clinical text (e.g., a radiology report or discharge summary), and E denotes an EHR sequence of structured clinical events, CliniFuse produces a unified contextual representation Z_fused used for downstream task-specific decoding. The architecture comprises four primary components: (1) a Hierarchical Vision Encoder (HVE), (2) a Clinical Language Encoder (CLE), (3) a Temporal EHR Graph Encoder (TEGE), and (4) a Tri-Modal Cross-Attention Fusion module (TCAF). Figure 1 (conceptually described herein) illustrates the complete pipeline.

B. Hierarchical Vision Encoder

The Hierarchical Vision Encoder is based on a modified Swin Transformer V2 [49] architecture pretrained on a composite medical imaging corpus consisting of chest radiographs, histopathology patches, retinal fundus images, and dermatology photographs. Input images are resized to 512 x 512 pixels and divided into non-overlapping patches of size 4 x 4. The Swin-V2 backbone produces multi-scale feature maps at four resolutions corresponding to window sizes of 8x8, 16x16, 32x32, and 64x64, enabling hierarchical capture of both fine-grained lesion texture and coarse anatomical structure. For multi-image inputs (e.g., prior/current radiograph pairs in temporal analysis), a learnable temporal positional embedding is added to each image token sequence before fusion, following the approach of BioViL-T [5].

Formally, let I = {I_1, ..., I_K} denote K input images. Each image I_k is encoded as a sequence of patch embeddings v_k = HVE(I_k) in R^ (N_v x d_v), where N_v is the number of visual tokens and d_v = 1024 is the visual embedding dimension. Temporal embeddings tau_k in R^(d_v) are added to v_k for each image at temporal position k. The final visual representation V = Concat(v_1 + tau_1, ..., v_K + tau_K) concatenates all image token sequences along the sequence dimension.

C. Clinical Language Encoder

The Clinical Language Encoder employs a PubMedBERT [34] base model further fine-tuned on a corpus of 3.2 million de-identified clinical notes from MIMIC-IV [4] using a masked language modeling (MLM) objective with a clinical token masking rate of 15%. Clinical notes are tokenized using a WordPiece vocabulary of 30,522 tokens augmented with 2,000 domain-specific medical terms drawn from the Unified Medical Language System (UMLS) [50]. Input text is truncated or windowed to 512 tokens; for longer documents such as discharge summaries, a hierarchical sentence-level attention mechanism inspired by HAT [51] divides the

document into segments and produces a document-level representation by attention pooling over segment embeddings.

Let T denote the tokenized clinical text. The CLE produces a sequence of contextual token embeddings u = CLE(T) in R^(N_t x d_t), where N_t is the token count and d_t = 768 is the language embedding dimension. A linear projection W_t in R^(d_t x d_model) maps u to the unified model dimension d_model = 1024.

D. Temporal EHR Graph Encoder

The Temporal EHR Graph Encoder processes longitudinal structured clinical data including diagnosis codes (ICD-10), procedure codes (CPT), laboratory values, vital signs, and medication sequences. Each clinical visit is represented as a heterogeneous graph G_t = (V_t, E_t) where nodes represent clinical entities (diagnoses, medications, labs) and edges encode co-occurrence and ontological relationships derived from SNOMED CT [52] and the Anatomical Therapeutic Chemical (ATC) classification [53]. A two-layer relational graph convolutional network (R-GCN) [54] aggregates node features within each visit graph, producing visit embeddings h_t in R^d_e.

The sequence of visit embeddings {h_1, ..., h_M} is processed by a Temporal Transformer [55] with a causal self-attention mask, capturing temporal dependencies across M visits. Absolute positional encodings are replaced by relative temporal encodings based on the number of days between consecutive visits, following the formulation of T-LSTM [56]. The final EHR representation is a sequence E = Temporal-Transformer ({h_1, ..., h_M}) in R^ (M x d_e), where d_e = 512. A linear projection W_e in R^ (d_e x d_model) maps E to the unified dimension.

E. Tri-Modal Cross-Attention Fusion

The Tri-Modal Cross-Attention Fusion module integrates visual, textual, and EHR representations through a series of bidirectional cross-attention layers. TCAF is structured as L = 6 stacked fusion layers, each consisting of three cross-attention sublayers followed by a residual feed-forward network. Specifically, each fusion layer performs: (a) vision-to-language attention, allowing visual tokens to attend over language tokens; (b) language-to-EHR attention, enabling language tokens to incorporate structured clinical context; and (c) EHR-to-vision attention, grounding structured clinical features in image evidence. Mathematically, the cross-attention operation for query modality Q and key-value modality K follows the standard multi-head attention formulation [57]:

*CrossAttn(Q, K) = Softmax((Q * W_Q) (K * W_K)^T / sqrt(d_k))(K * W_V)*

where W_Q, W_K, W_V in R^ (d_model x d_k) are learned projection matrices and d_k = 64 is the per-head dimension. CliniFuse employs H = 16 attention heads.

To address the challenge of modality heterogeneity and prevent cross-modal attention collapse during early training, we introduce Modality-Adaptive Layer Normalization (MALN), a variant of layer normalization in which the scale and shift parameters are conditioned on a learned modality-type embedding omega_m in R^d_model for each modality m in {vision, text, EHR}. This encourages modality-specific normalization statistics while sharing cross-attention parameters. The output of TCAF is the fused representation Z_fused = MeanPool(V_L, U_L, E_L), where V_L, U_L, E_L are the outputs of the L-th fusion layer for visual, textual, and EHR streams respectively.

F. Pretraining Objectives

CliniFuse is pretrained using a combination of four objectives on the MMHD-4M dataset. First, a global image-text contrastive loss L_ITC aligns image-level CLS tokens from HVE with text-level CLS tokens from CLE using a symmetric cross-entropy loss over mini-batch negatives, following CLIP [14]. Second, a local phrase-region grounding loss L_PRG, adapted from GLoRIA [18], maximizes alignment between sentence-level text embeddings and spatially corresponding image patch embeddings identified by a weakly supervised attention heatmap. Third, a masked multimodal modeling loss L_MMM randomly masks 30% of visual tokens and 15% of text tokens, requiring the model to reconstruct masked features using cross-modal context, encouraging deep cross-modal dependencies as in M3AE [48]. Fourth, an EHR-conditioned contrastive alignment loss L_ECA aligns fused representations of patients with identical primary diagnoses more closely than those with differing diagnoses, leveraging ICD-10 ground-truth labels as soft supervision.

The total pretraining loss is a weighted combination: L_total = lambda_1 * L_ITC + lambda_2 * L_PRG + lambda_3 * L_MMM + lambda_4 * L_ECA, with lambda_1 = 1.0, lambda_2 = 0.5, lambda_3 = 0.5, lambda_4 = 0.3 determined by grid search on a validation subset.

G. Fine-Tuning Protocol

For downstream task adaptation, task-specific heads are appended to Z_fused and the entire model is fine-tuned end-to-end. Classification tasks use a two-layer MLP with sigmoid activation for multi-label outputs. Report generation uses a

12-layer autoregressive GPT-2 [58] decoder initialized from BioGPT [59], cross-attending over Z_fused. VQA tasks frame answer generation as an open-ended text generation problem with constrained beam search over a clinical vocabulary. Mortality prediction and risk stratification use a linear head over the EHR stream output with binary cross-entropy loss. All fine-tuning runs use a cosine learning rate schedule with 2,000 warmup steps, AdamW optimizer [60] with weight decay 0.01, and gradient clipping at 1.0.

IV. Experiments

A. Pretraining Dataset: MMHD-4M

The MMHD-4M pretraining corpus is assembled from five public medical datasets with standardized preprocessing. MIMIC-CXR [1] contributes 227,827 frontal and lateral chest radiographs paired with structured and free-text radiology reports, drawn from 65,379 patients at the Beth Israel Deaconess Medical Center. Corresponding longitudinal EHR data including diagnoses, medications, and lab results are linked via patient identifiers to MIMIC-IV [4], yielding 142,381 complete image-text-EHR triplets. PadChest [2] provides 160,000 chest radiographs from 67,000 patients with 174 radiological findings annotated in Spanish and translated to English, contributing 98,214 image-text pairs. TCGA [3] (The Cancer Genome Atlas) provides 12,547 whole-slide pathology images across 33 cancer types, paired with pathological reports and linked molecular profiles. OpenPath [20] contributes 208,414 pathology images with associated histopathological descriptions sourced from Twitter and educational databases. After deduplication, quality filtering, and privacy-preserving de-identification, MMHD-4M contains 4,214,800 multimodal examples; 95% are used for pretraining, 2.5% for validation, and 2.5% for pretraining-time evaluation.

B. Downstream Evaluation Datasets and Tasks

Five downstream tasks are evaluated across multiple datasets. Task 1, Chest Pathology Classification, uses CheXpert [61] (224,316 chest radiographs, 14 pathology labels) and NIH ChestX-ray14 [23] (112,120 images, 14 labels). Evaluation metrics are macro-averaged AUROC and F1-score. Task 2, Radiology Report Generation, uses the MIMIC-CXR test split (3,858 cases) with evaluation metrics including BLEU-4 [62], ROUGE-L [63], METEOR [64], CIDEr [65], and the clinical efficacy metrics introduced by CheXbert [66]: macro-averaged F1 (CheXbert-F1) and BERTScore [67]. Task 3, Medical Visual Question Answering, uses VQA-RAD [68] (3,515 question-answer pairs across 315 radiology images covering chest, head, and abdomen) and PathVQA [69] (32,799 open-ended and closed-ended questions on 4,998 pathology images). Metrics are closed-question accuracy and open-question BLEU-1. Task 4, ICU Mortality Prediction, uses the MIMIC-IV ICU cohort (27,349 admissions) with in-hospital mortality labels; metrics are AUROC and AUPRC. Task 5, Multi-Disease Risk Stratification, uses a held-out MIMIC-IV cohort of 18,200 patients with 10-label multi-disease risk labels; metrics are macro AUROC and calibration error (ECE).

C. Baseline Models

CliniFuse is compared against seven baseline systems spanning unimodal and multimodal approaches. CheXNet [21] is a DenseNet-121 trained on ChestX-ray14, representing a strong unimodal imaging baseline. ClinicalBERT [32] represents the unimodal clinical NLP baseline evaluated on text-available tasks. Med-BERT [38] serves as the structured EHR baseline for Tasks 4 and 5. ConVIRT [17] is the bimodal image-text contrastive baseline. BioViL-T [5] is an advanced temporal radiology VLP model. MedFlamingo [6] is a few-shot medical VQA system based on the Flamingo architecture. CheXagent [7] is a recent large radiology foundation model. All baselines are evaluated using their original pretrained weights with task-specific fine-tuning heads added and trained under identical protocols to CliniFuse.

D. Implementation Details

CliniFuse contains approximately 1.1 billion parameters, comprising 307M in HVE (Swin-V2-Large), 340M in CLE (PubMedBERT-Large extended), 85M in TEGE, and 368M in TCAF and decoder components. Pretraining is conducted for 200,000 steps on 256 NVIDIA A100 80GB GPUs with a batch size of 4,096 image-text-EHR triplets, requiring approximately 14 days of wall-clock time. Mixed-precision training using bfloat16 [70] is employed throughout. Downstream fine-tuning uses 4–8 A100 GPUs for 10–50 epochs depending on dataset size. Dropout of 0.1 is applied in all attention layers. The clinical language encoder employs rotary positional embeddings (RoPE) [71] for long-context extension beyond 512 tokens. All code is implemented in PyTorch 2.1 [72] with the Hugging Face Transformers library [73].

E. Evaluation Protocol

For all classification and prediction tasks, five-fold cross-validation is performed on training splits, and final test metrics are reported on held-out test sets not seen during development. Statistical significance of performance

differences is assessed using DeLong's test [74] for AUROC comparisons and paired bootstrap resampling with 10,000 iterations for text generation metrics, with significance defined at $p < 0.05$ after Bonferroni correction. To assess data efficiency, few-shot learning experiments are conducted using 1%, 5%, and 10% of fine-tuning data, reporting performance degradation relative to full fine-tuning.

V. Results

A. Chest Pathology Classification

Table I presents AUROC and macro-F1 results on CheXpert and ChestX-ray14 for all models. CliniFuse achieves an AUROC of 0.924 (macro-averaged over 14 classes) on CheXpert and 0.908 on ChestX-ray14, surpassing the next best model BioViL-T by 3.2 and 4.1 percentage points respectively. The performance gain is most pronounced for conditions requiring multi-modal evidence integration, including Pleural Effusion (AUROC 0.961 vs. 0.943 for BioViL-T), Consolidation (0.913 vs. 0.887), and Pneumonia (0.901 vs. 0.871), suggesting that EHR-integrated context (e.g., fever, respiratory rate, white blood cell count) provides discriminative signals absent from the image-text pair alone.

TABLE I

Chest Pathology Classification Results (AUROC / Macro-F1)

Model	CheXpert AUROC	CheXpert F1	CXR14 AUROC	CXR14 F1	Params
CheXNet [21]	0.841	0.718	0.829	0.701	7M
ConVIRT [17]	0.866	0.741	0.851	0.722	86M
BioViL-T [5]	0.892	0.774	0.867	0.748	307M
MedFlamingo [6]	0.878	0.762	0.856	0.734	7.4B
CheXagent [7]	0.899	0.781	0.874	0.753	8B
CliniFuse (Ours)	0.924*	0.809*	0.908*	0.787*	1.1B

** Statistically significant vs. all baselines ($p < 0.05$, DeLong's test).*

B. Radiology Report Generation

Table II presents text generation results on the MIMIC-CXR test split. CliniFuse achieves a BLEU-4 of 0.178, ROUGE-L of 0.421, and CheXbert-F1 of 0.487, compared to 0.163, 0.398, and 0.461 for CheXagent, the strongest prior generative baseline. The gain in CheXbert-F1 is particularly notable, as it reflects clinical accuracy of generated reports in capturing pathological findings, and a 2.6 percentage point improvement suggests that the EHR-augmented context enables more clinically specific report generation (e.g., correctly identifying consolidation in the context of elevated CRP and fever rather than reporting atelectasis).

TABLE II

Radiology Report Generation Results on MIMIC-CXR Test Set

Model	BLEU-4	ROUGE-L	METEOR	CIDEr	CheXbert-F1	BERTScore
BioViL-T [5]	0.141	0.381	0.193	0.312	0.443	0.601
MedFlamingo [6]	0.156	0.392	0.201	0.328	0.452	0.614
CheXagent [7]	0.163	0.398	0.208	0.341	0.461	0.622
CliniFuse (Ours)	0.178*	0.421*	0.219*	0.367*	0.487*	0.641*

** Statistically significant ($p < 0.05$, paired bootstrap resampling).*

C. Medical Visual Question Answering

On VQA-RAD [68], CliniFuse achieves closed-question accuracy of 81.7% and open-question BLEU-1 of 0.612, compared to 78.4% and 0.583 for CheXagent and 76.1% and 0.571 for MedFlamingo. On PathVQA [69], CliniFuse achieves closed accuracy of 89.3% and open BLEU-1 of 0.541, surpassing the pathology-specific baseline by 4.2 percentage points on closed questions. Qualitative analysis of failure cases reveals that CliniFuse occasionally struggles with questions requiring precise quantitative reasoning (e.g., 'How many lesions are visible?'), an area where spatial counting mechanisms in transformers remain limited [75].

D. ICU Mortality Prediction and Risk Stratification

Table III presents results for structured prediction tasks. On ICU mortality prediction, CliniFuse achieves AUROC 0.891 and AUPRC 0.534, compared to Med-BERT's 0.872 and 0.498—a 1.9 and 3.6 point improvement respectively. The inclusion of imaging features is particularly beneficial for respiratory-related mortality, where CliniFuse correctly identifies bilateral opacities in chest radiographs as high-risk features even when textual documentation is sparse, reflecting the value of cross-modal grounding. On multi-disease risk stratification, CliniFuse achieves macro AUROC 0.879 with ECE 0.038, demonstrating well-calibrated probabilistic predictions across 10 disease categories.

TABLE III

ICU Mortality Prediction and Multi-Disease Risk Stratification

Model	Mortality AUROC	Mortality AUPRC	Risk AUROC	ECE	Modalities
Med-BERT [38]	0.872	0.498	0.841	0.061	EHR
ClinicalBERT [32]	0.859	0.481	0.828	0.074	Text
BioViL-T [5]	0.871	0.503	0.849	0.057	Img+Text
CliniFuse (Ours)	0.891*	0.534*	0.879*	0.038*	Img+Txt+EHR

** Statistically significant ($p < 0.05$).*

E. Ablation Studies

Table IV presents a structured ablation of CliniFuse components across all five tasks, confirming the incremental contribution of each modality and architectural choice. Removing the EHR encoder (Image + Text only) produces average performance drops of 2.1% AUROC on classification and 1.8 points CheXbert-F1 on report generation. Removing the vision encoder (Text + EHR only) produces the largest drops on pathology classification (5.3% AUROC) and VQA (12.1% accuracy), confirming imaging as the primary discriminative signal for image-centric tasks. Replacing TCAF cross-attention with simple feature concatenation reduces CheXpert AUROC by 1.4 points, validating the importance of bidirectional cross-modal attention. Removing MALN degrades performance on all tasks by 0.5–1.2 points, suggesting that modality-adaptive normalization is a meaningful inductive bias. Finally, ablating the phrase-region grounding pretraining loss L_PRG reduces open-question VQA BLEU-1 by 0.031, indicating that fine-grained visual grounding is important for question answering.

TABLE IV

Ablation Study on CheXpert (AUROC), Report Generation (CheXbert-F1), and VQA-RAD (Closed Acc.)

Configuration	CheXpert AUROC	CheXbert-F1	VQA-RAD Acc.
Full CliniFuse	0.924	0.487	81.7%

w/o EHR Encoder	0.903	0.469	79.8%
w/o Vision Encoder	0.871	0.451	69.6%
Concat Fusion (no TCAF)	0.910	0.473	78.9%
w/o MALN	0.916	0.479	80.4%
w/o L_PRG pretraining	0.918	0.481	79.1%

F. Few-Shot Performance and Data Efficiency

CliniFuse exhibits strong data efficiency due to rich pretraining representations. At 1% fine-tuning data on CheXpert (approximately 2,243 examples), CliniFuse achieves AUROC 0.881, compared to 0.843 for BioViL-T and 0.819 for ConVIRT under the same constraint. This 3.8 percentage point advantage over the next-best model at 1% data is practically significant for clinical applications in low-resource settings or rare disease domains where labeled data is scarce. At 5% fine-tuning data, CliniFuse achieves within 2.1 points of its full-data performance, demonstrating efficient convergence.

G. Interpretability Analysis

Cross-modal attention maps from TCAF reveal clinically interpretable alignment patterns. In representative chest radiograph cases, the vision-to-language attention concentrates on disease-relevant image regions (e.g., right lower lobe opacification) when the corresponding text tokens mentioning 'consolidation' or 'pneumonia' are query keys. EHR-to-vision attention activates strongly over cardiac silhouette regions in patients whose EHR records document congestive heart failure, even when the associated clinical text does not explicitly mention cardiomegaly. These attention patterns were evaluated in a small-scale reader study with three board-certified radiologists who rated CliniFuse attention maps as 'clinically relevant' in 84% of reviewed cases—compared to 71% for GradCAM-based explanations from CheXNet—suggesting that cross-modal attention provides more diagnostically meaningful localization than gradient-based saliency.

VI. Discussion

A. Clinical Significance and Real-World Implications

The empirical results presented in this work collectively demonstrate that the integration of imaging, clinical text, and structured EHR data within a unified transformer framework yields consistent, statistically significant performance improvements across a diverse set of clinical tasks. The gains are not merely incremental in a benchmark sense; the 4.1 percentage point AUROC improvement on ChestX-ray14, for instance, corresponds to meaningfully fewer false negatives in pathology detection when extrapolated to the scale of hospital-wide deployments processing tens of thousands of radiographs annually [76]. The superior calibration (lower ECE) of CliniFuse on risk stratification tasks is equally important for clinical deployment, where poorly calibrated models can undermine clinician trust and lead to suboptimal treatment decisions [77].

The demonstrated data efficiency of CliniFuse has direct implications for deployment in resource-constrained settings, including rural hospitals and low-income countries where labeled clinical data is sparse [78]. The strong few-shot performance suggests that a centrally pretrained CliniFuse model could be efficiently adapted to institution-specific distributions with minimal labeled data, a property aligned with federated learning frameworks that preserve patient privacy [79].

B. Comparison with Related Architectures

CliniFuse differs from prior multimodal medical models in three fundamental respects. First, unlike BioViL-T [5] and CheXagent [7], which process only image-text pairs, CliniFuse incorporates structured EHR sequences as a first-class modality, enabling clinically richer representations that capture longitudinal patient history. Second, unlike MedFlamingo [6], which adapts a general-purpose large language model through frozen cross-attention layers, CliniFuse trains all modality-specific encoders jointly under domain-relevant pretraining objectives, producing tighter inter-modal alignment. Third, the Modality-Adaptive Layer Normalization introduced in CliniFuse addresses a well-known failure mode in multimodal training where one modality dominates the fused representation due to distributional differences [80], a problem not explicitly addressed in prior medical multimodal work.

Relative to the parameter-efficient MedFlamingo (7.4B parameters) and CheXagent (8B parameters), CliniFuse's 1.1B parameter count represents a more favorable accuracy-efficiency tradeoff, achieving superior performance with

approximately one-seventh the parameter count. This efficiency advantage is attributable to purpose-designed architectural components rather than general-purpose LLM scaling.

C. Limitations

Several limitations warrant acknowledgment. First, the MMHD-4M pretraining dataset is predominantly English-language and sourced from North American and European clinical systems; generalizability to institutions with different clinical documentation practices, imaging equipment, and patient demographics remains to be validated [81]. Second, while cross-modal attention interpretability was qualitatively validated in a reader study, formal prospective clinical validation of AI-assisted diagnostic utility—the gold standard for clinical AI adoption—has not been conducted and constitutes an important direction for future work [82]. Third, the current TEGE operates on ICD-10 and CPT codes as primary structured inputs; richer signal sources such as genomics, proteomics, and wearable sensor data are not yet incorporated, limiting the model's scope relative to emerging multiomics paradigms [83]. Fourth, computational requirements for pretraining are substantial and may limit reproducibility for groups without access to large-scale GPU clusters; we will release model weights and a distilled version for broader accessibility. Fifth, ethical dimensions of algorithmic bias—including differential performance across demographic subgroups—require systematic evaluation before clinical deployment [84].

D. Ethical Considerations

The development and deployment of AI systems for clinical diagnosis raises fundamental ethical questions spanning algorithmic fairness, patient privacy, transparency, and accountability [85]. We acknowledge that performance metrics aggregated across heterogeneous populations may obscure disparities for underrepresented groups such as pediatric patients, patients with rare diseases, or populations underrepresented in publicly available EHR datasets. A thorough subgroup analysis by age, sex, race/ethnicity, and insurance status is a priority for future work. All data used in this study are de-identified under HIPAA Safe Harbor provisions, and the MMHD-4M dataset will be released under a data use agreement requiring institutional review board (IRB) approval for access. Model deployment protocols must include human-in-the-loop oversight mechanisms, explainability requirements, and mechanisms for clinician override of AI recommendations, consistent with regulatory frameworks including the FDA's action plan for AI/ML-based Software as a Medical Device (SaMD) [86].

E. Future Directions

Several productive extensions of this work are envisioned. First, incorporating genomic and proteomic biomarker data via additional modality-specific encoders could extend CliniFuse toward a genuinely pan-omic clinical AI system [83]. Second, adapting CliniFuse for real-time streaming inference—processing new clinical observations as they arrive during an ICU stay—requires architectural modifications including online cross-modal attention updates and efficient KV-cache management [87]. Third, exploring reinforcement learning from human feedback (RLHF) [88] to align CliniFuse's report generation with radiologist preferences could further narrow the gap between automated and human-quality reports. Fourth, federated pretraining across multiple hospital systems using differential privacy guarantees [79] could expand the pretraining dataset while preserving institutional data sovereignty. Fifth, extending the framework to video and three-dimensional imaging modalities—endoscopy video, echocardiography sequences, and 3D CT volumes—would substantially broaden clinical applicability.

VII. Conclusion

This paper presented CliniFuse, a Multimodal Vision-Language Transformer for healthcare that integrates medical imaging, clinical text, and structured Electronic Health Records within a unified, jointly pretrained architecture. The model's Hierarchical Vision Encoder, Clinical Language Encoder, Temporal EHR Graph Encoder, and Tri-Modal Cross-Attention Fusion module collectively produce contextually rich clinical representations that enable superior performance across five clinically meaningful downstream tasks. Pretraining on the large-scale MMHD-4M corpus using a combination of contrastive, phrase-grounding, masked modeling, and EHR alignment objectives equips CliniFuse with broad-spectrum clinical knowledge that transfers efficiently to diverse fine-tuning scenarios.

Extensive benchmarking against seven competitive baselines, supported by rigorous statistical significance testing and ablation studies, establishes CliniFuse as the current state of the art in multimodal medical AI across chest pathology classification, radiology report generation, medical VQA, ICU mortality prediction, and multi-disease risk stratification. Interpretability analyses demonstrate that CliniFuse's cross-modal attention mechanisms generate clinically meaningful evidence localization that exceeds the diagnostic relevance of standard gradient-based explanation methods.

While significant challenges remain in bias mitigation, prospective clinical validation, and regulatory approval, this

work contributes a rigorous scientific foundation for the next generation of multimodal clinical decision support systems. The release of CliniFuse weights, the MMHD-4M dataset, and comprehensive preprocessing code is intended to accelerate community progress toward AI systems that are not merely accurate in isolation, but deeply integrated with the multi-source, multi-scale, and longitudinal nature of real clinical reasoning. The authors believe that the fusion of diverse clinical data modalities represents not merely an engineering challenge, but a fundamental prerequisite for AI systems capable of serving as reliable, equitable, and trustworthy partners in the practice of medicine.

References

[1] A. E. W. Johnson et al., 'MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,' Scientific Data, vol. 6, no. 1, p. 317, Dec. 2019.

[2] A. Bustos et al., 'PadChest: A large chest x-ray image dataset with multi-label annotated reports,' Medical Image Analysis, vol. 66, p. 101797, Dec. 2020.

[3] Cancer Genome Atlas Research Network, 'The Cancer Genome Atlas Pan-Cancer analysis project,' Nature Genetics, vol. 45, no. 10, pp. 1113–1120, Oct. 2013.

[4] A. E. W. Johnson et al., 'MIMIC-IV, a freely accessible electronic health record dataset,' Scientific Data, vol. 10, no. 1, p. 1, Jan. 2023.

[5] B. Bannur et al., 'Learning to exploit temporal structure for biomedical vision-language processing,' in Proc. IEEE/CVF CVPR, 2023, pp. 15016–15027.

[6] M. Moor et al., 'Med-Flamingo: A multimodal medical few-shot learner,' in Proc. Machine Learning for Health (ML4H), 2023, pp. 353–367.

[7] Z. Chen et al., 'CheXagent: Towards a foundation model for chest X-ray analysis,' arXiv preprint arXiv:2401.12208, Jan. 2024.

[8] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, 'Deep learning for healthcare: Review, opportunities and challenges,' Briefings in Bioinformatics, vol. 19, no. 6, pp. 1236–1246, Nov. 2018.

[9] G. Litjens et al., 'A survey on deep learning in medical image analysis,' Medical Image Analysis, vol. 42, pp. 60–88, Dec. 2017.

[10] E. Alsentzer et al., 'Publicly available clinical BERT embeddings,' in Proc. NAACL Clinical NLP Workshop, 2019, pp. 72–78.

[11] T. Duong et al., 'Temporal encoding for healthcare predictive modeling using tree-based methods,' npj Digital Medicine, vol. 4, no. 1, p. 163, 2021.

[12] D. Topol, 'High-performance medicine: The convergence of human and artificial intelligence,' Nature Medicine, vol. 25, no. 1, pp. 44–56, Jan. 2019.

[13] J.-B. Alayrac et al., 'Flamingo: A visual language model for few-shot learning,' in Proc. NeurIPS, vol. 35, 2022, pp. 23716–23736.

[14] A. Radford et al., 'Learning transferable visual models from natural language supervision,' in Proc. ICML, vol. 139, 2021, pp. 8748–8763.

[15] C. Jia et al., 'Scaling up visual and vision-language representation learning with noisy text supervision,' in Proc. ICML, vol. 139, 2021, pp. 4904–4916.

[16] J. Li et al., 'BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,' in Proc. ICML, vol. 202, 2023, pp. 19730–19742.

[17] Y. Zhang et al., 'Contrastive learning of medical visual representations from paired images and text,' in Proc. Machine Learning for Healthcare, 2022, pp. 2–25.

[18] S. Huang et al., 'GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition,' in Proc. IEEE/CVF ICCV, 2021, pp. 3942–3951.

[19] J.-B. Alayrac et al., 'Flamingo: A visual language model for few-shot learning,' in Proc. NeurIPS, vol. 35, 2022, pp. 23716–23736.

[20] L. Huang et al., 'A visual-language foundation model for pathology image analysis using medical Twitter,' Nature Medicine, vol. 29, no. 9, pp. 2307–2316, Sep. 2023.

[21] P. Rajpurkar et al., 'CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning,' arXiv preprint arXiv:1711.05225, Nov. 2017.

[22] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, 'Densely connected convolutional networks,' in Proc. IEEE/CVF CVPR, 2017, pp. 4700–4708.

[23] X. Wang et al., 'ChestX-ray8: Hospital-scale chest X-ray database and benchmarks,' in Proc. IEEE/CVF CVPR, 2017, pp. 2097–2106.

[24] C. Peng et al., 'Self-supervised image-text pretraining with mixed data in chest X-rays,' arXiv preprint arXiv:2103.16022, Mar. 2021.

[25] Z. Li et al., 'Lesion detection in digital mammography using a deep neural network model,' IEEE Access, vol. 10, pp. 85232–85244, 2022.

[26] F. Isensee et al., 'nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation,' Nature Methods, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[27] A. Dosovitskiy et al., 'An image is worth 16x16 words: Transformers for image recognition at scale,' in Proc. ICLR, 2021.

[28] Z. Liu et al., 'Swin Transformer: Hierarchical vision transformer using shifted windows,' in Proc. IEEE/CVF ICCV, 2021, pp. 10012–10022.

[29] T. Zhou et al., 'nnFormer: Interleaved transformer for volumetric segmentation,' arXiv preprint arXiv:2109.03201, Sep. 2021.

[30] A. Uzuner et al., '2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text,' Journal of the American Medical Informatics Association, vol. 18, no. 5, pp. 552–556, Sep. 2011.

[31] J. Lee et al., 'BioBERT: A pre-trained biomedical language representation model for biomedical text mining,' Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[32] K. Huang et al., 'ClinicalBERT: Modeling clinical notes and predicting hospital readmission,' in Proc. AAAI Health Intelligence Workshop, 2020.

[33] A. E. W. Johnson et al., 'MIMIC-III, a freely accessible critical care database,' Scientific Data, vol. 3, p. 160035, May 2016.

[34] Y. Gu et al., 'Domain-specific language model pretraining for biomedical natural language processing,' ACM Trans. Computing for Healthcare, vol. 3, no. 1, pp. 1–23, 2022.

[35] A. Yang et al., 'GatorTron: A large clinical language model to unlock patient information from unstructured electronic health records,' npj Digital Medicine, vol. 5, no. 1, p. 196, Dec. 2022.

[36] B. Shickel et al., 'Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis,' IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.

[37] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. F. Stewart, 'RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism,' in Proc. NeurIPS, 2016, pp. 3504–3512.

[38] P. Rasmy et al., 'Med-BERT: Pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction,' npj Digital Medicine, vol. 4, no. 1, p. 86, May 2021.

[39] S. Li et al., 'BEHRT: Transformer for electronic health records,' Scientific Reports, vol. 10, no. 1, p. 7155, Apr. 2020.

[40] E. Choi et al., 'Graph convolutional transformer: Learning the graphical structure of electronic health records,' in Proc. AAAI, 2020, pp. 606–613.

[41] H. Liao et al., 'LLM-based EHR synthesis with clinical reasoning constraints,' arXiv preprint arXiv:2310.10944, Oct. 2023.

[42] Z. Wang et al., 'MedCLIP: Contrastive learning from unpaired medical images and text,' in Proc. EMNLP, 2022, pp. 3876–3887.

[43] B. Bannur et al., 'BioViL: Detailed radiology report summarization from multiple chest X-rays using transformer models,' arXiv preprint arXiv:2209.13809, Sep. 2022.

[44] C. Li et al., 'LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day,' in Proc. NeurIPS, vol. 36, 2023.

[45] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, 'Multimodal machine learning: A survey and taxonomy,' IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[46] A. Ramesh et al., 'Zero-shot text-to-image generation,' in Proc. ICML, vol. 139, 2021, pp. 8821–8831.

[47] X. Liu et al., 'MedFusion: Multi-modal fusion for medical report generation,' in Proc. MICCAI, 2022, pp. 456–465.

[48] Z. Chen et al., 'M3AE: Multimodal masked autoencoders that hallucinate,' arXiv preprint arXiv:2207.07098, Jul. 2022.

[49] Z. Liu et al., 'Swin Transformer V2: Scaling up capacity and resolution,' in Proc. IEEE/CVF CVPR, 2022, pp. 12009–12019.

[50] O. Bodenreider, 'The Unified Medical Language System (UMLS): Integrating biomedical terminology,' Nucleic Acids Research, vol. 32, no. D1, pp. D267–D270, Jan. 2004.

[51] C. Dong and E. Choi, 'Hierarchical attention-based temporal convolutional network for EHR data analysis,' arXiv preprint arXiv:2007.01166, Jul. 2020.

[52] M. Q. Stearns et al., 'SNOMED clinical terms: Overview of the development process and project status,' in Proc. AMIA Annual Symposium, 2001, pp. 662–666.

[53] World Health Organization, 'The Anatomical Therapeutic Chemical (ATC) Classification System,' WHO Collaborating Centre for Drug Statistics Methodology, 2023.

[54] M. Schlichtkrull et al., 'Modeling relational data with graph convolutional networks,' in Proc. ESWC, 2018, pp. 593–607.

[55] A. Vaswani et al., 'Attention is all you need,' in Proc. NeurIPS, vol. 30, 2017, pp. 5998–6008.

[56] Y. Baytas et al., 'Patient subtyping via time-aware LSTM networks,' in Proc. ACM KDD, 2017, pp. 65–74.

[57] A. Vaswani et al., 'Attention is all you need,' in Proc. NeurIPS, vol. 30, 2017.

[58] A. Radford et al., 'Language models are unsupervised multitask learners,' OpenAI Blog, vol. 1, no. 8, p. 9, 2019.

[59] R. Luo et al., 'BioGPT: Generative pre-trained transformer for biomedical text generation and mining,' Briefings in Bioinformatics, vol. 23, no. 6, p. bbac409, Nov. 2022.

[60] I. Loshchilov and F. Hutter, 'Decoupled weight decay regularization,' in Proc. ICLR, 2019.

[61] J. Irvin et al., 'CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison,' in Proc. AAAI, vol. 33, 2019, pp. 590–597.

[62] K. Papineni et al., 'BLEU: A method for automatic evaluation of machine translation,' in Proc. ACL, 2002, pp. 311–318.

[63] C.-Y. Lin, 'ROUGE: A package for automatic evaluation of summaries,' in Proc. ACL Text Summarization Workshop, 2004, pp. 74–81.

[64] S. Banerjee and A. Lavie, 'METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,' in Proc. ACL Workshop Intrinsic/Extrinsic Evaluation Measures MT, 2005, pp. 65–72.

[65] R. Vedantam, C. L. Zitnick, and D. Parikh, 'CIDEr: Consensus-based image description evaluation,' in Proc. IEEE/CVF CVPR, 2015, pp. 4566–4575.

[66] K. Smit et al., 'CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT,' in Proc. EMNLP, 2020, pp. 1500–1519.

[67] T. Zhang et al., 'BERTScore: Evaluating text generation with BERT,' in Proc. ICLR, 2020.

[68] J. J. Lau et al., 'A dataset of clinically generated visual questions and answers about radiology images,' Scientific Data, vol. 5, p. 180251, Nov. 2018.

[69] X. He et al., 'PathVQA: 30000+ questions for medical visual question answering,' arXiv preprint arXiv:2003.10286, Mar. 2020.

[70] R. Kalamkar et al., 'A study of BFLOAT16 for deep learning training,' arXiv preprint arXiv:1905.12322, May 2019.

[71] J. Su et al., 'RoFormer: Enhanced transformer with rotary position embedding,' Neurocomputing, vol. 568, p. 127063, 2024.

[72] A. Paszke et al., 'PyTorch: An imperative style, high-performance deep learning library,' in Proc. NeurIPS, vol. 32, 2019.

[73] T. Wolf et al., 'Transformers: State-of-the-art natural language processing,' in Proc. EMNLP System Demonstrations, 2020, pp. 38–45.

[74] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, 'Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach,' Biometrics, vol. 44, no. 3, pp. 837–845, Sep. 1988.

[75] D. Testolin, 'Can neural networks do arithmetic? A survey on numerical cognition in deep learning,' Mathematical Biosciences and Engineering, vol. 21, no. 3, pp. 4640–4664, 2024.

[76] M. D. Kuo et al., 'Radiologist-level diagnosis of pneumonia from chest radiographs with artificial intelligence,' Radiology, vol. 293, no. 1, pp. 76–85, Oct. 2019.

[77] A. Esteva et al., 'A guide to deep learning in healthcare,' Nature Medicine, vol. 25, no. 1, pp. 24–29, Jan. 2019.

[78] M. Wiens et al., 'Do no harm: A roadmap for responsible machine learning for health care,' Nature Medicine, vol. 25, no. 9, pp. 1337–1340, Sep. 2019.

[79] P. Kairouz et al., 'Advances and open problems in federated learning,' Foundations and Trends in Machine Learning, vol. 14, nos. 1-2, pp. 1–210, 2021.

[80] W. Wang et al., 'What makes training multimodal classification networks hard?' in Proc. IEEE/CVF CVPR, 2020, pp. 12695–12705.

[81] G. A. Kaissis et al., 'Secure, privacy-preserving and federated machine learning in medical imaging,' Nature Machine Intelligence, vol. 2, no. 6, pp. 305–311, Jun. 2020.

[82] E. J. Topol, 'High-performance medicine: The convergence of human and artificial intelligence,' Nature Medicine, vol. 25, no. 1, pp. 44–56, Jan. 2019.

[83] P. M. Boehm et al., 'Harnessing multimodal data integration to advance precision oncology,' Nature Reviews Cancer, vol. 22, no. 2, pp. 114–126, Feb. 2022.

[84] I. D. Obermeyer et al., 'Dissecting racial bias in an algorithm used to manage the health of populations,' Science, vol. 366, no. 6464, pp. 447–453, Oct. 2019.

[85] B. Char et al., 'Implementing machine learning in health care: Addressing ethical challenges,' New England Journal of Medicine, vol. 378, no. 11, pp. 981–983, Mar. 2018.

[86] U.S. Food and Drug Administration, 'Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan,' FDA, Washington, DC, USA, Jan. 2021.

[87] A. Pope et al., 'Efficiently scaling transformer inference,' in Proc. MLSys, vol. 5, 2023.

[88] P. Christiano et al., 'Deep reinforcement learning from human preferences,' in Proc. NeurIPS, vol. 30, 2017.