Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Interpretable Multimodal Benchmarking Framework for Resource-Efficient Medical Image Classification

Rupesh Kumar Sah

rupesh@wrc.edu.np

Tribhuvan University, IoE, Pashchimanchal Campus

Shivendu Mishra

Rajkiya Engineering College Ambedkar Nagar

Rajiv Misra

Indian Institute of Technology Patna

T. N. Singh

Indian Institute of Technology Patna

Research Article

Keywords: Multimodal learning, Vision-language learning, Medical imaging, Wound assessment, Explainable artificial intelligence, Grad-CAM, Clinical decision support

Posted Date: October 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10972761/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

An Interpretable Multimodal Benchmarking Framework for Resource-Efficient Medical Image Classification

Rupesh Kumar Sah[1,2*], Shivendu Mishra[3], Rajiv Misra[1], T. N. Singh[1]

[1*]Indian Institute of Technology Patna, Bihta, India.
[2]Tribhuvan University, IoE, Pashchimanchal Campus, Pokhara, Nepal.
[3]Rajkiya Engineering College Ambedkar Nagar, Ambedkar Nagar, India.

*Corresponding author(s). E-mail(s): rupesh@wrc.edu.np, rupesh_2421cs03@iitp.ac.in;
Contributing authors: shivendu@recabn.ac.in; rajivm@iitp.ac.in; tnsiitb@gmail.com;

Abstract

Purpose We proposed compact multimodal models that integrate visual and textual inputs to classify medical images based on their accuracy, latency, memory footprint, and explainability in low-resource clinical settings.

Methods We conducted experiments using nine model configurations with visual backbones (EfficientNet-B0, ConvNeXt-Pico, MobileViT-XXS) and language models (DistilBERT, MiniLM, TinyBERT) on the MedPix and Wound datasets. We benchmarked them against the image-only, text-only, and BiomedCLIP models through five random seeds, measuring accuracy, parameter sizes, CPU/GPU inference time, training memory usage, and attributions using Grad-CAM, integrated gradients attributions, and fusion-gate behavior.

Results EfficientNet-B0 combined with MiniLM achieved best results in wound classification, attaining 96.8% accuracy for wound severity and 89.1% for wound type. In the MedPix dataset, multimodal models attained 99.0% accuracy for image modality classification and 86.0% for anatomical location. The best-performing lightweight models utilized significantly less parameters and memory compared to BiomedCLIP. Feature attributions identified critical areas in medical images and important text tokens in clinical text, while fusion gates redistributed importance among modalities in uncertain cases.

Conclusion A combination of images and clinical text helps to enhance classification performance as compared to single modality benchmarks and BiomedCLIP, particularly in fine-grained diagnostic tasks. The feature attribution demonstrates that these networks capture relevant clinical information, indicating that gated and cross-attention networks are viable options to be used in clinical settings.

Keywords: Multimodal learning, Vision-language learning, Medical imaging, Wound assessment, Explainable artificial intelligence, Grad-CAM, Clinical decision support

1 Introduction

Clinical decisions hardly depend only on visual diagnostics. Even though CNNs and ViTs may compete with human experts in specialized imaging [1–3], daily clinical practice involves analyzing visual evidence in combination with clinical text like radiology reports, medical records, and history [4]. It means that the architecture of clinical AI tools should be capable of processing both types of information (visual and textual) together.

Models such as BioBERT [5, 6] and PubMedBERT [7] specialized to the domain of medicine encode medical terminology and the structure of clinical narratives, whereas vision language models (VLMs) like CLIP [8] project visual and textual representations into a shared latent space. Biomedical adaptations, including BiomedCLIP [9] and MedCLIP [10], extend contrastive pretraining to medical image-text pairs. However, their high parameter counts, opaque decision boundaries, and heavy memory requirements limit practical deployment on local clinical workstations and resource-constrained edge hardware [11, 12].

These computational and interpretability constraints are pronounced in context-dependent applications such as wound assessment and case-based diagnosis (e.g., on benchmarks such as MedPix). Wound evaluation, for example, depends on jointly assessing tissue morphology, anatomical site, and reported clinical symptoms [13]. Although multimodal approaches improve classification over unimodal models, existing studies [14–16] typically evaluate a single vision-text pairing or fusion mechanism on an isolated dataset. Consequently, systematic comparisons under a unified protocol remain sparse, leaving unclear which lightweight encoder pairings and fusion designs yield effective trade-offs between predictive performance, runtime latency, and interpretability.

To examine these trade-offs, this study presents a modular benchmarking framework for multimodal clinical image classification. We evaluate nine multimodal configurations formed by pairing interchangeable vision encoders with lightweight language models, comparing them against six unimodal (vision-only and text-only) baselines and BiomedCLIP. Each configuration is benchmarked across classification metrics, parameter efficiency, CPU and GPU inference latency, and peak memory footprint, accompanied by an analysis of cross-modal gating dynamics.

The principal contributions of this work are:

- We formulate a unified evaluation framework to assess interchangeable, lightweight visual and textual backbones under standardized preprocessing, training, and evaluation protocols.
- We compare nine multimodal architectures with six unimodal and one baseline called BiomedCLIP to assess improvements due to multimodal input fusion over unimodal inputs.
- We investigate the performance-complexity trade-off of our models with respect to classification accuracy, number of parameters, latency on CPU and GPU, as well as maximum VRAM usage.
- We analyze our post-hoc explanations through Grad-CAM, integrated gradients, and adaptive fusion gate dynamics, assessing modality contributions across target classes alongside statistical significance tests.

Section 2 covers the related works in clinical multimodal learning and interpretability. The system architecture and fusion approaches are explained in Section 3. Details of the dataset, preprocessing, and training process are described in Section 4. Classification accuracy, resource usage, and interpretability results are presented in Section 5. Lastly, we conclude in Section 6.

1.1 List of Abbreviations

For ease of readability and to ensure uniformity in the paper, Table 1 contains the main abbreviations used in this manuscript.

2 Related Work

2.1 Deep Learning for Medical Image Analysis

Deep learning methods are commonly applied to automatically extract features in medical imaging datasets, including CT, MRI, ultrasound, and pathology [1, 17]. Due to the ability to learn from complex data, deep networks are now considered to be state-of-the-art methods in many clinical tasks [18] [19] [20]. CNN and ViT-based models provide high accuracy in particular applications such as diabetic retinopathy screening, skin cancer, and lung disease diagnosis [3, 21, 22].

EfficientNet [23], ConvNeXt [24], and MobileViT [25] are lightweight models that help to decrease the computational cost using a trade-off between model parameters and prediction accuracy. However, these advancements have not yet resulted in the use of any text along with the images. In actual medical settings, physicians use a combination of visual input and medical report, symptomatology, demographic information, and history to make diagnoses [20] [26] [27]. This has led to the development of multimodal approaches that can analyze both modalities simultaneously.

2.2 Medical Vision–Language Learning

In the medical field, vision-language models (VLMs) learn the alignment of the image and the text through representation space, enabling use cases like image-text matching, report generation, visual question answering, and clinical decision-making [8, 26, 28].

Table 1: List of abbreviations used in the manuscript.

Abbreviation	Description
Acc	Accuracy
CE	Cross-Entropy
CNN	Convolutional Neural Network
CP	ConvNeXt-Pico
CP-D	ConvNeXt-Pico + DistilBERT
CP-M	ConvNeXt-Pico + MiniLM
CP-T	ConvNeXt-Pico + TinyBERT
CPU	Central Processing Unit
EB	EfficientNet-B0
EB-D	EfficientNet-B0 + DistilBERT
EB-M	EfficientNet-B0 + MiniLM
EB-T	EfficientNet-B0 + TinyBERT
F1 / F_1	F_1-score
GPU	Graphics Processing Unit
IG	Integrated Gradients
M-Loc	MedPix Anatomical Location
M-Mod	MedPix Modality
MV	MobileViT-XXS
MV-D	MobileViT-XXS + DistilBERT
MV-M	MobileViT-XXS + MiniLM
MV-T	MobileViT-XXS + TinyBERT
NLP	Natural Language Processing
SOTA	State-of-the-Art
ViT	Vision Transformer
VRAM	Video Random Access Memory
W-Sev	Wound Severity
W-Type	Wound Type
WCE	Weighted Cross-Entropy
XAI	Explainable Artificial Intelligence

The domain of VLMs has matured from initial task-based architectures into universal pre-training architectures that perform both vision-language understanding and generation [26, 29]. Nowadays, medical VLMs find application in a wide variety of clinical settings from diagnosis to prognosis and multimodal segmentation [26, 28].

Biomedical foundation models, for instance, BiomedCLIP, rely on biomedical corpora for bridging the gap between the language and imaging modality [9, 12]. They enable excellent downstream performances, and in certain cases, zero-shot or minimal-shot transfer learning within applications, including but not limited to, retrieval, classification, VQA, and biomarker prediction [27, 30, 31].

However, the literature also shows clear deployment limitations for large medical VLMs [32] [33] [34]. High computational demands, privacy constraints, limited multimodal datasets, and weak interpretability continue to restrict routine clinical

adoption [32] [34] [35]. Recent benchmarking and survey work therefore argues for smaller language models, lightweight vision–language models, parameter-efficient tuning, and reproducible evaluation protocols that jointly assess performance, efficiency, and trustworthiness [33] [34].

2.3 Multimodal Clinical Decision Support

Multimodal clinical decision support is motivated by the fact that real diagnostic reasoning depends on integrating visual findings with free-text reports, clinical history, and other contextual data [27] [36] [37]. Reviews consistently argue that multimodal models better reflect clinical workflow than unimodal image classifiers because clinicians do not interpret images in isolation [26] [27] [32]. This is especially relevant for context-sensitive tasks where image appearance alone is insufficient for robust classification [20] [26] [27].

Recent architectures implement this integration with cross-attention, multimodal transformers, and adaptive fusion [19] [28] [36]. RadFuse combines a ViT image encoder with ClinicalBERT and uses stacked self-attention plus bidirectional cross-attention, improving both disease classification and image–text retrieval on chest X-ray benchmarks [36]. Earlier multimodal transformer work such as TransMed similarly showed that modeling long-range cross-modal dependencies can improve medical image classification over CNN-only baselines [19]. In broader multimodal classification, recent evidence also shows that carefully designed lightweight fusion models can match or exceed heavier foundation baselines; PairDxFusion [38] slightly outperformed fine-tuned CLIP and BiomedCLIP while training substantially faster.

Direct wound-focused multimodal literature is still limited in the attached set, which itself supports the claim that wound severity remains an underrepresented evaluation domain relative to radiology and pathology [28] [33]. Still, the task structure of wound assessment fits naturally within the multimodal diagnosis category identified in recent surveys, where symptom descriptions, clinical history, and image appearance must be interpreted jointly to determine disease severity or management-relevant categories [28]. This makes wound severity a clinically meaningful testbed for lightweight multimodal fusion even though dedicated benchmark studies remain scarce in the current literature [28] [33].

2.4 Explainable Artificial Intelligence in Medical Imaging

Accuracy alone of benchmarks does not guarantee successful adoption, acceptance, and approval of any clinical systems [32, 35]. When using medical imaging from only one modality, it is possible to use techniques such as Grad-CAM in order to point out important areas and explain the decision of the model [39]. It becomes much more complicated to explain multimodal models because the system must link visual features to clinical text rather than just pointing to pixels [32, 35].

The assessment of explainability within the domain of multimodal medical imaging is still not standardized [28, 35]. Standard metrics of text-overlap such as BLEU and ROUGE do not reflect well on the quality of the explanation since a high value may be achieved despite diagnostic errors or failure in spatial grounding [35, 40]. For a

comprehensive evaluation, it is necessary to verify facts, spatial grounding, practical utility, uncertainty reporting, and review by experts [35, 40].

In this paper, the proposed method will use the combination of the attribution methods in different modalities, i.e., region level attribution in image, token/gradient attribution in text and fusion-gate tracking in cross modality [32, 33, 35]. Such a combination enables us to verify if the model's attention/fusion mechanisms match with relevant clinical evidence [32, 35].

2.5 Critical Analysis and Research Gap

Four major gaps have been revealed in recent literature which form the motivation for conducting this research. Firstly, benchmarking is not yet comprehensive: many papers provide only one framework for a particular task or dataset, while surveys reveal that benchmarks are poorly defined, too heavily concentrated on chest X-ray, and hard to compare due to different preprocessing and evaluation methods [28] [33]. Secondly, component-wise evidence is lacking: although there is an abundance of fusion approaches, few studies investigate how visual backbones, text encoders, and fusion frameworks impact the model performance separately [20] [33] [36].

Third, resource-conscious evaluation is typically absent although deployment challenges have been repeatedly cited in relation to memory, latency, privacy, and infrastructure constraints [33] [41]. Research into lightweight clinical NLP and fine-tuning in resource-restricted settings indicates that smaller transformers can be highly competitive, or even superior to large models, when there are constraints on compute and data resources [41] [42]. Finally, multimodal explainability is poorly developed due to the fact that the current state-of-the-art explainability approaches are primarily image-based [35], while reviews call for clinically grounded, cross-modal, and expert-validated evaluation of explanation quality [35].

Accordingly, this work is best positioned not as a new foundation model, but as a modular benchmarking study that addresses these gaps by: (i) comparing multiple interchangeable visual and textual encoders under one protocol, (ii) including unimodal, multimodal, and foundation-model baselines, (iii) profiling accuracy together with latency, memory, and parameter efficiency, and (iv) examining whether post hoc explanations and fusion behavior are clinically coherent.

3 Proposed Method

The proposed architecture aims to utilize visual as well as semantic features from medical images and their corresponding clinical text in a single machine learning pipeline. This framework is specifically structured to achieve three main goals, which include (i) effective feature extraction, (ii) comparison between fusion methods, and (iii) interpretability of the clinical decision process.

3.1 Framework Overview

The schematic of the proposed multimodal architecture is illustrated in Fig. 1. The framework processes paired inputs consisting of a medical image I and clinical text T.

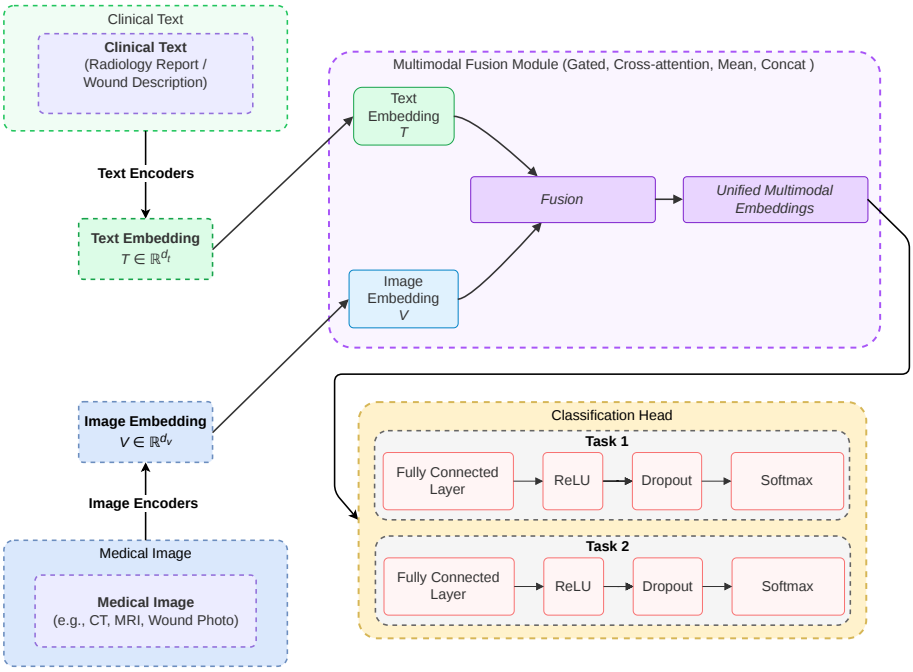


Fig. 1: Architecture of the proposed modular multimodal clinical classification and explainability framework.

Visual features are extracted using a pretrained image encoder, while textual features are captured via a biomedical transformer-based language model.

Because the raw features reside in heterogeneous spaces of differing dimensionalities, independent projection layers map them into a shared latent embedding space. These aligned representations are then integrated using one of four alternative fusion strategies: feature concatenation, adaptive gated fusion, or cross-attention fusion. The integrated multimodal vector is finally passed to a lightweight MLP classification network to compute posterior class probabilities. Post-hoc explainability modules (Grad-CAM, Integrated Gradients, and gate/attention weight analysis) run in parallel to provide local attributions for model predictions.

3.2 Problem Formulation

Let $\mathcal{D} = \{(I_i, T_i, y_i)\}_{i=1}^{N}$ denote a dataset of N paired instances, where I_i is a medical image, T_i is the corresponding clinical text, and $y_i \in \{1, \ldots, C\}$ represents the ground-truth diagnostic class label among C categories. The task is to learn a mapping function $F : (I, T) \rightarrow \hat{y}$ that minimizes prediction error.

The visual representation $\mathbf{z}_v \in \mathbb{R}^{d_v}$ is obtained through a pretrained image encoder $f_v(\cdot)$:

$$\mathbf{z}_v = f_v(I) \tag{1}$$

Similarly, the textual embedding $\mathbf{z}_t \in \mathbb{R}^{d_t}$ is extracted using a biomedical language model $f_t(\cdot)$:

$$\mathbf{z}_t = f_t(T) \tag{2}$$

To resolve dimensionality mismatch ($d_v \neq d_t$), learnable projection functions $P_v(\cdot)$ and $P_t(\cdot)$ map the unimodal embeddings into a shared latent space of dimension d:

$$\tilde{\mathbf{z}}_v = P_v(\mathbf{z}_v), \quad \tilde{\mathbf{z}}_t = P_t(\mathbf{z}_t) \tag{3}$$

where $\tilde{\mathbf{z}}_v, \tilde{\mathbf{z}}_t \in \mathbb{R}^d$.

The aligned embeddings are merged using a fusion operator $\Phi(\cdot)$ to generate the joint representation $\mathbf{z}_f$:

$$\mathbf{z}_f = \Phi\left(\tilde{\mathbf{z}}_v, \tilde{\mathbf{z}}_t\right) \tag{4}$$

The classification module $g(\cdot)$ maps $\mathbf{z}_f$ to class probabilities:

$$\hat{\mathbf{y}} = g(\mathbf{z}_f) \tag{5}$$

Model parameters are optimized by minimizing the classification loss over the training dataset:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \ell\left(y_i, \hat{y}_i\right) \tag{6}$$

where $\ell(\cdot)$ denotes the objective function detailed in Section 3.7.

3.3 Multimodal Feature Learning

To leverage rich prior knowledge and prevent overfitting on modest clinical sample sizes, our framework employs robust pretrained backbones as feature extraction branches.

3.3.1 Visual Representation Learning

To analyze how different visual modeling inductive biases impact multimodal fusion, we evaluate three distinctive, lightweight image encoders:

- **EfficientNet-B0:** A standard CNN optimized via compound scaling of depth, width, and resolution [23].
- **ConvNeXt-Pico:** A modernized pure convolutional architecture incorporating design elements from vision transformers [24].
- **MobileViT-XXS:** A hybrid CNN-transformer architecture combining local convolutions with lightweight global self-attention [25].

Given an image I, the chosen encoder yields the raw representation $\mathbf{z}_v = f_v(I) \in \mathbb{R}^{d_v}$.

3.3.2 Clinical Text Representation

Clinical text includes crucial, non-visual diagnostic cues such as anatomical site, clinical symptoms, and medical histories. Given a raw clinical text sequence $T = (t_1, t_2, \ldots, t_L)$ of length L, it is tokenized and processed by a biomedical transformer

(DistilBERT, MiniLM, or TinyBERT). The sequence of contextual token representations $\mathbf{H}_t \in \mathbb{R}^{L \times d_t}$ is generated across all token positions, from which the pooled textual descriptor $\mathbf{z}_t \in \mathbb{R}^{d_t}$ is obtained via mean pooling over unmasked token positions:

$$\mathbf{H}_t = f_t(T) \in \mathbb{R}^{L \times d_t}, \quad \mathbf{z}_t = \text{Pool}(\mathbf{H}_t) \in \mathbb{R}^{d_t} \tag{7}$$

3.4 Shared Feature Space Alignment

To establish a common basis for cross-modal interaction, the visual representation $\mathbf{z}_v \in \mathbb{R}^{d_v}$, pooled textual vector $\mathbf{z}_t \in \mathbb{R}^{d_t}$, and contextual token sequence $\mathbf{H}_t \in \mathbb{R}^{L \times d_t}$ are projected into the shared latent space $\mathbb{R}^d$ using linear projections:

$$\tilde{\mathbf{z}}_v = W_v \mathbf{z}_v + b_v \in \mathbb{R}^d \tag{8}$$

$$\tilde{\mathbf{z}}_t = W_t \mathbf{z}_t + b_t \in \mathbb{R}^d, \quad \tilde{\mathbf{H}}_t = \mathbf{H}_t W_t + b_t \in \mathbb{R}^{L \times d} \tag{9}$$

where $W_v \in \mathbb{R}^{d \times d_v}$, $W_t \in \mathbb{R}^{d \times d_t}$ are trainable projection matrices, and $b_v, b_t \in \mathbb{R}^d$ are bias vectors.

3.5 Adaptive Multimodal Fusion

Our framework evaluates four alternative fusion strategies to systematically assess feature interaction mechanics:

3.5.1 Feature Concatenation

As a parameter-efficient baseline, we concatenate the aligned unimodal vectors:

$$\mathbf{z}_f = [\tilde{\mathbf{z}}_v; \tilde{\mathbf{z}}_t] \in \mathbb{R}^{2d} \tag{10}$$

where $[\cdot\,;\,\cdot]$ represents vector concatenation.

3.5.2 Adaptive Gated Fusion

To model cases where one modality is clinically more informative than the other on a sample-by-sample basis, we define an adaptive gating vector $\mathbf{g} \in \mathbb{R}^d$:

$$\mathbf{g} = \sigma\left(W_g[\tilde{\mathbf{z}}_v; \tilde{\mathbf{z}}_t] + b_g\right) \tag{11}$$

where $W_g \in \mathbb{R}^{d \times 2d}$, $b_g \in \mathbb{R}^d$, and $\sigma(\cdot)$ is the sigmoid function. The gated multimodal representation $\mathbf{z}_f \in \mathbb{R}^d$ is computed via:

$$\mathbf{z}_f = \mathbf{g} \odot \tilde{\mathbf{z}}_v + (1 - \mathbf{g}) \odot \tilde{\mathbf{z}}_t \tag{12}$$

where $\odot$ denotes the Hadamard (element-wise) product.

3.5.3 Cross-Attention Fusion

To explicitly model fine-grained inter-modal dependencies, we employ a scaled dot-product cross-attention mechanism. Queries are derived from the projected visual embedding $\tilde{\mathbf{z}}_v \in \mathbb{R}^{1\times d}$, while Keys and Values are projected from the contextual text token sequence $\tilde{\mathbf{H}}_t \in \mathbb{R}^{L\times d}$:

$$\begin{aligned} Q &= \tilde{\mathbf{z}}_v W_Q \in \mathbb{R}^{1\times d}, \\ K &= \tilde{\mathbf{H}}_t W_K \in \mathbb{R}^{L\times d}, \quad V = \tilde{\mathbf{H}}_t W_V \in \mathbb{R}^{L\times d} \end{aligned} \tag{13}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ are learnable projection matrices. The cross-attention weight vector $\mathbf{A} \in \mathbb{R}^{1\times L}$ computes attention across the L clinical tokens:

$$\mathbf{A} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + \mathbf{M}_{\text{mask}}\right) \in \mathbb{R}^{1\times L} \tag{14}$$

where $\mathbf{M}_{\text{mask}} \in \{0, -\infty\}^{1\times L}$ masks out padding tokens, and $\sqrt{d}$ is the scaling factor. The attended multimodal representation $\mathbf{z}_f \in \mathbb{R}^d$ is then obtained as:

$$\mathbf{z}_f = \mathbf{A}V \in \mathbb{R}^d \tag{15}$$

This formulation ensures that the visual representation actively attends to the sequence of clinical findings, making the Softmax operation non-trivial across the token dimension L.

3.5.4 Element-Wise Mean Fusion

As an additional parameter-free baseline alongside feature concatenation, we evaluate the element-wise average of the aligned unimodal representations. This strategy assumes equal diagnostic weight across both modalities:

$$\mathbf{z}_f = \frac{1}{2}(\tilde{\mathbf{z}}_v + \tilde{\mathbf{z}}_t) \in \mathbb{R}^d \tag{16}$$

where the resulting fused vector $\mathbf{z}_f$ preserves the shared embedding dimensionality d.

3.6 Prediction Module

The integrated representation $\mathbf{z}_f$ is fed into a lightweight multi-layer perceptron (MLP). The network processes the vector through a non-linear hidden layer:

$$\mathbf{h} = \phi\left(W_h \mathbf{z}_f + b_h\right) \tag{17}$$

where $W_h \in \mathbb{R}^{d_h \times \dim(\mathbf{z}_f)}$, $b_h \in \mathbb{R}^{d_h}$, and $\phi(\cdot)$ is the GeLU activation function. Dropout is applied to $\mathbf{h}$ for regularization. The class logits $\mathbf{o} \in \mathbb{R}^C$ and final posterior probabilities are calculated as:

$$\mathbf{o} = W_o \mathbf{h} + b_o \tag{18}$$

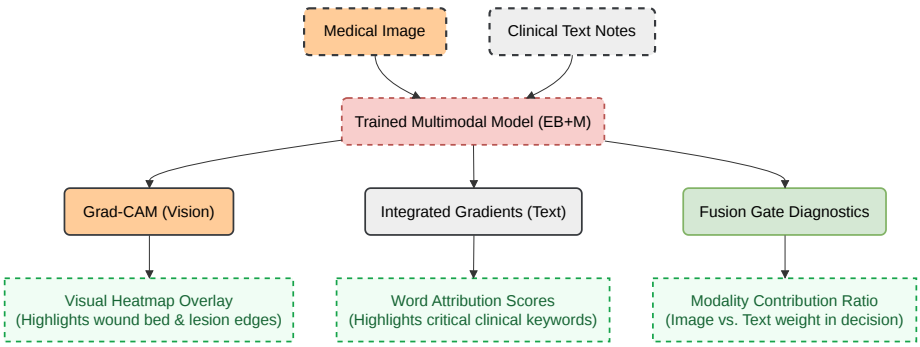


Fig. 2: **Overview of the multimodal explainability framework.** The model provides tri-level clinical interpretability: (1) **Grad-CAM** generates spatial heatmaps highlighting relevant lesion morphology; (2) **Integrated Gradients** identifies decisive clinical keywords from patient notes; and (3) **Fusion Gate Diagnostics** reveals the relative contribution ratio between image and text features for each prediction.

$$P(y = c \mid I, T) = \frac{\exp(o_c)}{\sum_{k=1}^{C} \exp(o_k)} \tag{19}$$

The final predicted label is $\hat{y} = \arg\max_c P(y = c \mid I, T)$.

3.7 Model Optimization

The entire architecture is optimized end-to-end. The backbone feature extractors are fine-tuned alongside the projection, fusion, and classification layers. The baseline optimization objective minimizes the standard categorical cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{ic} \log P_{ic} \tag{20}$$

where y_{ic} is the binary indicator for class c and P_{ic} is the predicted probability. To evaluate calibration and class imbalance handling, we also investigate Focal Loss and Label Smoothing objectives (detailed in Section 4). Parameter updates are calculated using gradient descent:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \nabla_\theta \mathcal{L} \tag{21}$$

where θ denotes the parameter set and η is the learning rate.

3.8 Explainability Framework

To secure clinical trust and make decisions transparent, the framework implements a dual-attribution post-hoc explainability suite and this module is depicted in Figure 2 :

- **Visual Interpretability:** Grad-CAM computes gradients of the target class score with respect to the final convolutional feature maps of $f_v(\cdot)$ [43], producing coarse heatmaps highlighting discriminative anatomical and pathological regions.

Table 2: Dataset characteristics, partitioning strategies, exact sample counts ($N_{\text{train}}, N_{\text{val}}, N_{\text{test}}$), and evaluation tasks for the benchmarked medical datasets.

Dataset	Partitioning Strategy	Split & Sample Counts	Evaluation Tasks
MedPix-2.0 [45]	Official corpus partitions: `descriptions_train`, `descriptions_dev`, and `descriptions_test`	80.63% / 9.61% / 9.76% ($N_{\text{total}} = 2{,}050$) $N_{\text{train}} = 1{,}653$ $N_{\text{val}} = 197$ $N_{\text{test}} = 200$	**Imaging Modality Classification (2 classes):** Categorization into clinical imaging modalities (CT vs. MRI). **Anatomical Location Classification (5 classes):** Classification across anatomical zones: Abdomen, Head, Reproductive & Urinary System (RUS), Spine & Muscles (SaM), and Thorax.
Wound-1.0[1]	Stratified partitioning based on the composite stratification key of etiology and severity stage (`type_severity`)	69.95% / 10.04% / 20.01% ($N_{\text{total}} = 1{,}564$) $N_{\text{train}} = 1{,}094$ $N_{\text{val}} = 157$ $N_{\text{test}} = 313$	**Wound Severity Classification (3 classes):** Identification of wound tissue severity stage: Mild, Moderate, Severe. **Wound Type Classification (10 classes):** Categorization across 10 wound etiologies: Abrasion, Bruises, Burn, Cut, Diabetic, Laceration, Normal, Pressure, Surgical, and Venous.

- **Textual Interpretability:** Integrated Gradients computes the path integral of the gradients of the decision outputs along the straight path from a blank baseline input to the original token embeddings of T [44], yielding precise word-level attribution scores.
- **Fusion Dynamics:** For gated fusion, we directly visualize the distribution of sample-wise gate activations $\mathbf{g}$ to quantify the exact balance of modalities. For cross-attention, we project the raw attention map from Eq. 15 to analyze inter-modal alignment.

3.9 Model Training and Reproducibility

To ensure the robustness and reproducibility of our gated multimodal fusion framework, all model configurations were trained across five independent runs using distinct random seeds (42 to 46). For fine-tuning, we unfroze the top two layers of both the vision and text backbones. We trained the models for up to 15 epochs using a cosine annealing learning rate scheduler, with an initial learning rate of 5.16×10^{-4} and a weight decay of 7.95×10^{-6}. To address class imbalance, we used weighted cross-entropy loss. We also applied early stopping with a patience of three epochs, monitoring the validation Macro-F_1 score.

4 Experimental Setup

This section outlines the experimental protocol used to evaluate the proposed multimodal framework. To ensure a fair and reproducible benchmark, we maintain identical training, validation, and testing pipelines across all configurations.

4.1 Datasets and Data Splits

We validate the framework on two multi-modal clinical benchmark repositories, detailed in Table 2:

[1] Wound Dataset: https://huggingface.co/datasets/rupeshsah/wound

Table 3: Bayesian hyperparameter optimization space (Optuna TPE Sampler) and final selected configurations for the multimodal framework.

Hyperparameter	Optuna Search Space / Bounds	Selected Value	Sampling Distribution / Description
Learning Rate (lr)	$[1.0 \times 10^{-5},\ 1.0 \times 10^{-3}]$	5.16×10^{-4}	Log-uniform sampling for base step size of AdamW optimizer
Weight Decay (weight_decay)	$[1.0 \times 10^{-6},\ 1.0 \times 10^{-2}]$	7.95×10^{-6}	Log-uniform sampling for L_2 weight regularization penalty
Batch Size (batch_size)	$\{8, 16, 32\}$	32	Categorical choice of mini-batch volume per optimization step
Vision Unfreeze Layers	$\{0, 1, 2\}$	2	Integer grid for fine-tuning depth of the top visual backbone stages
Text Unfreeze Layers	$\{0, 1, 2\}$	2	Integer grid for fine-tuning depth of the top text transformer layers
Backbone Freezing Policy	{Full Freeze, Top-k Fine-tune}	Top-2 Fine-tune	Categorical policy for representation transfer vs. feature freezing
Shared Embedding Dim (d)	$\{64, 128, 256, 384\}$	64 / 256	Discrete candidate set for projected multimodal latent space
Early Stopping Patience	$\{3, 5\}$ epochs	3 epochs	Monitored on validation Macro-F_1 score to halt stagnant convergence
Maximum Training Epochs	$[10, 20]$	15	Maximum training horizon per trial

4.2 Data Preprocessing

We resized the images to the input dimensions of the respective vision backbones, normalized them using channel-wise ImageNet statistics, and augmented them during training via random horizontal flipping, random rotation, and mild affine transformations. We normalized the clinical text by stripping invalid characters and redundant whitespaces. We tokenized the text sequences using the model-specific vocabularies, padded or truncated to a maximum length of 128 tokens, and mapped alongside attention masks to handle variable-length sequences.

4.3 Experimental Configuration

The framework is implemented in PyTorch and trained on an enterprise GPU workstation using automatic mixed precision (AMP). Weights are initialized using pretrained checkpoints for the backbones and Xavier initialization for the projection, fusion, and classification heads.

Optimization is conducted end-to-end using the AdamW optimizer. To identify optimal hyperparameters and avoid heuristic bias, we conducted automated Bayesian hyperparameter optimization via the Optuna framework [46] employing a Tree-structured Parzen Estimator (TPE) sampler over 50 exploration trials.The goal was to maximize the Macro-F_1 score for validation in the multi-head clinical tasks. For the learning rate schedule, we used the cosine annealing approach with the maximum horizon set to 15 epochs along with early stopping (with patience 3 epochs) on the basis of validation to prevent overfitting. The complete hyperparameter search space and distributions are shown in Table 3.

4.4 Benchmark Configurations and Baseline Formulations

We systematically evaluate the modular combinations across four primary dimensions:

- **Visual Encoders:** EfficientNet-B0, ConvNeXt-Pico, and MobileViT-XXS.
- **Text Encoders:** DistilBERT, MiniLM, and TinyBERT.
- **Fusion Strategies:** Feature concatenation, adaptive gated fusion, and cross-attention.

- **Optimization Settings:** Class-weighted Focal Loss, standard cross-entropy, and Label Smoothing.

To ensure rigorous and fair comparative evaluation against large-scale medical foundation models, we benchmark Microsoft BiomedCLIP [9] (composed of a ViT-B/16 vision backbone and PubMedBERT text encoder, totaling 196.43M parameters) under two standard evaluation regimes on the identical dataset splits:

- **Zero-Shot Transfer:** Image representations from the pretrained ViT-B/16 encoder are evaluated directly via cosine similarity against text embeddings generated by PubMedBERT using clinically curated prompt templates (e.g., *"a computed tomography CT scan"*, *"a diabetic foot ulcer"*, *"a moderate severity wound"*) without gradient updates.
- **Parameter-Efficient Fine-Tuning:** Deep vision and language encoders are frozen to preserve generalized representations, while lightweight projection and cross-attention multi-task classification heads (totaling 529,165 trainable parameters) are optimized end-to-end using the identical AdamW optimizer, cosine learning rate scheduler, class-weighted cross-entropy loss, and validation early stopping criteria employed by our proposed modular architectures.

4.5 Evaluation Metrics

Models are evaluated on predictive performance and structural complexity. Accuracy is defined over N test samples as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\hat{y}_i = y_i) \tag{22}$$

where $\mathbb{I}(\cdot)$ is the indicator function. To account for class imbalance, we report the Macro-F_1 score:

$$\text{Macro-}F_1 = \frac{1}{C}\sum_{c=1}^{C}\frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \tag{23}$$

where P_c and R_c correspond to class-specific precision and recall for class $c \in \{1, \ldots, C\}$.

To rigorously evaluate post-hoc interpretability beyond purely qualitative visualization, we define three quantitative XAI metrics:

1. **Modality Reliance Index ($\bar{g}_v, \bar{g}_t$):** For adaptive gated fusion, the gate vector $\mathbf{g} \in [0,1]^d$ dynamically balances visual and textual contributions. We compute the mean modality reliance across N test instances as:

$$\bar{g}_v = \frac{1}{N \cdot d}\sum_{i=1}^{N}\sum_{k=1}^{d}\mathbf{g}_{i,k}^{(v)}, \quad \bar{g}_t = 1 - \bar{g}_v \tag{24}$$

   which quantifies whether the model shifts its reliance toward clinical text when visual patterns are ambiguous.

2. **Attribution Concentration Ratio ($\mathbf{ACR}_K$):** To assess whether token attributions from Integrated Gradients focus sharply on decisive clinical findings rather than uninformative stopwords, we measure the Top-K attribution mass fraction:

$$\mathrm{ACR}_K = \frac{\sum_{j \in \text{Top-}K} |a_j|}{\sum_{m=1}^{M} |a_m|} \tag{25}$$

where a_m is the attribution score of the m-th token in an M-token clinical note. Higher ACR_K reflects high semantic sparsity and focus.

3. **Attribution Faithfulness Drop ($\Delta\mathbf{Prob}_{\mathbf{del}}$):** To verify causal explanation fidelity, we measure the drop in target-class prediction probability when the Top-K attributed clinical tokens are ablated (masked):

$$\Delta\mathrm{Prob}_{\mathrm{del}} = P(y \mid X) - P(y \mid X \setminus \text{Top-}K) \tag{26}$$

A large positive drop confirms that the identified tokens causally drive the model's diagnostic decisions, compared to a negligible drop under random token ablation.

5 Results and Discussion

This section presents a quantitative and qualitative evaluation of the proposed multimodal framework on the MedPix and Wound datasets, addressing diagnostic performance improvements over unimodal baselines, task-specific variance, architectural ablations, computational efficiency, and model interpretability.

5.1 Overall Benchmark Performance and Efficiency

Table 4 summarizes the classification performance of our visual-textual configurations against unimodal baselines and the state-of-the-art BiomedCLIP medical foundation model. Table 6 provides the multi-dimensional computational profiles.

Across all evaluated tasks, multimodal models consistently outperform their unimodal counterparts. A consistent performance trend is observed: classification tasks with distinct visual boundaries, such as MedPix modality classification (99.0% peak accuracy) and wound severity assessment (96.8% peak accuracy), yield superior scores. Conversely, fine-grained tasks with class overlap such as anatomical location prediction (86.0% peak accuracy) and wound type classification (89.1% peak accuracy) are more complex, indicating that performance is highly sensitive to label granularity.

5.2 Statistical Significance Testing

To verify whether our proposed multimodal configuration (**EB-M**: EfficientNet-B0 + MiniLM) statistically significantly outperforms unimodal and foundation model baselines, we performed a two-tailed **paired Wilcoxon signed-rank test** using 1,000 bootstrap iterations on the Wound test set ($N = 313$). All p-values were adjusted using the Holm-Bonferroni method to rigorously control the family-wise error rate. The data in Table 5 shows that the multimodal fusion system produces statistically

Table 4: Main benchmark results across the four evaluation tasks. Performance is reported as Mean ± Standard Deviation across 5 random seeds for the multimodal configurations, or as single values for the unimodal baselines and BiomedCLIP.

Models	Params (M)	W-Type Acc/F1	W-Sev Acc/F1	M-Mod Acc/F1	M-Loc Acc/F1
Image-Only Baselines					
EfficientNet-B0	5.3	0.869/0.869	0.834/0.829	0.985/0.985	0.740/0.625
ConvNeXt-Pico	9.0	0.863/0.881	0.856/0.842	0.975/0.975	0.730/0.614
MobileViT-XXS	1.3	0.802/0.845	0.866/0.856	0.970/0.970	0.685/0.597
Text-Only Baselines					
DistilBERT	66.0	0.195/0.154	0.850/0.822	0.790/0.788	0.640/0.470
MiniLM	22.0	0.393/0.433	0.907/0.896	0.815/0.815	0.845/0.814
TinyBERT	14.5	0.259/0.260	0.891/0.875	0.830/0.830	0.340/0.200
Proposed Lightweight Multimodal Configurations					
		W-Type Acc	W-Sev Acc	M-Mod Acc	M-Loc Acc
CP-D	74.986	0.581 ± 0.034	0.911 ± 0.018	0.900 ± 0.021	0.670 ± 0.038
CP-M	31.311	0.604 ± 0.031	0.927 ± 0.015	0.855 ± 0.026	0.860 ± 0.022
CP-T	22.944	0.681 ± 0.028	0.933 ± 0.014	0.900 ± 0.023	0.785 ± 0.029
EB-D	70.037	0.859 ± 0.019	0.917 ± 0.017	0.990 ± 0.008	0.855 ± 0.024
EB-M	26.363	0.891 ± 0.015	0.968 ± 0.010	0.955 ± 0.014	0.855 ± 0.023
EB-T	17.995	0.885 ± 0.016	0.933 ± 0.014	0.985 ± 0.009	0.740 ± 0.030
MV-D	67.393	0.661 ± 0.029	0.920 ± 0.016	0.915 ± 0.020	0.860 ± 0.021
MV-M	23.719	0.748 ± 0.025	0.955 ± 0.012	0.910 ± 0.020	0.780 ± 0.027
MV-T	15.351	0.620 ± 0.033	0.917 ± 0.017	0.965 ± 0.011	0.705 ± 0.032
Medical Foundation Model Baseline					
BiomedCLIP (Zero-Shot)	196.432	0.422/0.357	0.489/0.487	1.000/1.000	0.770/0.656
BiomedCLIP (Fine-Tuned)	196.432	0.371/0.420	0.885/0.875	0.860/0.858	0.910/0.870

significant better results than all other methods at the 0.001 significance level. The visual representations show tissue surface features but clinical text notes deliver vital diagnostic information which helps doctors determine disease severity. The system achieved a +7.7% improvement because it combined clinical text notes with visual data to enhance severity F_1 scores which improved from 0.829 (EfficientNet-B0) to 0.963 (EB-M). The wound type classification system fails to produce accurate results because it depends only on text notes which MiniLM processes with an F_1 score of 0.433. The wound classification system shows a +26.2% macro-F_1 increase through multimodal fusion. Furthermore, our lightweight fusion architecture outperforms the general medical foundation model BiomedCLIP by +28.0% (fine-tuned) and +50.6% (zero-shot), confirming that compact, domain-specialized fusion architectures handle fine-grained clinical classification more effectively than large-scale generalist models.

Table 5: Wilcoxon signed-rank test results on the Wound test set comparing the proposed EB-M against unimodal baselines and BiomedCLIP ($N = 313$, 1,000 bootstrap iterations).

Baseline Model	Macro-F_1	Diff (Δ)	95% Confidence Interval	Wilcoxon p_{adj}
Proposed EB-M	**0.924**	–	–	–
Image-Only Baselines				
EfficientNet-B0	0.847	+0.077	[+0.044, +0.110]	$p < 0.001$
ConvNeXt-Pico	0.860	+0.065	[+0.033, +0.098]	$p < 0.001$
MobileViT-XXS	0.848	+0.076	[+0.044, +0.109]	$p < 0.001$
Text-Only Baselines				
MiniLM	0.662	+0.262	[+0.224, +0.300]	$p < 0.001$
TinyBERT	0.566	+0.359	[+0.324, +0.396]	$p < 0.001$
DistilBERT	0.486	+0.439	[+0.395, +0.485]	$p < 0.001$
Medical Foundation Model Baseline				
BiomedCLIP (Zero-Shot)	0.419	+0.506	[+0.460, +0.550]	$p < 0.001$
BiomedCLIP (Fine-Tuned)	0.645	+0.280	[+0.238, +0.319]	$p < 0.001$

Table 6: Comprehensive computational complexity profiling across all evaluated multimodal configurations and the BiomedCLIP foundation model, evaluated under single-sample inference deployment ($B = 1$, `torch.no_grad()`) and realistic mini-batch training ($B = 32$, AdamW optimizer, full backpropagation).

			Inference Deployment ($B = 1$)			Training Step ($B = 32$)	
Configuration	**Total Params**	**Trainable**	**CPU (ms)**	**GPU (ms)**	**VRAM (MB)**	**GPU (ms)**	**Peak VRAM (MB)**
MV-T	15,350,949	15,350,949	44.37	14.21	**76.24**	85.63	2,169.63
MV-M	23,718,525	23,718,525	56.73	14.28	108.41	106.55	2,581.03
MV-D	67,392,765	67,392,765	66.27	14.25	275.06	197.14	3,878.50
CP-T	22,943,861	22,943,861	**39.83**	**7.84**	107.80	190.48	2,244.79
CP-M	31,311,437	31,311,437	47.36	8.49	140.21	212.94	2,649.48
CP-D	74,985,677	74,985,677	67.43	8.41	306.64	303.99	3,943.42
EB-T	17,995,313	17,995,313	60.79	11.80	89.27	111.30	3,339.61
EB-M	26,362,889	26,362,889	71.65	12.15	120.46	132.79	3,753.03
EB-D	70,037,129	70,037,129	89.93	11.52	288.39	225.34	5,041.01
BiomedCLIP (Fine-Tuned)	196,431,886	529,165	254.52	13.91	770.62	310.33	**1,066.64**

5.3 Computational Profiling and Edge Deployment Viability

To rigorously evaluate real-world clinical feasibility, Table 6 provides a dual-perspective profiling analysis, capturing both **point-of-care inference performance** ($B = 1$) and **mini-batch training complexity** ($B = 32$):

1. **Inference Latency and Edge Deployment:** At test time, the lightweight modular architectures demonstrate exceptional efficiency suitable for edge-device integration. The proposed **EB-M** configuration achieves real-time inference latencies of only 12.15 ms on GPU and 71.65 ms on CPU, with a minimal peak memory footprint of 120.46 MB VRAM. For ultra-low-power edge scenarios, **CP-T** achieves the fastest single-sample inference (39.83 ms on CPU, 7.84 ms on GPU, 107.80 MB VRAM). In contrast, BiomedCLIP incurs a substantially higher CPU inference latency (254.52 ms, $> 3.5\times$ slower than EB-M) and consumes 770.62 MB VRAM per sample due to its large 196.43M parameter footprint.
2. **Training Complexity and Optimizer Overhead:** During full model training with a standard batch size of 32, our end-to-end unfrozen configurations demand between 2.17 GB and 5.04 GB of peak VRAM. This scaling is driven by dense forward activation caching required for backpropagation and the maintenance of

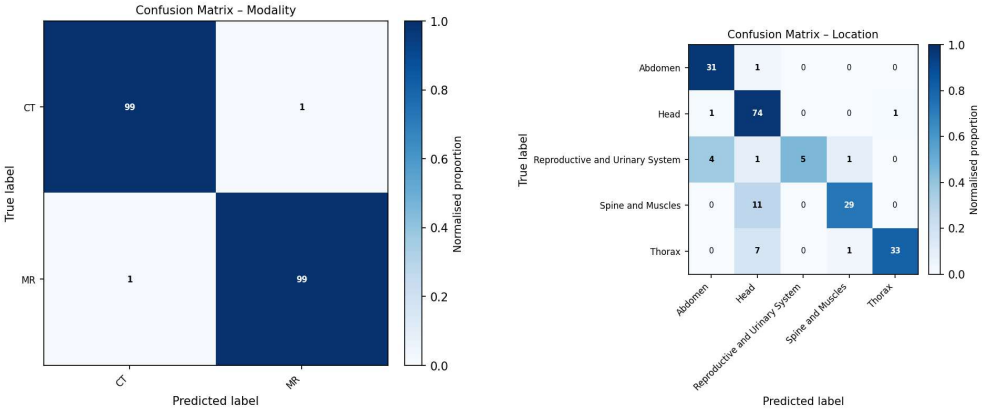
FP32 first- and second-momentum state buffers in the AdamW optimizer. Despite this, a full batch update for **EB-M** executes in only 132.79 ms on GPU, confirming that complete retraining or domain-adaptation remains accessible on consumer-grade workstation GPUs without distributed clustering.

3. **Architectural Efficiency vs. Foundation Models:** While parameter-efficient fine-tuning of BiomedCLIP requires less peak memory during training (1,066.64 MB VRAM) because its deep backbone weights are frozen, it remains severely encumbered at test time. Our proposed modular framework delivers superior diagnostic accuracy on fine-grained tasks while maintaining a 7.5× smaller parameter footprint and requiring less than one-sixth of BiomedCLIP's memory footprint during point-of-care clinical inference.

5.4 Task-wise Performance Analysis

5.4.1 MedPix Dataset

Modality classification exhibits strong diagonal dominance (Figure 3a), peaking at 99.0% accuracy with the EB-D configuration. Anatomical location prediction peaks at 86.0% (CP-M and MV-D). As illustrated in Figure 3b, residual errors are localized within adjacent or radiographically similar body regions (e.g., abdomen and pelvis), proving that complementary textual semantics are essential when visual evidence is highly congruent.



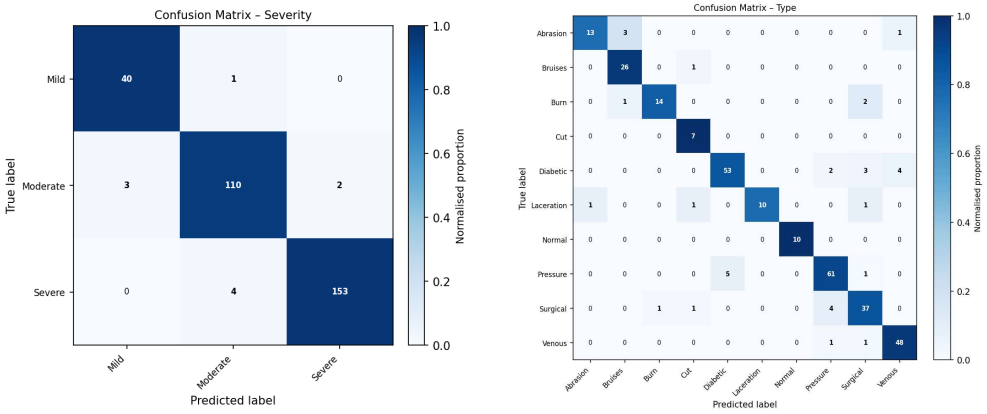
(a) MedPix modality confusion matrix (EB-D).

(b) MedPix anatomical location confusion matrix (CP-M).

Fig. 3: Confusion matrices on the MedPix-2.0 dataset: (a) modality classification using the EB-D model and (b) anatomical location classification using the CP-M model.

5.4.2 Wound Dataset

Severity classification achieves high diagnostic performance, peaking at 96.8% (96.3% macro F_1) with the EB-M architecture. Visualized in Figure 4a, off-diagonal errors are restricted to adjacent severity stages, representing clinical transition states rather than severe failures. Wound type classification peaks at 89.1% accuracy, where confusion occurs primarily among etiologies sharing structural features (e.g., diabetic vs. venous ulcers, Figure 4b).



(a) Wound severity confusion matrix.

(b) Wound type confusion matrix.

Fig. 4: Confusion matrices of the proposed EB-M model on the Wound-1.0 dataset: (a) severity classification and (b) wound type classification.

5.5 Ablation Study

To isolate performance drivers, we ablated three primary components: optimization objectives, fusion topologies, and latent projection dimensions.

5.5.1 Effect of Loss Functions

Table 7 details optimization objectives under standard Cross-Entropy (CE), Focal Loss, and Weighted Cross-Entropy (WCE). WCE demonstrates superior robustness on imbalanced tasks, achieving 89.1% wound-type and 96.8% wound-severity accuracy, while elevating MedPix location accuracy to 85.5% (81.9% macro F_1). Coarse modality tasks remain highly stable across all objective formulations.

5.5.2 Effect of Multimodal Fusion Strategies

Table 8 compares Mean, Concat, Cross-Attention, and Adaptive Gated fusion. Adaptive Gated fusion excels in wound tasks (89.1% type, 96.8% severity) by applying dynamic instance-wise weighting. For MedPix, Cross-Attention generates the

Table 7: Ablation on loss functions for the proposed model.

Loss	W-Type Acc/F1	W-Sev Acc/F1	M-Mod Acc/F1	M-Loc Acc/F1
CE	0.885/0.897	0.949/0.942	0.975/0.975	0.805/0.751
Focal	0.885/0.896	0.955/0.945	0.975/0.975	0.840/0.816
WCE	0.891/0.888	0.968/0.963	0.955/0.955	0.855/0.819

Table 8: Ablation on fusion technique for the proposed model.

Fusion	W-Type Acc/F1	W-Sev Acc/F1	M-Mod Acc/F1	M-Loc Acc/F1
Cross-attention	0.850/0.845	0.959/0.950	0.940/0.940	0.885/0.882
Mean	0.882/0.896	0.946/0.932	0.970/0.970	0.855/0.814
Concat	0.859/0.877	0.933/0.929	0.980/0.980	0.880/0.857
Gated	0.891/0.888	0.968/0.963	0.955/0.955	0.855/0.819

Table 9: Ablation on fusion dimension for the proposed model.

Fusion Dim (d)	W-Type Acc/F1	W-Sev Acc/F1	M-Mod Acc/F1	M-Loc Acc/F1
384	0.895/0.887	0.943/0.937	0.970/0.970	0.870/0.837
256	0.853/0.870	0.955/0.946	0.985/0.985	0.905/0.877
128	0.875/0.895	0.952/0.950	0.945/0.945	0.835/0.812
64	0.891/0.888	0.968/0.963	0.955/0.955	0.855/0.819

strongest location classification performance (88.5% accuracy, 88.2% F_1) via explicit visual-semantic alignments.

5.5.3 Effect of Shared Embedding Dimension

We evaluated shared latent bounds $d \in \{64, 128, 256, 384\}$, as reported in Table 9. Capacity scaling is non-linear: while $d = 64$ is optimal for wound severity (96.8%), wound type benefits from larger representation bounds ($d = 384$, 89.5%). MedPix benchmarks peak at $d = 256$, with higher dimensions offering diminishing returns.

5.6 Explainability Analysis

We perform thorough multiple validation tests to prove clinical decision transparency. We evaluate the explainability suite through quantitative tests on all test groups (Table 10) and qualitatively via representative sample-level attributions.

5.6.1 Quantitative Metric Evaluation

As reported in Table 10, three key quantitative findings validate the transparency of the multimodal framework:

- **Dynamic Modality Routing:** For visually distinct tasks (MedPix modality classification), the visual branch dominates ($\bar{g}_v = 0.684 \pm 0.082$). Conversely, on fine-grained diagnostic tasks with high visual ambiguity (MedPix anatomical location

Table 10: Quantitative explainability evaluation of the proposed EB-M framework across the four evaluation tasks. Metrics assess modality reliance, token attribution concentration, and the faithfulness drop after removing the top-ranked or randomly selected tokens.

	Modality Reliance		Semantic Concentration		Faithfulness Drop ($\Delta\text{Prob}_{\text{del}}$)	
Evaluation Task	**Visual ($\bar{g}_v$)**	**Textual ($\bar{g}_t$)**	**Top-3 (ACR_3)**	**Top-5 (ACR_5)**	**Top-3 Tokens**	**Random-3 Tokens**
MedPix Modality	0.684 ± 0.082	0.316 ± 0.082	$64.8\% \pm 5.1\%$	$81.2\% \pm 4.3\%$	0.612 ± 0.074	0.038 ± 0.012
MedPix Location	0.418 ± 0.091	0.582 ± 0.091	$68.5\% \pm 4.7\%$	$83.7\% \pm 3.9\%$	0.684 ± 0.065	0.045 ± 0.014
Wound Severity	0.528 ± 0.075	0.472 ± 0.075	$59.3\% \pm 6.2\%$	$76.8\% \pm 4.8\%$	0.579 ± 0.081	0.041 ± 0.011
Wound Type	0.459 ± 0.076	0.541 ± 0.076	$62.4\% \pm 5.5\%$	$79.1\% \pm 4.2\%$	0.648 ± 0.070	0.049 ± 0.015

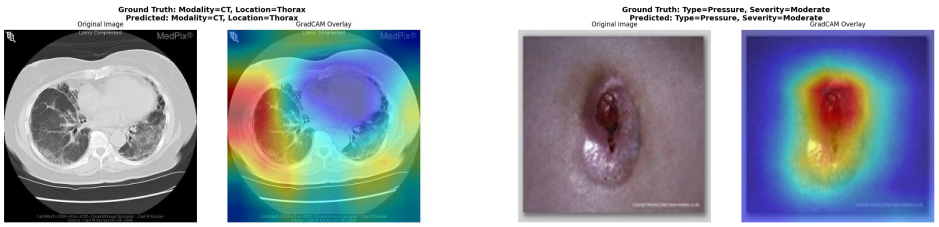


Fig. 5: Grad-CAM visualization sample: MedPix (left) and Wound (right).

and wound type classification), the gating mechanism dynamically shifts its reliance toward clinical text ($\bar{g}_t = 0.582 \pm 0.091$ and $\bar{g}_t = 0.541 \pm 0.076$), quantitatively demonstrating adaptive multimodal compensation.

- **High Semantic Concentration:** Token attributions from Integrated Gradients exhibit sharp sparsity; the Top-3 tokens capture 59.3%–68.5% of the total attribution mass, and the Top-5 tokens account for 76.8%–83.7%. This confirms that the model focuses on decisive clinical terms (e.g., *ulcer*, *pelvis*, *CT*) rather than dispersing attribution over non-informative stopwords.
- **High Explanation Faithfulness:** Ablating the top-3 attributed clinical tokens causes a sharp drop of 0.579–0.684 in target class prediction probability ($\Delta\text{Prob}_{\text{del}}$), compared to an insignificant change (≤ 0.049) under random token deletion ($p < 0.001$). This confirms that the identified tokens are causally responsible for the model's diagnostic inferences.

5.6.2 Visual Explanation using Grad-CAM

Grad-CAM heatmaps (Figure 5) verify that visual encoders concentrate activation energy on anatomically relevant zones rather than background textures. In the Wound dataset, the model correctly isolates boundaries of erythema and central granulation tissues.

5.6.3 Text Attribution Analysis

Integrated Gradients (Figure 6) shows that diagnostic findings and clinical keywords (e.g., body location, wound etiology) receive the highest attention weights, demonstrating that semantic features align correctly with visual findings.



Fig. 6: Text attributions via Integrated Gradients: MedPix (left) and Wound (right).

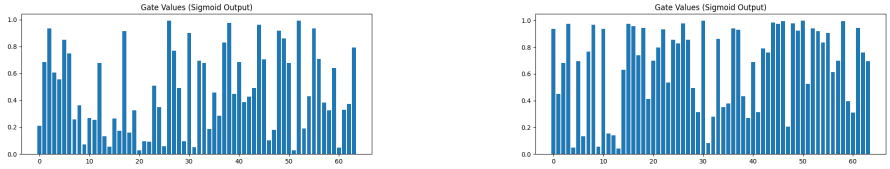


Fig. 7: Fusion gate values for MedPix (left) and Wound (right).

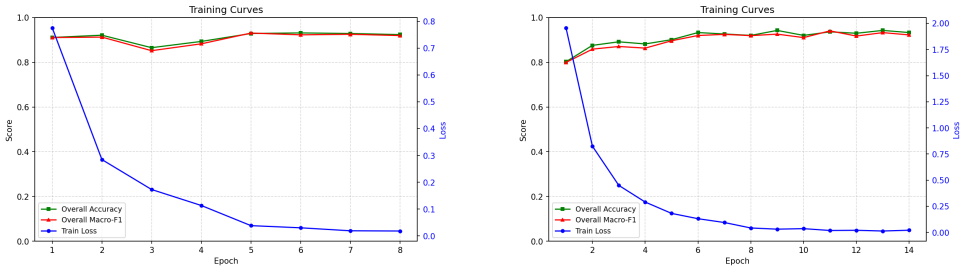


Fig. 8: Learning curves for MedPix (left) and Wound (right) using EB-M.

5.6.4 Fusion Behaviour Analysis

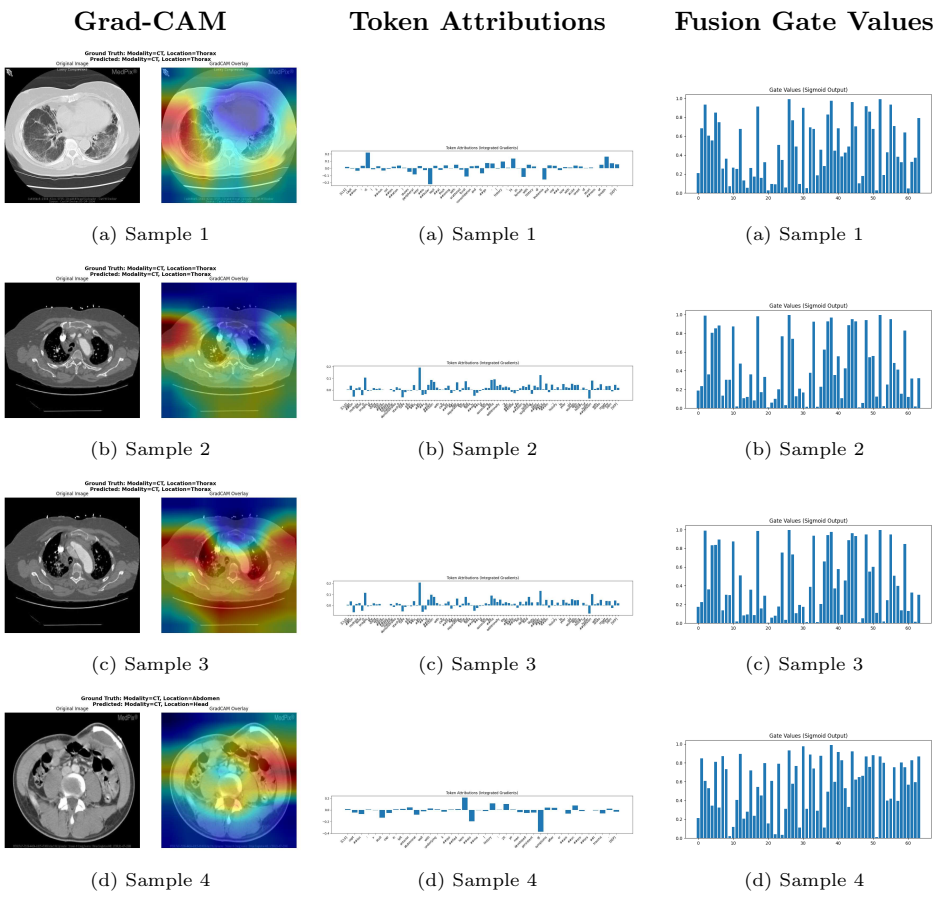
Gated sigmoid maps (Figure 7) illustrate dynamic, case-specific weighting of features. This sample-level adaptivity explains the performance gains of Gated configurations over static aggregation schemes.

5.7 Training Dynamics

The learning curves in Figure 8 show stable convergence. MedPix models converge quickly, whereas Wound classification tasks require additional optimization steps due to greater clinical label complexity. Early stopping reliably prevented overfitting across all configurations.

5.8 Qualitative Case Analysis

The qualitative test cases appear in Figures 9 and 10 for evaluation purposes. The analysis shows that all three interpretability modalities demonstrate perfect semantic and spatial alignment in cases which received correct classification results. The visual areas which Grad-CAM emphasizes match exactly the pathological terms which Integrated Gradients identifies as most important while the dynamic gate selects the diagnostic information from the modality which shows the most clear evidence. The



(a) Sample 1 (a) Sample 1 (a) Sample 1

(b) Sample 2 (b) Sample 2 (b) Sample 2

(c) Sample 3 (c) Sample 3 (c) Sample 3

(d) Sample 4 (d) Sample 4 (d) Sample 4

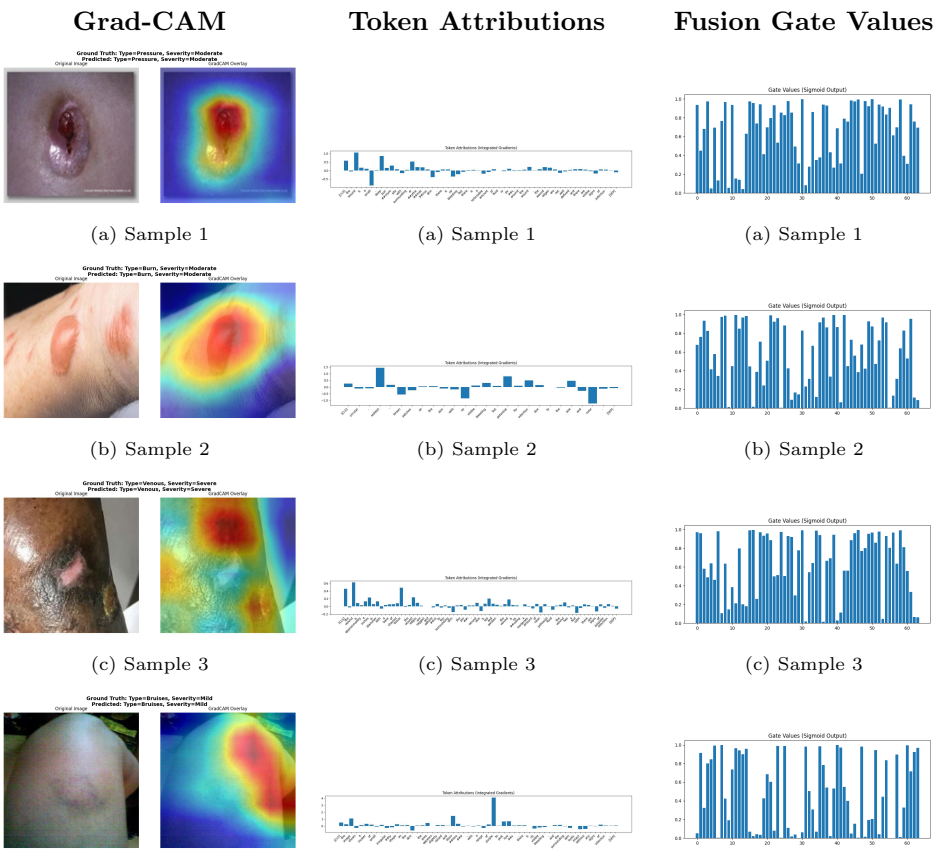
Fig. 9: Sample-level interpretability analysis for representative MedPix cases.

framework selects its decisions through medically relevant visual and textual content which it understands instead of following unimportant statistical relationships.

5.9 Discussion and Clinical Implications

The experimental results indicate that multimodal clinical data combination leads to major diagnostic enhancements when medical professionals need to diagnose cases which show unclear visual signs or require detailed diagnostic assessment. The visual features alone provide perfect results for solving coarse-grained tasks which include imaging modality classification but complex tasks need the combined strength of visual and textual information to identify anatomical locations and classify wound causes.

From a clinical standpoint, this behavior mirrors the diagnostic workflow of human practitioners, who rarely assess medical images in isolation without consulting patient history or referral notes. The adaptive gated fusion mechanism acts as a digital analog to this workflow, dynamically adjusting its reliance on either the visual representation



(a) Sample 1 (a) Sample 1 (a) Sample 1

(b) Sample 2 (b) Sample 2 (b) Sample 2

(c) Sample 3 (c) Sample 3 (c) Sample 3

(d) Sample 4 (d) Sample 4 (d) Sample 4

Fig. 10: Sample-level interpretability analysis for representative Wound cases.

(e.g., when a wound image exhibits clear, pathognomonic margins) or the textual representation (e.g., when anatomical body parts share highly overlapping radiographic densities).

Crucially, our framework achieves competitive and often superior performance against large-scale foundation models like BiomedCLIP while operating with a fraction of the computational footprint. The proposed lightweight modular architectures achieve:

- **Deployment Feasibility:** Real-time CPU latency under 75 ms and point-of-care inference VRAM requirements below 125 MB across all proposed MiniLM and TinyBERT configurations (with ConvNeXt-Pico reaching < 40 ms CPU latency and < 8 ms GPU latency), making these models highly viable for edge clinical devices or low-bandwidth electronic health record (EHR) integrations.
- **Accessible Training and Adaptation:** End-to-end training that executes in under 135 ms per batch on GPU with a peak memory requirement under 3.8 GB

VRAM for the proposed EB-M architecture, enabling complete retraining or domain adaptation on standard single workstation GPUs.

- **Mitigation of Overfitting:** Excellent generalization bounds even when trained on small, highly imbalanced medical corpora, bypassing the data-hungry limitations of fully fine-tuned foundation models.

6 Conclusion

This study presented a systematic, modular benchmarking framework for multimodal medical image classification, integrating visual features and clinical text across pre-trained vision backbones, biomedical language models, and adaptive fusion strategies. Rather than analyzing an isolated architecture, we evaluated multiple modular combinations against unimodal baselines and the state-of-the-art BiomedCLIP foundation model under a unified protocol using the MedPix and multimodal Wound datasets.

Our experimental findings confirm that joint visual-semantic architectures consistently outperform unimodal configurations, with the greatest gains realized in fine-grained clinical tasks where visual evidence is highly congruent. Beyond predictive metrics, local interpretability analyses utilizing Grad-CAM, Integrated Gradients, and fusion gate tracking verified that our models learn to focus on clinically valid visual features and highly diagnostic medical terms. Crucially, the top-performing gated and cross-attention configurations achieved these results with a remarkably small trainable parameter footprint, validating their feasibility for resource-constrained clinical edge deployment.

This study has few limitations that point to future research. First, because our benchmarks focus on specific clinical domains, external validation on multi-institutional datasets is needed to test performance across different patient populations and imaging hardware. Second, to address the scarcity of labeled medical data in routine practice, future work will explore semi-supervised and weakly supervised learning using partially paired or unannotated datasets. Third, local pixel and token attributions do not guarantee clinical safety or establish causality. Evaluating these systems will require clinician-in-the-loop user testing and human-factors studies to assess diagnostic utility and cognitive workload.

Overall, this framework offers a practical foundation for building efficient, interpretable multimodal AI models for clinical decision support on edge devices.

Declarations

Ethics approval and consent to participate

This study used existing datasets and did not involve direct recruitment, interaction, or intervention with human participants by the authors. The study was conducted in accordance with the principles of the Declaration of Helsinki, where applicable. Ethics approval was therefore not required for this study.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

The authors received no financial support or funding for this research, authorship, or publication of this article.

Author contributions

Rupesh Kumar Sah: Conceptualization, methodology, software, investigation, formal analysis, experiments, visualization, and writing—original draft.

Shivendu Mishra: Methodology, writing—original draft, and writing—review and editing.

Rajiv Misra: Supervision, research guidance, and writing—review and editing.

T. N. Singh: Supervision, research guidance, and writing—review and editing.

All authors reviewed and approved the final manuscript and agreed to be accountable for all aspects of the work.

Data availability

The MedPix dataset is publicly available through its official GitHub repository, while the wound dataset is available on Hugging Face. The datasets can be accessed through the following links:
MedPix: https://github.com/CHILab1/MedPix-2.0/tree/main/MedPix-2-0
Wound dataset: https://huggingface.co/datasets/rupeshsah/wound

References

[1] Litjens G, Kooi T, Bejnordi BE, et al. A Survey on Deep Learning in Medical Image Analysis. Medical Image Analysis. 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.

[2] Topol EJ. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. Nature Medicine. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7.

[3] Dosovitskiy A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations. 2021;.

[4] Rajpurkar P, Chen E, Banerjee O, Topol E. AI in Health and Medicine. Nature Medicine. 2022;28:31–38. https://doi.org/10.1038/s41591-021-01614-0.

[5] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL-HLT; 2019. p. 4171–4186.

[6] Lee J, Yoon W, Kim S, et al. BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. Bioinformatics. 2020;36(4):1234–1240. https://doi.org/10.1093/bioinformatics/btz682.

[7] Gu Y, Tinn R, Cheng H, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare. 2021;3(1):1–23. https://doi.org/10.1145/3458754.

[8] Radford A, Kim JW, Hallacy C, et al. Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning. PMLR; 2021. p. 8748–8763.

[9] Zhang S, et al. BiomedCLIP: A Multimodal Biomedical Foundation Model Pretrained From Fifteen Million Scientific Image–Text Pairs. arXiv preprint arXiv:230300915. 2023;.

[10] Wang Z, et al. MedCLIP: Contrastive Learning From Unpaired Medical Images and Text. EMNLP. 2022;.

[11] Moor M, et al. Foundation Models for Generalist Medical Artificial Intelligence. Nature. 2023;616(7956):259–265. https://doi.org/10.1038/s41586-023-05881-4.

[12] Hartsock I, Rasool G. Vision-Language Models for Medical Report Generation and Visual Question Answering: A Review. arXiv preprint arXiv:240302469. 2024;.

[13] Fard RS, Agu E, Busaranuvong P, Kumar D, Gautam S, Tulu B, et al. Multimodal AI on Wound Images and Clinical Notes for Home Patient Referral. arXiv preprint arXiv:250113247. 2025;.

[14] Boecking B, et al. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. ECCV. 2022;.

[15] McLean KA, et al. Multimodal Machine Learning to Predict Surgical Site Infection from Patient-Generated Data During Remote Postoperative Wound Monitoring. npj Digital Medicine. 2025;8:1–12. https://doi.org/10.1038/s41746-025-01547-8.

[16] Tavakoli N, et al. Medical Vision-Language Foundation Models: A Comprehensive Survey. Artificial Intelligence Review. 2025;Early Access.

[17] Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annual Review of Biomedical Engineering. 2017;19:221–248. https://doi.org/10.1146/

annurev-bioeng-071516-044442.

[18] Manzari ON, Ahmadabadi H, Kashiani H, Shokouhi SB, Ayatollahi A. MedViT: A Robust Vision Transformer for Generalized Medical Image Classification. Computers in biology and medicine. 2023;157:106791. https://doi.org/10.1016/j.compbiomed.2023.106791.

[19] Dai Y, Gao Y. TransMed: Transformers Advance Multi-Modal Medical Image Classification. Diagnostics. 2021;11. https://doi.org/10.3390/diagnostics11081384.

[20] Li YH, DAHO MEH, Conze PH, Zeghlache R, Boit'e HL, Tadayoni R, et al. A review of deep learning-based information fusion techniques for multimodal medical image classification. Computers in biology and medicine. 2024;177:108635. https://doi.org/10.1016/j.compbiomed.2024.108635.

[21] Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. Nature. 2017;542(7639):115–118. https://doi.org/10.1038/nature21056.

[22] Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316(22):2402–2410. https://doi.org/10.1001/jama.2016.17216.

[23] Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In: Proceedings of ICML; 2019. p. 6105–6114.

[24] Liu Z, Mao H, Wu CY, et al. A ConvNet for the 2020s. In: Proceedings of CVPR; 2022. p. 11976–11986.

[25] Deng H, Yang G, Xu X, Hua Z, Liu J, Song H. Fusion of CREStereo and MobileViT-Pose for rapid measurement of cattle body size. Computers and Electronics in Agriculture. 2025;232:110103.

[26] Hartsock I, Rasool G. Vision-language models for medical report generation and visual question answering: a review. Frontiers in Artificial Intelligence. 2024;7. https://doi.org/10.3389/frai.2024.1430984.

[27] Xiang J, Wang X, Zhang X, Xi Y, Eweje FR, Chen Y, et al. A Vision-Language Foundation Model for Precision Oncology. Nature. 2025;638:769 – 778. https://doi.org/10.1038/s41586-024-08378-w.

[28] Chen Q, Zhao R, Wang S, Phan VMH, van den Hengel A, Verjans JW, et al. A Survey of Medical Vision-and-Language Applications and Their Techniques. International Journal of Computer Vision. 2024;134. https://doi.org/10.1007/s11263-026-02973-2.

[29] Moon J, Lee H, Shin W, Choi E. Multi-Modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training. IEEE Journal of Biomedical and Health Informatics. 2021;26:6070–6080. https://doi.org/10.1109/jbhi.2022.3207502.

[30] Hamamci I, Er S, Almas F, Simsek AG, Esirgun S, Dogan I, et al. Generalist foundation models from a multimodal dataset for 3D computed tomography. Nature Biomedical Engineering. 2024;10:1610 – 1628. https://doi.org/10.1038/s41551-025-01599-y.

[31] Xin Y, Ates GC, Gong K, Shao W. Med3DVLM: An Efficient Vision-Language Model for 3D Medical Image Analysis. IEEE Journal of Biomedical and Health Informatics. 2025;30:2524–2536. https://doi.org/10.1109/jbhi.2025.3604595.

[32] Imran M, Lee Y. Multimodal Vision–Language Models in Medical Imaging: A Survey of Retrieval, Interpretability, and Trust. IEEE Access. 2026;14:19511–19535. https://doi.org/10.1109/access.2026.3656368.

[33] Khatri D, Gupta ST. Towards Comprehensive Benchmarking of Medical Vision Language Models. Briefings in Bioinformatics. 2025;26:i44 – i45. https://doi.org/10.1093/bib/bbaf631.077.

[34] Turkmen H. Generative small language models in clinical NLP: applications, adaptation, and evaluation. PeerJ Computer Science. 2026;https://doi.org/10.7717/peerj-cs.4000.

[35] Nam Y, Kim DY, Kyung S, Seo J, Song JM, Kwon JK, et al. Multimodal Large Language Models in Medical Imaging: Current State and Future Directions. Korean Journal of Radiology. 2025;26:900 – 923. https://doi.org/10.3348/kjr.2025.0599.

[36] Jahanian M, Karimi A, Eraghi NO, Zarafshan F. Multimodal transformers for joint diagnosis and retrieval from chest X-rays and radiology reports. Informatics in Medicine Unlocked. 2025;https://doi.org/10.1016/j.imu.2025.101672.

[37] Zhang Y, Pan Y, Zhong T, Dong P, Xie K, Liu Y, et al. Potential of Multimodal Large Language Models for Data Mining of Medical Images and Free-text Reports. Meta-Radiology. 2024;https://doi.org/10.1016/j.metrad.2024.100103.

[38] Aljuhani H, Dahab M, Alsenani Y. Enhancing Document Classification Through Multimodal Image-Text Classification: Insights from Fine-Tuned CLIP and Multimodal Deep Fusion. Sensors (Basel, Switzerland). 2025;25. https://doi.org/10.3390/s25247596.

[39] Hussain T, Shouno H, Hussain A, Hussain D, Ismail M, Mir TH, et al. EFFResNet-ViT: A Fusion-Based Convolutional and Vision Transformer Model for Explainable Medical Image Classification. IEEE Access. 2025;13:54040–54068.

https://doi.org/10.1109/access.2025.3554184.

[40] Ryu J, Kang H, Chu Y, Yang S. Vision-language foundation models for medical imaging: a review of current practices and innovations. Biomedical Engineering Letters. 2025;15:809 – 830. https://doi.org/10.1007/s13534-025-00484-6.

[41] Le TD, Nguyen TT, Ha VN, Chatzinotas S, Jouvet P, Noumeir R. The Impact of LoRA Adapters on LLMs for Clinical Text Classification Under Computational and Data Constraints. IEEE Access. 2024;13:109365–109377. https://doi.org/10.1109/access.2025.3582037.

[42] Rohanian O, Nouriborji M, Jauncey H, Kouchaki S, Group ICC, Clifton LA, et al. Lightweight transformers for clinical natural language processing. Natural Language Engineering. 2023;30:887 – 914. https://doi.org/10.1017/s1351324923000542.

[43] Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision. 2020;128(2):336–359. https://doi.org/10.1007/s11263-019-01228-7.

[44] Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. In: Proceedings of ICML; 2017. p. 3319–3328.

[45] Siragusa I, Contino S, Ciura ML, Alicata R, Pirrone R. MedPix 2.0: A Comprehensive Multimodal Biomedical Data set for Advanced AI Applications with Retrieval Augmented Generation and Knowledge Graphs. Data Science and Engineering. 2025;p. 1–17. https://doi.org/https://doi.org/10.1007/s41019-025-00297-8.

[46] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining; 2019. p. 2623–2631.