RESEARCH PAPER

A Modality-Dropout Cross-Attention Transformer for Robust Pneumonia Detection Using Chest X-Ray Images and Clinical Notes

Sowjanya Jindam[1], Aniketh Chittiprolu[2], Sathwika Reddy Gundra[3], Hari Shankar Punna[4] and Rachamalla Surya Prakash[5]

[1]Sr. Asst. Professor, M.V.S.R Engineering College, Hyderabad, India
[2]Student, M.V.S.R Engineering College, Hyderabad, India
[3]Student, M.V.S.R Engineering College, Hyderabad, India
[4]Department of CSE (DS), CVR College of Engineering, Hyderabad, India
[5]Student, M.V.S.R Engineering College, Hyderabad, India
Email: [1]sowjanya_it@mvsrec.edu.in, [2]aniketh.chittiprolu@gmail.com, [3]sathwikareddygundra@gmail.com, [4]harishankar@cvr.ac.in and [5]suryarachamalla25@gmail.com
ORCID: [1]0000-0001-6959-2110, [2]0009-0005-3328-7364, [3]0009-0009-8338-6394, [4]0009-0009-1142-6254 and [5]0009-0007-3274-8098

Received: 28th May, 2026; Revised: 6th June 2026; Accepted: 7th June, 2026; Available Online: 20th July, 2026

ABSTRACT
Automated pneumonia detection from chest radiographs is clinically important, but image-only systems can be brittle when radiographic findings are subtle, labels are report-derived, or contextual clinical evidence is unavailable at deployment. We propose a Modality-Dropout Cross-Attention Transformer (MDCAT), a multimodal architecture that combines chest X-ray image tokens with temporally eligible clinical-note or ED-text tokens using bidirectional cross-attention and trains under explicit missing-modality simulation. The model includes modality-availability embeddings, learned absent-modality tokens, auxiliary unimodal heads, and a consistency term that aligns masked-modality predictions with full-input predictions when both modalities are observed. Because pneumonia labels in common chest X-ray datasets are often derived from radiology reports, we specify a leakage-controlled data protocol that separates pre-imaging clinical text from same-study radiology reports and treats report text only as a secondary ablation. The evaluation design covers full multimodal inference, image-only inference, text-only inference, stochastic modality dropout, real missingness, image corruption, note truncation, calibration, and external image-only validation

Keywords: Pneumonia detection, chest X-ray, clinical notes, multimodal learning, cross-attention transformer, modality dropout, missing-modality robustness, medical AI

How to cite this article:
Jindam J, Chittiprolu A, Gundra SR, Punna HS, Prakash RS, A Modality-Dropout Cross-Attention Transformer for Robust Pneumonia Detection Using Chest X-Ray Images and Clinical Notes. Int J Drug Deliv Technol. 2026;16(6): 1154-1162. DOI: 10.25258/ijddt.16.6.155
Source of support: Nil.
Conflict of interest: None

1. INTRODUCTION

Pneumonia remains a major cause of morbidity and mortality, and chest radiography is one of the most common first-line imaging studies used in suspected lower respiratory infection. The World Health Organization reports that pneumonia is the single largest infectious cause of death in children worldwide, accounting for 740,180 deaths among children under five in 2019 [1]. In hospitals and emergency departments, clinicians rarely interpret radiographs in isolation: symptoms, acuity, vital signs, clinical history, laboratory data, prior diagnoses, and radiographic findings are combined to decide whether an opacity is likely infectious, atelectatic, edematous, chronic, or incidental.

Deep learning systems for chest X-ray (CXR) interpretation have advanced substantially through large public datasets such as ChestX-ray8 [2], CheXpert [3], MIMIC-CXR [4], MIMIC-CXR-JPG [5], and expert-annotated pneumonia resources from the Radiological Society of North America (RSNA) challenge [6]. Image-only pneumonia classifiers, including DenseNet-style approaches such as CheXNet on ChestX-ray14 [7], can achieve strong benchmark results, but benchmark performance does not automatically imply clinical robustness. Dataset labels differ, pneumonia prevalence varies, radiology reports may be used to create image labels, and many systems are not tested under missing contextual information or deployment-style data incompleteness.

**Author for Correspondence: sowjanya_it@mvsrec.edu.in*

Multimodal learning offers a natural path forward. Clinical notes can encode symptoms, triage context, prior medical his-tory, and temporal cues that complement radiographic patterns. Prior medical image-text work has shown that paired reports can improve representation learning [8]–[10]; cross-attention transformers can model interactions between heterogeneous token streams [11], [12]; and missing-modality studies show that multimodal transformers are often sensitive to absent modalities [13]. However, a pneumonia detector that uses clinical notes must avoid an important leakage trap: if the text input is the same radiology report used to define the pneumonia label, the model may learn report label extraction rather than diagnostic reasoning from independently available clinical evidence.

This paper proposes MDCAT, a Modality-Dropout Cross-Attention Transformer for robust pneumonia detection using chest X-ray images and temporally eligible clinical notes. The contribution is fourfold:

1) We define a leakage-controlled multimodal task in which text inputs are restricted to notes or emergency department (ED) data available at or before the CXR acquisition time, while same-study radiology reports are reserved for label generation or explicitly labelled ablations.

2) We introduce a bidirectional cross-attention fusion architecture with modality-availability embeddings and learned absent-modality tokens.

3) We train with modality dropout, auxiliary unimodal supervision, and full-to-masked prediction consistency so that the same model can operate with image plus text, image only, or text only.

4) We provide a reproducible evaluation protocol covering missing-modality robustness, external image-only validation, calibration, subgroup analysis, and statistically principled comparisons.

The manuscript is written as a methods and protocol paper. No numerical superiority, clinical utility, or radiologist-level claim is made without running the experiments.

2. RELATED WORK

A. Chest X-Ray Pneumonia Detection

Large chest radiograph datasets enabled modern deep learning benchmarks. ChestX-ray8 introduced a hospital-scale weakly labeled database with text-mined thoracic disease labels [2]. CheXNet popularized DenseNet-121 for pneumonia detection on ChestX-ray14, but its radiologist-level framing depends on a narrow annotation and comparison setup [7]. Kermany et al. demonstrated image-based diagnosis for pedi-atric pneumonia and retinal disease, providing a widely used pediatric pneumonia dataset [14]. CheXpert expanded large-scale CXR classification with uncertainty labels and expert comparison [3]. MIMIC-CXR provides Digital Imaging and Communications in Medicine (DICOM) radiographs paired with deidentified reports [4]; MIMIC-CXR-JPG provides processed JPG images and structured labels derived from reports [5]. RSNA pneumonia challenge annotations added expert labels for possible pneumonia on National Institutes of Health (NIH)-derived images [6]. VinDr-CXR further supports external validation with radiologist annotations [15].

These resources are valuable but not interchangeable. Report-derived labels encode radiological language, while clinical pneumonia may require symptoms, microbiology, treatment, discharge diagnoses, or adjudication. A robust multimodal study should therefore report label provenance and avoid treating all pneumonia labels as definitive clinical pneumonia.

B. Medical Image-Text Learning

Medical image-text learning has moved from task-specific report generation and fusion toward scalable representation learning. TieNet jointly modeled chest X-rays and text for classification and reporting [16]. ConVIRT used paired medical images and reports for bidirectional contrastive representation learning [8]. GLoRIA added global-local alignment between image regions and report words for label-efficient recogni-tion [9]. MedCLIP extended contrastive learning to unpaired medical images and text using semantic matching [10]. These methods motivate multimodal representations but are often built on radiology reports rather than independent clinical notes. In diagnostic prediction, this distinction matters because report text may contain the target finding.

C. Cross-Attention and Missing Modalities

Transformer architectures [17]–[19] support flexible token interactions across image patches and text tokens. Multimodal Transformer (MulT) showed that directional pairwise cross-modal attention can model interactions across unaligned sequences [11], while ViLBERT used co-attentional streams for vision-language representation learning [12]. Missing modalities remain a central deployment problem. ModDrop introduced modality-level dropout to encourage robustness to absent channels [20], and Ma et al. showed that multimodal transformers can be sensitive to missing inputs and that fusion strategy affects robustness [13]. MDCAT adapts these ideas to chest radiography and clinical notes with explicit availability conditioning and a leakage-aware evaluation design.

3. PROBLEM FORMULATION

Let each patient-level study be indexed by i. The observed image modality is x_i^I , a chest radiograph or a set of radiographs from one study. The observed text modality is x_i^T , , defined in this manuscript as pre-imaging ED or non-radiology clinical text available no later than the primary prediction timestamp t_i^*.

This timestamp is prespecified as the CXR study/acquisition time. Discharge summaries, same-study radiology reports, and untimestamped post-evaluation diagnosis fields are excluded from the primary text

modality. The primary prediction target

is $y_i \in \{0, 1\}^K$, where $K = 1$ for binary pneumonia detection or $K > 1$ for multilabel thoracic findings including pneumonia; uncertainty sensitivity analyses may use $y_i \in [0, 1]^K$ soft labels. The model outputs $\hat{y}_i \in [0, 1]^K$.

The modality availability vector is $m_i = (m_i^I, m_i^T)$ with $m_i^I, m_i^T \in \{0, 1\}$. Valid inference settings are (1, 1) for full multimodal input, (1, 0) for image-only input, and (0, 1) for text-only input. The all-missing setting (0, 0) is rejected and should trigger abstention.

Leakage constraint. If labels are derived from radiology reports, the same report cannot be used as x_i^T in the primary experiment. Same-study report text is permitted only in a clearly labeled report-text ablation that tests image-report consistency, not independent clinical-note prediction.

4. PROPOSED MODEL

MDCAT has four components: an image encoder, a text encoder, bidirectional cross-attention fusion, and modality-aware prediction heads. The architecture is summarized in Fig. 1.

A. Image and Text Encoders

The image encoder maps a radiograph into patch-level embeddings:

$$H_I^0 = [h_{I,\text{cls}}; h_{I,1}, \ldots, h_{I,N_I}] = \text{LN}(W_I f_I(x^I)) + P_I + E_I, \quad (1)$$

where f_I is a CXR-pretrained or ImageNet-pretrained vision encoder, W_I projects features to shared dimension d, P_I is a positional embedding, and E_I is an image-modality embedding.

The text encoder maps tokenized clinical text into contextual embeddings:

$$H_T^0 = [h_{T,\text{cls}}; h_{T,1}, \ldots, h_{T,N_T}] = \text{LN}(W_T f_T(x^T)) + P_T + E_T, \quad (2)$$

where f_T is a clinical-domain language encoder such as ClinicalBERT or BioClinicalBERT [21], [22], W_T maps hidden states to dimension d, and E_T is a text-modality embedding.

B. Bidirectional Cross-Attention Fusion

At fusion layer ℓ, image tokens attend to text tokens and text tokens attend to image tokens. Let

$$\mathcal{A}(Q, K, V) = \text{MHA}(\text{LN}(Q), \text{LN}(K), \text{LN}(V)). \quad (3)$$

Here, MHA denotes multi-head attention and LN denotes layer normalization. The gated cross-attention updates are

$$\tilde{H}_I^\ell = H_I^{\ell-1} + m_I m_T \mathcal{A}(H_I^{\ell-1}, H_T^{\ell-1}, H_T^{\ell-1}), \quad (4)$$

$$\tilde{H}_T^\ell = H_T^{\ell-1} + m_T m_I \mathcal{A}(H_T^{\ell-1}, H_I^{\ell-1}, H_I^{\ell-1}). \quad (5)$$

The availability gates prevent cross-attention updates whenever the source or target modality is absent. Pooling also receives e_m, masks absent real tokens, and includes a single learned missing indicator vector for each absent modality. Thus learned missing tokens act as missingness indicators rather than observed clinical or imaging evidence. Each stream then passes through feed-forward residual blocks:

$$H_I^\ell = \tilde{H}_I^\ell + \text{FFN}(\text{LN}(\tilde{H}_I^\ell)), \quad H_T^\ell = \tilde{H}_T^\ell + \text{FFN}(\text{LN}(\tilde{H}_T^\ell)). \quad (6)$$

An attention-pooling token aggregates the final representation:

$$g = \text{AttnPool}\left(z_F, [H_I^L; H_T^L; e_m]\right), \quad (7)$$

where e_m is an embedding of the availability pattern. The primary prediction head is

$$\hat{y} = \sigma(W_o g + b_o). \quad (8)$$

Auxiliary pre-fusion unimodal heads are

$$\hat{y}_I = \sigma(W_{oI} h_{I,\text{cls}}^0 + b_{oI}), \quad \hat{y}_T = \sigma(W_{oT} h_{T,\text{cls}}^0 + b_{oT}). \quad (9)$$

(1)

C. Modality Dropout and Missing Inputs

Let $a = (a_I, a_T)$ denote true modality availability before simulated dropout, and let $s = (s_I, s_T)$ denote the sampled training keep pattern. When both modalities are truly observed,

(2 $$s \sim \begin{cases} (1,1) & \text{with probability } 1 - p_I - p_T, \\ (1,0) & \text{with probability } p_T, \\ (0,1) & \text{with probability } p_I. \end{cases} \quad (10)$$

If only one modality is truly available, s keeps that modality and does not synthesize the missing one. The effective model availability is $m = a \odot s$. This distribution requires $p_I + p_T \leq 1$ and excludes all-missing training views. A conservative default is $p_I = p_T = 0.15$, optionally annealed from zero during the first epochs. Evaluation-time random missingness uses separate stress-test probabilities q_I and q_T over truly observed modalities; all-missing draws are resampled or reported as abstentions.

When a modality is dropped or naturally unavailable, its token sequence is replaced by learned absent-modality tokens:

$$\bar{H}_I = \begin{cases} H_I^0, & m_I = 1, \\ R_I, & m_I = 0, \end{cases} \quad \bar{H}_T = \begin{cases} H_T^0, & m_T = 1, \\ R_T, & m_T = 0, \end{cases} \quad (11)$$

where R_I and R_T are single-token learned missing-modality embeddings. The explicit e_m embedding lets the network distinguish missingness from observed negative evidence. Fusion layers consume these tensors by setting $H_I^0 \leftarrow \bar{H}_I$ and $H_T^0 \leftarrow \bar{H}_T$ for the sampled availability pattern before the first cross-attention layer.

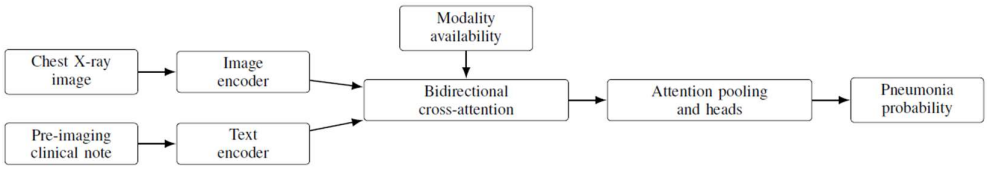

Fig. 1. MDCAT processes image and text streams separately, exchanges information through gated bidirectional cross-attention, and conditions predictions on explicit modality availability.

D. Training Objective

The supervised loss is class-weighted binary cross-entropy:

$$\mathcal{L}_{\text{sup}} = -\frac{1}{K}\sum_{k=1}^{K}\left[w_k^+ y_k \log \hat{y}_k + w_k^- (1-y_k)\log(1-\hat{y}_k)\right]. \tag{12}$$

Auxiliary unimodal supervision keeps each stream predictive:

$$\mathcal{L}_{\text{aux}} = \mathbb{1}[m_I = 1]\text{BCE}(\hat{y}_I, y) + \mathbb{1}[m_T = 1]\text{BCE}(\hat{y}_T, y). \tag{13}$$

For samples where both modalities are truly observed, MDCAT first computes a full-view prediction with $m = (1, 1)$. If the sampled training view drops one modality, the stopped-gradient full-view prediction $\hat{y}^{11}$ regularizes the masked prediction $\hat{y}^m$.

Let $a_I, a_T \in \{0, 1\}$ denote true modality availability before simulated dropout; the consistency term is

$$\mathcal{L}_{\text{cons}} = \mathbb{1}[a_I a_T = 1]\mathbb{1}[m \neq (1,1)] \times \frac{1}{K}\sum_{k=1}^{K} D_{\text{KL}}^{\text{Bern}}\left(\text{sg}(\hat{y}_k^{11}) \| \hat{y}_k^m\right), \tag{14}$$

where sg(·) stops gradients and D_{KL}^{Bern} is the Bernoulli KL divergence. The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}. \tag{15}$$

5. DATA CONSTRUCTION PROTOCOL

Table I summarizes recommended datasets and their role, and Fig. 2 shows the leakage-controlled pairing logic. The primary development route links MIMIC-CXR/MIMIC-CXR-JPG [4], [5], [23] to temporally eligible clinical context from MIMIC-IV-ED [24], using MIMIC-IV identifiers for linkage [25]. The default independent text source is ED chief complaint and other timestamped pre-CXR ED text; vitals and structured ED metadata are treated as optional structured baselines rather than note text. MIMIC-IV-Note [26] is not used as a primary independent note source in the default MIMIC route because its discharge summaries and radiology reports are excluded. It is reserved for explicitly labeled sensitivity or report-text analyses if note type and availability are appropriate. MIMIC-CXR and MIMIC-IV share *subject_id*, but only studies with verified patient or stay linkage are used for the paired multimodal cohort. CXR studies should be paired by *subject_id* plus *study_id/dicom_id* for imaging, then aligned to ED or hospital context using verified *subject_id* and available *hadm_id/stay_id*. Eligible text must have a document creation time, *storetime*, or ED event time no later than the CXR study/acquisition time t_i^*; charttime-only notes are excluded from the primary analysis unless availability before t^* can be verified. Order-time prediction can be reported as a stricter sensitivity analysis if an order timestamp is available. Interpretation-time text, discharge summaries, and same-study radiology reports are excluded from the primary text input.

A. Inclusion and Exclusion Criteria

The study unit is a patient-level CXR study. Inclusion criteria are: (i) frontal CXR image available, (ii) pneumonia label available from a documented label source, (iii) patient identifier available for leakage-safe splitting, and (iv) for full multimodal training, at least one eligible clinical-note or ED-text source available no later than t^*. Exclusion criteria are: (i) missing image data, (ii) duplicated or conflicting identifiers that cannot be resolved, (iii) note text or structured fields written or made available after t^* for the primary text modality, (iv) charttime-only notes whose availability before t^* cannot be verified, (v) untimestamped ED diagnosis fields in the primary input, and (vi) same-study radiology report as text input for the primary diagnostic experiment.

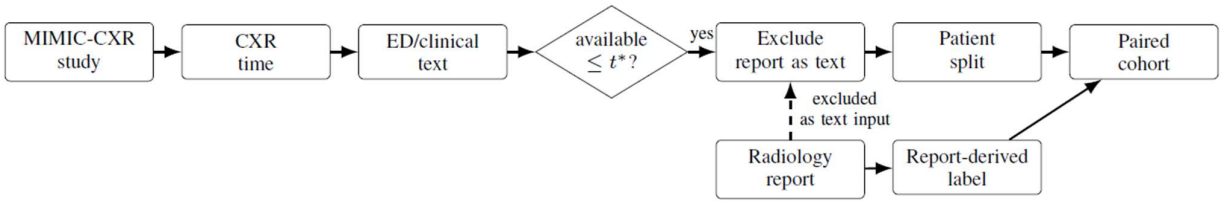

Fig. 2. Leakage-controlled cohort construction. Chest radiographs are paired only with clinical text available at or before the prespecified CXR acquisition time t^*. Same-study radiology reports may define labels but are excluded from the primary text input when labels are report-derived. Splits are performed at the patient level.

TABLE I
RECOMMENDED DATASETS AND THEIR INTENDED ROLE.

Dataset	Relevant content	Recommended use
MIMIC-CXR / MIMIC-CXR-JPG [4], [5], [23]	Chest radiographs, reports, study identifiers, structured report-derived labels	Primary source for CXR images and labels; reports excluded from primary text input
MIMIC-IV-ED [24]	ED stays, triage data, chief complaint, vitals, and timestamped pre-prediction events	Source for ED chief complaint and text-valued pre-CXR context; vitals and structured ED metadata are optional structured baselines; untimestamped diagnoses are excluded or ablated separately
MIMIC-IV-Note [26]	Discharge and radiology notes, linkable to MIMIC-IV	Not a primary independent text source in the default MIMIC route; use only for explicitly labeled sensitivity or report-text analyses when timing and note type are appropriate
CheXpert [3]	CXR images, reports, uncertainty labels	External or pretraining source; report text is not independent clinical evidence
RSNA Pneumonia Detection Challenge [6]	Expert pneumonia opacity annotations on NIH-derived images	External image-only validation with text unavailable
VinDr-CXR [15]	Radiologist-annotated chest X-rays	External image-only validation and label-definition stress test
PadChest / OpenI [27], [28]	Image-report pairs	Secondary report-pairing or pretraining experiments

B. Label Handling

For binary pneumonia detection, labels may be derived from MIMIC-CXR-JPG CheXpert / NegBio labels, RSNA expert annotations, or VinDr-CXR annotations. The primary MIMIC endpoint should be described as report-derived pneumonia mention or finding. The RSNA endpoint should be described as expert-labeled opacity possibly representing pneumonia, and the VinDr-CXR endpoint as radiologist-labeled pneumonia finding. These endpoints should be reported separately unless a prespecified harmonization rule is used. The primary analysis should either exclude uncertain labels or treat them as indeterminate. Sensitivity analyses should map uncertainty to positive, negative, and soft labels. When using report-derived labels, the manuscript should state that the target is radiological pneumonia mention or opacity compatible with pneumonia rather than confirmed clinical pneumonia.

6. EXPERIMENTAL PROTOCOL

A. Splits and Preprocessing

All splits must be patient-level. All studies and images from the same patient must remain in a single split. The CXR pipeline should resize images to 224 or 320 pixels, normalize intensities, and use clinically conservative augmentations such as small rotations, mild random crops, and contrast jitter. Vertical flips and aggressive transformations should be avoided. The text pipeline should remove protected placeholders as needed, segment long notes into clinically meaningful sections, and restrict tokens to information available at prediction time. A maximum length of 256 or 512 tokens is a practical default; longer notes can be represented by section pooling or hierarchical encoding.

B. Baselines

The evaluation should compare MDCAT against:

- **Image-only models:** DenseNet-121, ResNet-50, Efficient-Net, ConvNeXt, Swin Transformer, or Vision Transformer (ViT).
- **Text-only models:** term frequency-inverse document fre-quency (TF-IDF) with logistic regression, ClinicalBERT, BioClinicalBERT, PubMedBERT, and hierarchical clinical-note encoders.
- **Structured clinical baselines:** logistic regression, XG-Boost, or multilayer perceptron over age, sex, permitted race/ethnicity variables, triage vitals, chief complaint em-beddings, and timestamped pre-prediction ED metadata.
- **Fusion baselines:** late-fusion probability averaging, con-catenation fusion, gated multimodal units, cross-attention without modality dropout, and missing-aware late-fusion ensembles.

C. Missing-Modality Robustness

Fig. 3 summarizes the required missing-modality evaluation settings. The central robustness experiment evaluates the same trained model under:

*Author for Correspondence: sowjanya_it@mvsrec.edu.in

1) full input: CXR plus eligible clinical text;

2) image only: clinical text missing;

3) text only: CXR missing;

4) random missingness with modality drop probabilities 0.10, 0.25, 0.50, and 0.75;

5) real missingness from naturally absent notes or incomplete studies;

6) external image-only validation with text unavailable;

7) acquisition subgroup analysis: portable AP versus PA radiographs;

8) image corruption tests: lower resolution and brightness/contrast shift;

9) text corruption tests: note truncation, entity masking, and negation-preserving token masking.

Setting	CXR	Note	Use
Full input	✓	✓	reference
Image only	✓	×	missing text
Text only	×	✓	missing image
Stochastic	drop q_I	drop q_T	stress test
External	✓	×	validation

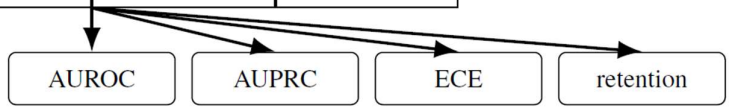

Fig. 3. Missing-modality evaluation matrix. The same trained model is evaluated under full input, image-only input, text-only input, stochastic missingness, and external image-only validation. Robustness is reported using absolute performance drop, retention relative to full-input performance, calibration error, and worst-case missing-modality performance.

The principal metrics are area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), and expected calibration error (ECE).

Report absolute metric drop from full-input inference, performance retention ratio, worst-case missing-modality score, and area under the missingness-performance curve. Retention is interpreted only together with absolute AUROC/AUPRC, absolute drop, and a prespecified non-inferiority or accept-ability margin; otherwise high retention with poor absolute performance is descriptive rather than evidence of robustness.

D. Metrics and Statistical Analysis

The prespecified primary robustness endpoint is MDCAT AUROC retention under image-only inference relative to its own full-input inference, because the main claim concerns missing clinical-note robustness. Between-model comparisons against image-only baselines are secondary and should use image-only AUROC and AUPRC rather than retention. Other secondary metrics are full-input AUROC, AUPRC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, balanced accuracy, Matthews correlation coefficient, Brier score, calibration slope/intercept, and ECE. AUPRC must be reported with pneumonia prevalence. Clinically relevant operating points include sensitivity at fixed 90% or 95% specificity and specificity at fixed 90% sensitivity. Thresholds are selected on validation data only.

Report 95% confidence intervals using patient-clustered bootstrap resampling as the primary uncertainty estimate. The primary contrast is within-model retention for MDCAT full-input versus MDCAT image-only inference. Paired DeLong tests for AUROC and McNemar tests for thresholded predictions should be used only after aggregating to one prediction per patient or study; otherwise use paired bootstrap or permutation tests that preserve patient clustering. Use paired bootstrap or permutation tests for AUPRC; secondary model, metric, missingness, and subgroup comparisons should be labeled exploratory unless corrected with Holm or Benjamini-Hochberg procedures.

E. Ablation Plan

Table II specifies the minimum ablations needed to isolate architectural and training contributions before any empirical performance claim is made.

TABLE II
ABLATIONS REQUIRED TO ISOLATE MODEL CONTRIBUTIONS.

Ablation	Purpose
Image-only encoder	Measures radiograph-only performance
Text-only encoder	Measures independent clinical-note signal
Late fusion without cross-attention	Tests whether token-level interaction is useful
Early concatenation transformer	Compares against simple joint-token fusion
Cross-attention without modality dropout	Isolates robustness contribution of modality dropout
Modality dropout without consistency loss	Tests full-to-masked regularization
No auxiliary unimodal heads	Tests branch-level supervision
Zero missing tokens vs learned missing tokens	Tests absent-modality representation
Image-to-text attention only	Tests directionality of cross-attention
Text-to-image attention only	Tests the opposite direction
Report-text variant	Leakage-prone secondary analysis, not primary diagnostic claim

7. IMPLEMENTATION DETAILS

A practical MDCAT implementation uses a ViT, DenseNet, or Swin image encoder pretrained on CXR data, a clinical-domain transformer text encoder, $L = 2$ cross-attention layers, hidden size $d = 512$, eight attention heads, feed-forward dimension 2048, dropout 0.1, and pre-normalization. Train with AdamW, image/text encoder learning rates of 1×10^{-5} to 3×10^{-5}, a fusion/head learning rate of approximately 1×10^{-4}, weight decay 0.01, mixed precision, gradient clipping at 1.0, and early stopping on validation AUROC or AUPRC. Hyperparameters should be selected on validation data and reported in full.

Class imbalance can be handled with class-weighted binary cross-entropy. Focal loss is acceptable only if chosen before test evaluation or selected using validation data. For multi-label prediction, thresholds should be selected per label. For binary pneumonia prediction, threshold selection should target clinically relevant sensitivity-specificity tradeoffs.

8. ETHICS, REPRODUCIBILITY, AND DEPLOYMENT CONSIDERATIONS

This study uses deidentified public clinical datasets that require compliance with data use agreements and credentialing where applicable. Raw data from MIMIC or CheXpert, including images, reports, labels, and clinical-note text where applicable, must not be redistributed. Reproducibility artifacts should include cohort extraction scripts, split generation scripts, preprocessing configuration, model code, hyperparameters, random seeds, and trained checkpoints where licenses permit. A cohort flow diagram should report included studies, excluded uncertain labels, missing notes, missing images, duplicate patients, image views, and external validation counts.

Clinical deployment should not be inferred from retrospective benchmark performance. The model should be considered decision support only after prospective validation, local calibration, failure-mode analysis, and workflow review. Subgroup performance should be audited across age, sex, race/ethnicity where available and permitted, care setting, scanner or view type, and patient acuity. Because notes can encode social and demographic information, subgroup auditing must examine whether text improves robustness or amplifies bias.

9. LIMITATIONS

This protocol does not report empirical results. Therefore it cannot support claims of superiority, state-of-the-art performance, radiologist-level performance, clinical utility, or deployment readiness. Public CXR datasets often use report-derived labels; such labels may not equal adjudicated clinical pneumonia. In the public MIMIC route, independent pre-CXR free text is limited mainly to ED chief complaint and related text-valued ED fields, while richer pre-CXR clinician notes require local institutional data or another verified source. Temporally eligible clinical notes may be missing, sparse, or available for only a subset of studies, creating selection bias. MIMIC data come from a single health system, and external image-only validation cannot fully test multimodal generalization. Finally, modality dropout simulates missingness, but real missingness can be informative, nonrandom, and institution-specific.

10. CONCLUSION

MDCAT is a leakage-aware multimodal transformer for robust pneumonia detection from chest X-rays and temporally eligible clinical notes. The architecture uses bidirectional cross-attention, explicit modality availability, modality dropout, auxiliary unimodal supervision, and prediction consistency to support full-input, image-only, and text-only inference. The proposed protocol emphasizes patient-level splitting, strict note timing, report-text leakage controls, missing-modality robustness, external validation, calibration, and subgroup analysis. These design choices make the framework suitable for a defensible empirical study once dataset extraction and model training are completed.

DATA AND CODE AVAILABILITY

No new dataset is released with this manuscript. The proposed experiments rely on public or credentialed-access datasets cited in the manuscript. Code should be released with scripts that regenerate cohorts from the original sources, subject to each dataset's license and data use agreement.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

[1] World Health Organization, "Pneumonia in children," 2022, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia. Accessed: Apr. 24, 2026.

[2] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2097–2106.

[3] J. Irvin, P. Rajpurkar, M. Ko et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 590–597.

[4] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," Scientific Data, vol. 6, no. 317, 2019.

[5] A. Johnson, M. Lungren, Y. Peng et al., "MIMIC-CXR-JPG - chest radiographs with structured labels," PhysioNet, Mar. 2024, version 2.1.0.

[6] G. Shih, C. C. Wu, S. S. Halabi et al., "Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia," Radiology: Artificial Intelligence, vol. 1, no. 1, p. e180041, 2019.

[7] P. Rajpurkar, J. Irvin, K. Zhu et al., "CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning," 2017, arXiv:1711.05225.

[8] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in Proceedings of the 7th Machine Learning for Healthcare Conference, ser. Proceedings of Machine Learning Research, vol. 182. PMLR, 2022, pp. 2–25.

[9] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, "GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3942–3951.

[10] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 3876–3887.

[11] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2019, pp. 6558–6569.

[12] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining taskagnostic visiolinguistic representations for vision-and-language tasks," in Advances in Neural Information Processing Systems, vol. 32, 2019.

[13] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multi-modal transformers robust to missing modality?" in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[14] D. S. Kermany, M. Goldbaum, W. Cai et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," Cell, vol. 172, no. 5, pp. 1122–1131.e9, 2018.

[15] H. Q. Nguyen, K. Lam, L. T. Le et al., "VinDr-CXR: An open dataset of chest x-rays with radiologist's annotations," Scientific Data, vol. 9, no. 429, 2022.

[16] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9049–9058.

[17] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in International Conference on Learning Representations, 2021.

[20] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "ModDrop: Adaptive multi-modal gesture recognition," 2015, arXiv:1501.00102.

[21] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clin-ical notes and predicting hospital readmission," 2019, arXiv:1904.05342.

[22] E. Alsentzer, J. R. Murphy, W. Boag et al., "Publicly available clinical BERT embeddings," in Proceedings of the 2nd Clinical Natural Language Processing

Workshop. Association for Computational Linguistics, 2019, pp. 72–78.

[23] A. Johnson, T. Pollard, R. Mark, S. Berkowitz, and S. Horng, “MIMIC-CXR Database,” PhysioNet, Jul. 2024, version 2.1.0.

[24] A. Johnson, L. Bulgarelli, T. Pollard, L. A. Celi, R. Mark, and S. Horng, “MIMIC-IV-ED,” PhysioNet, Jan. 2023, version 2.2.

[25] A. E. W. Johnson, L. Bulgarelli, L. Shen et al., “MIMIC-IV, a freely accessible electronic health record dataset,” Scientific Data, vol. 10, no. 1, 2023.

[26] A. Johnson, T. Pollard, S. Horng, L. A. Celi, and R. Mark, “MIMIC-IV-Note: Deidentified free-text clinical notes,” PhysioNet, Jan. 2023, version 2.2.

[27] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vaya, “PadChest: A large chest x-ray image dataset with multi-label annotated reports,” Medical Image Analysis, vol. 66, p. 101797, 2020.

[28] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman et al., “Preparing a collection of radiology examinations for distribution and retrieval,” Journal of the American Medical Informatics Association, vol. 23, no. 2, pp. 304–310, 2016.