Exploring the Capabilities of Large Language Model Encoders for Image-Text Retrieval in Chest X-rays

Hanbin Ko, Rong Yang, Gihun Cho, Inhyeok Baek, Donguk Kim, Joonbeom Koo, Changi Kim, Dongheon Lee, *Member, IEEE*, and Chang Min Park

***Abstract*— Multimodal learning from paired medical images and clinical text is a central challenge in medical data-driven informatics, where effective cross-modal alignment is critical for scalable analysis and retrieval. In chest radiography, vision–language pretraining is constrained by heterogeneous radiology reports that contain abbreviations, impression-only notes, and institution-specific writing styles. Unlike general-domain settings, naively aggregating large collections of noisy reports can plateau or even degrade multimodal learning when reporting styles differ substantially. We propose a domain-adapted bidirectional large language model text encoder for chest radiograph reports, trained with masked token prediction and supervised contrastive learning on stylistically diverse but clinically equivalent report variants to produce robust, generalizable text embeddings. We then integrate this encoder into a dual-tower contrastive vision–language framework using parameter-efficient adaptation to improve image–text alignment. Across 1.6 million paired studies from public datasets and a de-identified hospital cohort, the proposed models improve bidirectional retrieval accuracy and external generalization, achieving GREEN scores of 0.308 on MIMIC-CXR and 0.618 on Open-I, while reducing the degradation observed when abbreviation-rich, impression-only hospital reports are added to training. Significance: Robust cross-modal embeddings enable scalable retrieval and multimodal representation learning from routine clinical data for biomedical and health informatics.**

Corresponding authors: Dongheon Lee (dhlee13@snu.ac.kr) and Chang Min Park (morphius@snu.ac.kr).

H. Ko, G. Cho, and C. Kim are with the Interdisciplinary Program in Bioengineering, Seoul National University Graduate School, Seoul, Republic of Korea; and the Integrated Major in Innovative Medical Science, Seoul National University Graduate School, Seoul, Republic of Korea (e-mail: lucasko1994@snu.ac.kr; gihuncho@snu.ac.kr; fr2zyroom@snu.ac.kr).

R. Yang is with the Department of Radiology, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China (e-mail: dryangrong@zju.edu.cn).

I. Baek, D. Kim, and J. Koo are with the Seoul National University College of Medicine, Seoul, Republic of Korea (e-mail: bih1122@snu.ac.kr; drinkuranium@snu.ac.kr; ciderest@snu.ac.kr).

D. Lee and C. M. Park are with the Department of Radiology, Seoul National University College of Medicine, Seoul National University Hospital, Seoul, Republic of Korea; the Institute of Medical and Biological Engineering, Seoul National University Medical Research Center, Seoul, Republic of Korea; and the Institute of Radiation Medicine, Seoul National University Medical Research Center, Seoul, Republic of Korea (e-mail: dhlee13@snu.ac.kr; morphius@snu.ac.kr).

***Index Terms*— CXR retrieval, Large language model encoder, Medical CLIP, Vision-language pretraining**

I. Introduction

Early advances in medical image analysis primarily focused on single-modality tasks such as classification, detection, and segmentation [1]–[3]. While these models achieved strong performance on curated datasets, they required extensive expert annotations and captured only visual patterns, overlooking the clinical information described in radiology reports. This limitation has motivated a shift toward methods that can jointly learn from visual and textual data.

Vision-language pretraining (VLP) has recently emerged as an effective solution for integrating these two modalities. By learning from paired image–text data, VLP models reduce reliance on manual labels and enable tasks such as report generation, retrieval, and cross-modal reasoning. Among these, **CLIP** [4] is particularly influential for its simple contrastive learning framework and strong zero-shot performance, motivating adaptations for radiology [5], [6].

Adapting CLIP-style frameworks to radiology, however, introduces distinct challenges. Radiology reports differ substantially from natural-language captions, featuring specialized terminology, abbreviations, and linguistic patterns such as negation and uncertainty [7]. They also vary across institutions in section composition, ranging from full Findings–Impression reports to impression-only summaries, and in institution-specific shorthand, such as “BLLF” for bilateral lower lung field and “PTX” for pneumothorax. In addition, temporal expressions may refer to prior studies that are not visible in the current image, further complicating image–text alignment [8], [9]. These characteristics hinder the direct transfer of general-domain VLP models and necessitate domain-specific adaptation. Most existing medical VLP systems employ domain-specialized BERT-based text encoders [8], [10], [11], which outperform generic encoders but remain vulnerable to stylistic variability, abbreviation-heavy reporting, and section-level mismatch in radiology reports [9], [12]. Consequently, simply aggregating stylistically divergent report corpora does not necessarily produce monotonic gains [13]; as shown in Section V, adding impression-only or abbreviation-heavy datasets can degrade retrieval performance unless the text encoder is robust to such report-style heterogeneity.

Large language model (LLM) based embedding frameworks have been reported to provide a stronger foundation for text representation than BERT based encoders [14], [15]. These models produce high capacity embeddings that capture nuanced semantic variation across different phrasings [14], [16], such as "no pleural effusion" and "pleural spaces are clear," while effectively handling longer clinical contexts. These properties make LLM based embeddings well suited for radiologic reports, where clinically equivalent findings are frequently expressed using diverse phrasing and structure. We illustrate this setting using chest X-ray (CXR) imaging, one of the most commonly performed radiologic examinations worldwide, whose reports exhibit substantial variability in structure, length, and expression. Motivated by these gaps, this study asks whether a domain-adapted bidirectional LLM encoder can provide more robust CXR report representations than BERT-based encoders, and whether such improved text representations can be transferred to a CLIP-style framework to improve clinically faithful image–text retrieval under heterogeneous reporting styles.

To address this gap, we propose two complementary components: an LLM-based text encoder (*LLM2VEC4CXR*) and a multimodal framework (*LLM2CLIP4CXR*) that integrates the text encoder with a vision encoder for improved image–text alignment.

(1) LLM2VEC4CXR: a domain-adapted LLM encoder for CXR reports that is robust to abbreviation-heavy and stylistically heterogeneous clinical language. It combines masked token prediction, supervised contrastive learning, and latent attention pooling to produce stable, clinically meaningful text embeddings across both Findings and Impression sections.

(2) LLM2CLIP4CXR: a multimodal framework that integrates LLM2VEC4CXR with a vision encoder for image–text alignment. Using LoRA-based parameter-efficient fine-tuning, it transfers improved text understanding to the image domain, enhancing retrieval accuracy and cross-dataset generalization.

Trained on **1.6M CXRs** from public and private sources, our models outperform BERT-based and prior medical CLIP frameworks on both standard and clinically oriented metrics. Together, these results emphasize that advancing clinical text comprehension is more decisive for multimodal generalization than simply increasing data volume.

II. Related works

Our work builds on two primary research areas: medical vision-language modeling and the use of large language models as text encoders. We review both areas and then position our contributions within this context.

A. Medical vision-language pretraining

Contrastive learning on paired images and text, as introduced by CLIP [4], has proven effective for transferable multimodal representations. Several medical adaptations follow this paradigm, including ConVIRT [17], CheXzero [6], BioViL/BioViL-T [8], [18], and MedCLIP [19], which pair chest X-rays with reports in CLIP-style frameworks. These models typically replace CLIP's text encoder with biomedical BERT variants such as BioClinicalBERT [11], PubMedBERT [20], or CXR-BERT [8]. While these encoders improve biomedical language coverage, they can still be challenged by sectioned report structure, institution-specific abbreviations, negation, uncertainty, and the semantic relationship between *Findings* and *Impression* sections. Thus, the text encoder remains a key bottleneck in medical VLP.

CXR report generation represents a related but distinct line of work [21]. Conventional transformer-based and recent LLM/VLM-based systems typically couple visual encoders with autoregressive language decoders to synthesize free-text radiology reports, as in MAIRA-2 [22], CheXagent [23], CXR-LLAVA [24], and META-CXR [25]. These methods primarily optimize image-to-report decoding, whereas our study addresses the input-side representation problem by adapting an LLM into a bidirectional, domain-specialized report encoder for image–text retrieval and multimodal alignment under heterogeneous CXR reporting styles.

B. Evaluation challenges in CXR retrieval

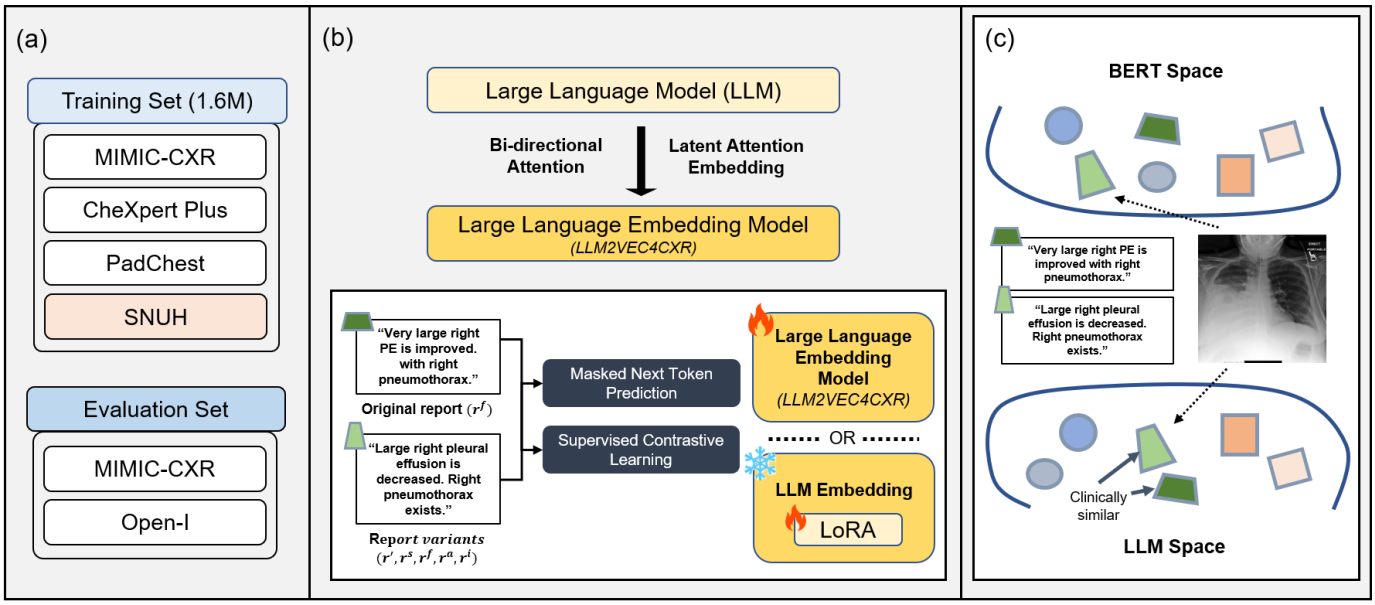
Most clinical CLIP retrieval studies report recall at top-k [13], [17], which measures exact string matching between queries and reports. Yet radiology reports frequently contain multiple semantically equivalent phrasings, so retrieving a clinically correct but textually different report is penalized as an error. To address this, radiology report generation research [22], [23], [26] has proposed clinically oriented metrics such as CheXbert F1 [27], RadGraph F1 [28], and GREEN [29], which assess whether retrieved text captures the correct entities and relations. We adopt this direction, applying clinically relevant metrics to retrieval and focusing on *clinical faithfulness* rather than surface-level similarity.

C. LLMs as text encoders

Beyond generation, LLMs are increasingly adapted for encoding tasks. Methods such as LLM2VEC [14] and NV-Embed [16] show that LLM representations can outperform specialized encoders when fine-tuned for embeddings. Their use has expanded to domains such as table analysis [30] and recommendation systems [31]. Early biomedical efforts include BMRetriever [32], which tunes LLMs for clinical retrieval, but applications to multimodal medical tasks remain limited. In radiology, there has been little exploration of LLM encoders for chest X-ray report alignment, despite their potential to capture semantic variation across diverse reporting styles.

D. Positioning and contributions

In summary, prior studies have primarily adapted BERT-style text encoders for medical CLIP frameworks, but these models still lack strong clinical alignment and rely heavily on recall-based evaluation. Moreover, most existing work has been trained on curated public corpora, whereas real-world private reports are substantially noisier, abbreviated, and often restricted to shorter sections such as *Impression* notes.



SNUH: Seoul National University Hospital, LoRA: Low-Rank Adaptation

Fig. 1. Overview of the proposed framework. (a) Training and evaluation datasets combining public and private data for 1.6M CXR image–report pairs. (b) **LLM2VEC4CXR**: domain-adapted LLM trained with masked token prediction and contrastive learning. (c) Compared to BERT, LLM-based embeddings cluster clinically similar phrases, yielding more robust representations for multimodal learning.

Our work addresses these limitations through *LLM2VEC4CXR* and *LLM2CLIP4CXR*, which leverage LLM encoders to capture richer clinical semantics and enable more robust image-text alignment. Together, they advance medical vision-language pretraining by improving cross-dataset generalization and enhancing the clinical relevance of retrieval performance.

III. Methods

We introduce two main models: *LLM2VEC4CXR*, a specialized LLM encoder for CXR reports, and *LLM2CLIP4CXR*, which integrates this encoder with a vision backbone in a two-tower CLIP-style framework. An overview of the proposed framework is shown in Fig. 1.

A. LLM2VEC4CXR

CXR reports contain long sentences, abbreviations, temporal references, and variable formatting, which make standard BERT encoders brittle. *LLM2VEC4CXR* addresses the complexity of radiology language. Following LLM2VEC [14], we remove the causal attention mask from decoder-only architectures to enhance bidirectional context encoding, and adopt latent attention pooling (LAP) for more informative global embeddings. The model is trained with two objectives: Masked token prediction (MTP) and supervised contrastive learning.

1) Data generation: To expose the model to clinically equivalent but stylistically diverse inputs, we generate multiple variants of each report (r) using *Gemini2.0-Flash* [33] and *Deepseek-R1-Distill-Qwen-14B* [34]. These include:

- **Clinically similar reports (r'):** Rephrasings that preserve meaning while altering style.
- **Sentence splitting (r^s):** Decomposition of multi-finding sentences into atomic observations [22].
- **Omitting temporal references (r^o):** Removal of temporal expressions and change descriptors.
- **Anatomical partitioning (r^d):** Segmentation by anatomical region, followed by structured recombination.
- **Summarization pairs:** Linking *Findings* (r^f) with *Impression* (r^i) sections.

To assess whether these synthetic variants were clinically valid as positive pairs, we performed an expert review of 50 source–synthetic report pairs sampled from the training-positive pool. A thoracic radiologist with approximately 15 years of clinical experience assessed whether each synthetic report preserved the current clinically relevant content, including findings, negation or uncertainty, anatomical location, laterality, and severity. Each pair was categorized as clinically acceptable or clinically unacceptable. The radiologist judged 46 of 50 pairs (92%) as clinically acceptable positive pairs for contrastive training, supporting the use of these synthetic variants as training positives.

All variants also participate in MTP pretraining, strengthening robustness to heterogeneous report styles. To capture clinical shorthand, we additionally include MIMIC's *Indication* fields, which contain frequent acronyms. A summary is given in Table I.

2) Bidirectional adaptation of a decoder-only LLM: Let a report be tokenized into a sequence $\mathbf{x} = (x_1, \ldots, x_L)$. We convert a decoder-only transformer into a bidirectional encoder by removing the causal (triangular) attention mask. Concretely, in each self-attention layer we compute

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}\right)\mathbf{V}, \qquad (1)$$

where d is the head dimension and $\mathbf{M}$ is an attention mask. In a standard decoder, $\mathbf{M} = \mathbf{M}^{\mathrm{causal}}$ with $(\mathbf{M}^{\mathrm{causal}})_{ij} = 0$ if

$j \le i$ and $-\infty$ otherwise. We set $\mathbf{M} = \mathbf{0}$ to allow each token to attend to both preceding and following context, producing bidirectional hidden states $\mathbf{H} = f_\theta(\tilde{\mathbf{x}}) \in \mathbb{R}^{L \times d_{\text{model}}}$. This conversion is required because the decoder-only LLM is used as a report encoder rather than as an autoregressive generator. For CXR reports, full-context attention is appropriate because findings are commonly organized by anatomy rather than by a fixed causal order, so the final report embedding should integrate observations, modifiers, and section-level information across the entire report.

3) Masked token prediction: We pretrain the bidirectional text encoder using masked token prediction (MTP), following LLM2VEC [14]. Given a report, a subset of tokens is replaced with a mask symbol, and the model is trained to recover the original tokens from bidirectional context:

$$\mathcal{L}_{\text{MTP}} = -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \log p_\theta(x_i \mid \tilde{\mathbf{x}}). \tag{2}$$

Because the causal mask is removed, masked tokens are predicted from both left and right context, encouraging global report understanding rather than next-token continuation.

4) LAP for global text embeddings: For contrastive learning, token-level hidden states must be converted into a single report embedding. We use latent attention pooling (LAP) [16], in which learnable latent queries attend to token representations and produce a pooled report vector. The pooled vector is passed through a projection head and ℓ_2-normalized before contrastive learning. Compared with mean pooling, LAP allows the model to learn which report tokens contribute most to the global clinical representation.

5) Supervised contrastive learning on report variants: After MTP pretraining, we fine-tune the encoder with supervised contrastive learning to cluster clinically equivalent texts (e.g., original reports and their generated variants) while separating unrelated reports. Let a mini-batch contain B input texts, each mapped to a normalized embedding $\{\mathbf{u}_i\}_{i=1}^{B}$ by the text encoder and projection head. Each sample i has a supervision label y_i indicating membership in a positive group (e.g., same underlying report instance, same finding label, or an instruction-aligned pair). Define the positive set for anchor i as

$$\mathcal{P}(i) = \{\, p \in \{1, \dots, B\} \setminus \{i\} \mid y_p = y_i \,\}. \tag{3}$$

Using cosine similarity (equivalently dot product of unit vectors), $\text{sim}(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{u}_i^\top \mathbf{u}_j$, the supervised contrastive (InfoNCE-style) loss is

$$\mathcal{L}_{\text{C}} = -\frac{1}{B} \sum_{i=1}^{B} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp\left(\mathbf{u}_i^\top \mathbf{u}_p / \tau\right)}{\sum_{a \in \{1,\dots,B\} \setminus \{i\}} \exp\left(\mathbf{u}_i^\top \mathbf{u}_a / \tau\right)} \tag{4}$$

where $\tau > 0$ is a temperature hyperparameter. This objective explicitly pulls together embeddings of clinically consistent rephrasings/sections while pushing apart embeddings from different studies.

B. LLM2CLIP4CXR

LLM2CLIP4CXR integrates the domain-adapted text encoder *LLM2VEC4CXR* with a vision encoder in a dual-tower CLIP-style framework. The vision tower encodes the CXR image and the text tower encodes the paired report, with both embeddings projected into a shared normalized representation space.

Given a batch of image–report pairs, we optimize the standard symmetric CLIP contrastive loss, which aligns matched image–report pairs and separates mismatched pairs within the batch. During training, the vision encoder is fully optimized, the projection layers of both encoders are trained from scratch, and the text encoder is adapted using LoRA-based updates. This design transfers clinically informed report representations into the multimodal space while keeping adaptation of the LLM backbone efficient.

IV. Experiments

We describe the datasets, implementation details, preprocessing steps, and evaluation methodology used in our experiments.

A. Datasets

LLM2VEC4CXR is trained on the training splits of MIMIC [35], CheXpert-plus [36], and the generated variants described in Table I. For *LLM2CLIP4CXR*, we start with paired samples from MIMIC-CXR and incrementally add CheXpert-plus, PadChest [37], and a dataset from Seoul National University Hospital (SNUH; IRB no. 2405-070-1536), resulting in a combined training set of around 1.6M studies. Only frontal X-rays are used. To exploit the LLM's ability to handle longer text sequences, we retain full *Findings* or *Impression* sections when constructing non-sectioned training corpora. Notably, these datasets differ substantially in reporting style: MIMIC provides full structured reports, PadChest and SNUH contain impression-only summaries, and SNUH reports often consist of abbreviated notes. This heterogeneity provides a natural testbed for robustness to noise and style variation.

TABLE I
COUNTS OF PREPROCESSED REPORTS USED FOR LLM2VEC4CXR TRAINING.

Pair Type	Count
Original reports	319,564
Split reports	178,993
Prior-omitted reports	169,491
Anatomical-partitioned reports	178,096
Clinically similar reports	272,230
Total	1,118,374

We evaluate on two datasets: a held-out MIMIC-CXR internal validation split that is disjoint from training and tuning at the study level, and the *Open-I* [38] external validation set. To reduce temporal ambiguity, we exclude Open-I reports containing temporal expressions. The exclusion keywords are adopted from BioViL-T [8], but we apply them only as a

filter: any report containing one or more of these keywords is removed from the test set. We then retain a single comprehensive report per study to avoid multiple time points. Dataset statistics are summarized in Table II.

TABLE II
OVERVIEW OF THE DATASETS.

Dataset	Train	Tune	Int. Val.	Ext. Val.
MIMIC-CXR [35]	247,648	2,032	3,394	–
CheXpert-plus [36]	190,668	202	–	–
PadChest [37]	60,261	1,604	–	–
Open-I [38]	–	–	–	1,825
SNUH	1,136,410	1,328	–	–

Abbreviations: Int. Val., internal validation; Ext. Val., external validation; SNUH, Seoul National University Hospital.

B. Additional input processing

To improve section awareness, we use two complementary input-side mechanisms. First, we prepend inline section placeholders (`[FINDINGS]`, `[IMPRESSION]`) to the corresponding text span when training section-aware variants of *LLM2CLIP4CXR*. Second, we prepend a fixed task-type instruction prompt indicating the training objective (semantic similarity, summarization, or finding-level classification; see Table IX). The two mechanisms are orthogonal – placeholders mark report sections, prompts mark the training objective — and may co-occur in a single input. This section-aware prompting was particularly important for integrating datasets with impression-only reports (e.g., PadChest, private corpus), where otherwise the model could underfit detailed findings and overfit condensed styles.

C. Implementation details

1) Model architecture: For the text tower, we build on the LLM2VEC [14] encoder and construct *LLM2VEC4CXR* in two variants: (i) a *LoRA-based parameter-efficient update*, where only low-rank adapters are trained while keeping the backbone frozen, and (ii) a *fully fine-tuned version*, where all model parameters are updated. For *LLM2CLIP4CXR*, we extend the text encoder with an additional LoRA layer and a projection head. The vision tower is an EVA02-CLIP-L/14 backbone, resized for 448×448 pixels, with its own projection head of dimension 1280. Both projection layers are trained from scratch.

As a baseline, we trained a standard BERT [39] model under the same settings and integrated it into a CLIP-style framework, referred to as *BERT4CXR* and *CLIP4CXR*, respectively. Our *LLM2VEC4CXR* models use either *Llama3.2-1B* or *Llama3-8B* [40] as backbone LLMs. All standard medical CLIP variants are retrained on our dataset for consistent comparison.

2) Training configuration: All models are trained on four NVIDIA A6000 GPUs. *LLM2VEC4CXR* is pretrained with MTP for one epoch and further optimized with supervised contrastive learning. Latent attention pooling [16] is used to form global text embeddings. CLIP training configurations follow LLM2CLIP [15] with a per-GPU batch size of 256. A computational cost comparison of the 1B and 8B LLM2VEC4CXR variants is provided in Appendix II-A. All experiments were implemented in PyTorch (v2.4.1) using Python 3.8. More details on model configurations and versions are provided in our public repository: `https://github.com/lukeingawesome/llm2clip4cxr.git`.

D. Text-only evaluation

To evaluate the text encoder independently from the vision backbone, we designed five text-only tasks that probe different aspects of linguistic and clinical robustness. These tasks collectively assess the model's ability to handle reporting style variation, detect subtle semantic inconsistencies, and recognize clinically equivalent expressions across datasets. Each task isolates a specific challenge commonly encountered in radiology reporting—such as abbreviation comprehension, omission of temporal context, or paraphrased impressions—to provide a fine-grained analysis of text understanding. A detailed overview of the five evaluation tasks is provided in Table III.

E. Multimodal evaluation

For *LLM2CLIP4CXR*, we align frontal CXRs with reports. We first evaluate Top-k retrieval within each test set. To assess clinical correctness, we further apply the same metrics used in Task 5 (CheXbert F1, RadGraph F1, RaTEScore, and GREEN) when retrieving reports from the MIMIC train/validation pool. This avoids biases introduced by external datasets whose test reports may not cover all clinical variations.

Automated metrics may miss clinically important aspects of report quality. To complement them, we conducted a qualitative ranking study on 200 Open-I cases with three medical students (3, 8, and 44 months of training) and four large LLMs (*GPT-4o* [41], *Gemini 1.5 Pro* [33], *DeepSeek-V3* [42], and *DeepSeek-R1* [34]). We excluded half of the normal studies to ensure case diversity and compared ground-truth reports with retrieved and generated outputs. Candidate outputs were randomized to avoid order bias. Raters were instructed to rank the candidate reports by clinical accuracy relative to the ground truth, with ties permitted. All raters followed the LLM-RadJudge [43] protocol, which prioritizes critical clinical findings over minor or irrelevant details.

To anchor this qualitative evaluation to expert clinical judgment, a thoracic radiologist additionally reviewed a 72-case subset sampled from the same 200 Open-I cases. To reduce expert-rater burden while preserving representative model coverage, this subset included three systems: *LLM2CLIP4CXRall*, representing LLM-encoder retrieval; BC2C [44], representing BERT-based CLIP retrieval; and MAIRA-2 [22], representing report generation. The radiologist used the same randomized ranking protocol.

V. RESULTS

We report results for both the **text-only** setting, implemented using *LLM2VEC4CXR*, and the **multimodal** setting, implemented using *LLM2CLIP4CXR*. The evaluation focuses

TABLE III
OVERVIEW OF TEXT-ONLY EVALUATION TASKS FOR ASSESSING LLM2VEC4CXR. TOP-k RETRIEVAL IS USED UNLESS OTHERWISE NOTED.

Task	Name	Description	Metric(s)
1	Prior-omitted - Original matching	Given a prior-omitted report (r^o), the model retrieves the original report to evaluate sensitivity to omitted context.	Top-k retrieval
2	Report summarization	Given the *Findings* section, the model retrieves the corresponding *Impression*, assessing summarization and abstraction ability.	Top-k retrieval
3	Report error discrimination	Following *ReXErr* [45], three erroneous impressions (false negation, severity change, etc.) are synthesized per true report. The model must identify the correct impression given the *Findings* section.	Accuracy
4	Medical acronym understanding	Reports containing acronyms (e.g., "BLLF", "PTX") are paired with expert-refined, fully expanded versions (e.g., "bilateral lower lung field"). Tests robustness to lexical variation.	Top-k retrieval
5	Clinical similarity matching	Given a findings section from Open-I, the model retrieves the most similar MIMIC-CXR findings report to assess clinical equivalence.	RadGraph F1, CheXbert F1, RaTEScore, GREEN

on retrieval performance and clinically oriented measures. Overall, the LLM-based text encoders demonstrate stronger semantic alignment and generalizability than BERT-based and existing CLIP-style approaches, with consistent improvements on external validation.

A. Text-only results

Table IV compares *LLM2VEC4CXR* with baseline text encoders across the five evaluation tasks. Baselines include generic BERT [39], medical BERT variants (*PubMedBERT* [20], *BioClinicalBERT* [11], *CXR-BERT* [8], *ClinicalBERT* [10]), and two general-domain LLM encoders (*LLM2VEC* 1B and 8B). We also evaluate our proposed variants, *LLM2VEC4CXR* (1B/8B) and *LLM2VEC4CXR+* (fully fine-tuned), under the same protocol.

1) Task 1: Prior-omitted Original matching: *LLM2VEC4CXR* variants reach near-perfect top-5/top-10 performance. Top-1 exceeds 0.9 for the 8B and fully fine-tuned models, clearly surpassing *BERT4CXR* under the same protocol.

2) Task 2: Report summarization: General *LLM2VEC* already surpasses all BERT baselines, which struggle to link *Findings* and *Impression*, especially under style shifts in Open-I. Domain adaptation in *LLM2VEC4CXR* yields further gains, showing that LLM encoders capture summary relationships more reliably.

3) Task 3: Error discrimination: Most BERT variants operate below chance (≈ 0.25) while CXR-BERT performs slightly better. In contrast, *LLM2VEC4CXR+* attains 0.826 accuracy and *LLM2VEC4CXR (8B)* reaches 0.841. Notably, even the *LLM2VEC (1B)* model, without radiology-specific tuning, surpasses all BERT baselines. These results indicate that LLM-derived embeddings more reliably encode contradictions and subtle semantic edits (e.g., false negation, severity/location changes), enabling robust detection of clinically erroneous or different text.

4) Task 4: Acronym Comprehension: BERT-based encoders frequently fail to expand acronyms, whereas *LLM2VEC* and *LLM2VEC4CXR* both excel. Notably, the general *LLM2VEC* performs on par with BERT4CXR despite no explicit exposure to medical abbreviations.

5) Task 5: Clinical similarity: In cross-dataset matching, LLM-based encoders consistently outperform BERT variants on clinically oriented metrics. Domain-adapted *LLM2VEC4CXR* further improves alignment. A small drop in RadGraph F1 with full fine-tuning suggests sensitivity to local structure; however, GREEN remains high, indicating preservation of overall clinical correctness. Qualitative examples in Table V illustrate that *LLM2VEC4CXR* retrieves key findings reliably, whereas BERT-based models often hallucinate or misstate details.

6) Summary of text-only results: *LLM2VEC*-style encoders exhibit a stronger grasp of report structure and clinical semantics than BERT variants. Domain adaptation in *LLM2VEC4CXR* yields the best scores across most tasks. The 8B model achieves the highest absolute performance, while the fully fine-tuned 1B(*LLM2VEC4CXR+*) offers a favorable accuracy–efficiency trade-off (Appendix II-A); we therefore adopt *LLM2VEC4CXR+* for multimodal experiments.

B. Multimodal results

Given that LLM-based text encoders outperform BERT-based models in text-only tasks, we next examine whether *LLM2VEC4CXR* can transfer its domain-specific knowledge to the vision encoder within *LLM2CLIP4CXR*. Table VI summarizes retrieval performance and clinical evaluation metrics on the MIMIC (internal) and Open-I (external) test sets.

1) Comparison with CLIP baselines: Among models trained solely on MIMIC, BC2C [44] achieves strong clinical metrics, while CXR-CLIP [13] obtains the highest Top-k retrieval. On MIMIC, *LLM2CLIP4CXR* achieves comparable Top-k retrieval but clearly exceeds baselines in clinical efficacy measures, reflecting stronger alignment with clinically meaningful content. On Open-I, it surpasses all baselines on both Top-k retrieval and clinical metrics. This substantial gain in external validation underscores the advantage of LLM-based encoders,

TABLE IV
TEXT-ONLY EVALUATION OF *LLM2VEC4CXR* AND BASELINES. TASKS 1–2: TOP-k RETRIEVAL, TASK 3: ACCURACY, TASK 4: TOP-k RETRIEVAL, TASK 5: CLINICAL EFFICACY METRICS.

Model	Prior-omitted Original matching			Report summarization			Report error discrimination	Medical acronym understanding			Clinical similarity matching			
	@1	@5	@10	@1	@5	@10	Acc	@1	@3	@5	RadGraph	MF1	RaTE	GREEN
Base-BERT [39]	0.520	0.664	0.708	0.016	0.039	0.054	0.092	0.137	0.384	0.513	0.324	0.206	0.693	0.632
PubMedBERT [20]	0.447	0.585	0.638	0.030	0.066	0.097	0.223	0.274	0.470	0.598	0.309	0.227	0.696	0.638
BioClinicalBERT [11]	0.543	0.690	0.743	0.017	0.038	0.054	0.100	0.231	0.385	0.556	0.298	0.207	0.681	0.616
ClinicalBERT [10]	0.546	0.711	0.767	0.016	0.039	0.054	0.110	0.376	0.692	0.795	0.336	0.284	0.711	0.665
CXR-BERT-Specialized [8]	0.395	0.537	0.600	0.082	0.155	0.196	0.484	0.248	0.513	0.650	0.284	0.297	0.678	0.664
BERT4CXR	0.786	0.821	0.875	0.048	0.093	0.118	0.419	0.385	0.710	0.804	0.313	0.275	0.708	0.642
LLM2VEC(1B) [15]	0.576	0.766	0.826	0.053	0.093	0.116	0.552	0.368	0.675	0.769	0.392	0.682	0.724	0.662
LLM2VEC(8B) [15]	0.703	0.823	0.871	0.090	0.145	0.172	0.637	0.368	0.735	0.846	**0.404**	0.689	0.722	0.667
LLM2VEC4CXR(1B)	0.894	0.912	0.918	0.162	0.225	0.261	0.672	0.556	0.795	0.870	0.396	0.705	0.723	0.653
LLM2VEC4CXR(8B)	**0.933**	**0.989**	**0.994**	**0.212**	**0.305**	**0.352**	**0.841**	**0.611**	**0.875**	**0.926**	0.402	0.712	**0.727**	0.676
LLM2VEC4CXR+(1B)	0.918	0.984	0.989	0.203	0.295	0.339	0.826	0.607	0.854	0.910	0.373	**0.715**	0.724	**0.686**

Bold indicates the best result; underline indicates the second-best result.

TABLE V
TASK 5 QUALITATIVE EXAMPLE: OPEN-I QUERY VS. TOP-1 RETRIEVED MIMIC REPORT.

Model	Text
Ground Truth (Query)	Cardiomegaly is noted, with central pulmonary vascular prominence and coarsened interstitial markings, suspicious for developing interstitial pulmonary edema. No focal consolidation, pneumothorax, or definite effusion identified. No acute bony abnormality seen.
Base BERT	Patient is status post median sternotomy and CABG. Heart size is normal. Mediastinal contour is unremarkable. Fullness of the right hilum likely reflects known lung cancer. Mild upper zone pulmonary vascular redistribution suggests mild pulmonary vascular congestion. Coarse interstitial opacities are noted in the lung bases which may reflect chronic interstitial abnormality. Additionally, more focal opacity is seen within the periphery of the right lung base, a nonspecific finding. No pleural effusion or pneumothorax is identified. No acute osseous abnormalities seen.
ClinicalBERT	Heart size is normal. Mediastinal and hilar contours are unchanged with architectural distortion, superior bilateral hilar retraction, traction bronchiectasis, nodularity and fibrosis involving both upper lobes. No new focal consolidation, pleural effusion or pneumothorax is present. No pulmonary edema is detected, though there may be mild pulmonary vascular congestion. No acute osseous abnormality is visualized.
LLM2VEC (1B)	The heart is severely enlarged. There are perihilar hazy opacities with vascular indistinctness compatible with mild to moderate pulmonary edema. Assessment of the lung bases is limited due to technique, but there may be atelectasis. No large pleural effusion or pneumothorax is seen.
LLM2VEC4CXR+	The heart size is borderline enlarged. Mediastinal contours are unremarkable. Hilar contours are similar compared to the prior exam. Diffuse increased interstitial markings bilaterally suggest mild interstitial pulmonary edema. No pleural effusion or pneumothorax is identified. No acute osseous abnormality seen.

Blue indicates correct findings; red indicates incorrect/hallucinated statements; orange indicates uncertain/partially wrong statements.

which provide richer and more generalized text representations to guide the vision encoder.

2) Effect of additional training data: Several CLIP baselines show little gain—or even declines on MIMIC—after adding CheXpert-plus, likely due to formatting and style mismatches between sources. These results indicate that increasing data volume alone does not guarantee improved retrieval under report-style heterogeneity; robust text modeling is critical. In contrast, *LLM2CLIP4CXR* maintains or improves performance on both MIMIC and Open-I, demonstrating that the LLM-based text encoder transfers cross-domain semantics more robustly. Importantly, our model can encounter reports written in diverse styles and still learn the underlying clinical facts efficiently, leading to stable or improved clinical alignment. Overall, clinically oriented metrics remain strong and are comparable to those reported by recent generative systems.

Adding PadChest has mixed effects: while some metrics improve, top-k retrieval on MIMIC decreases, likely due to information density differences, since PadChest reports typically contain only the impression. This discrepancy is important because our private dataset (1.1M studies) also consists largely of impression-only reports. To mitigate this, we introduce section placeholders and instructions (Section IV-B), producing $\text{LLM2CLIP4CXR}_{\text{section}}$. This variant better distinguishes between *Findings* and *Impression*, mitigating performance drops and improving external generalization. By contrast, the $\text{CLIP4CXR}_{\text{section}}$ baseline with BERT shows little benefit, further underscoring the importance of stronger text encoders.

3) Private dataset integration: Finally, we augment training with a private corpus containing shorter, abbreviated, impression-only reports. Since *LLM2VEC4CXR* and *LLM2VEC* already demonstrate strong handling of abbreviations (see Table IV), we expected them to adapt well. Indeed, *LLM2CLIP4CXR* maintains or slightly improves performance on MIMIC and Open-I, while BERT-based models degrade more noticeably. This result highlights the adaptability of LLM-based encoders to diverse and noisy clinical reporting styles without compromising retrieval quality.

TABLE VI
MULTIMODAL IMAGE–TEXT RETRIEVAL RESULTS ON MIMIC AND OPEN-I.

		MIMIC						Open-I					
Model	**Dataset**	**@1**	**@10**	**RadGraph**	**MF1**	**RaTE**	**GREEN**	**@1**	**@10**	**RadGraph**	**MF1**	**RaTE**	**GREEN**
MAIRA2(7B) [22]	M, P, U			0.163	0.343	0.538	0.272			0.201	0.260	0.635	0.607
CheXagent(8B) [23]	M, C, P, B, Pu			0.154	0.248	0.510	0.256			0.154	0.186	0.604	0.539
CXR-LLAVA [24]	M, C, P, B			0.092	0.116	0.487	0.143			0.187	0.052	0.559	0.357
BC2C [44]	M	0.076	0.336	0.133	0.374	**0.512**	0.235	0.017	0.081	0.148	0.179	0.554	0.483
CXR-CLIP [13]	M	**0.175**	**0.553**	0.138	0.342	0.494	0.232	0.029	0.101	0.133	0.171	0.537	0.421
GLoRIA [46]	M	0.027	0.149	0.106	0.293	0.477	0.163	0.011	0.079	0.129	0.128	0.505	0.381
BioViL [18]	M	0.022	0.143	0.112	0.331	0.490	0.190	0.001	0.051	0.160	0.178	0.569	0.492
BioViL-T [8]	M	0.03	0.177	0.106	0.330	0.486	0.178	0.014	0.060	0.170	0.200	0.578	0.516
MedCLIP [19]	M	0.001	0.011	0.041	0.287	0.403	0.078	0.002	0.012	0.031	0.112	0.309	0.053
ConVIRT [17]	M	0.114	0.437	0.130	0.353	0.506	0.232	0.019	0.085	0.148	0.165	0.558	0.482
CLIP4CXR$_{\text{base}}$	M	0.132	0.468	0.137	0.365	0.508	0.226	0.021	0.089	0.145	0.155	0.562	0.488
LLM2CLIP4CXR$_{\text{base}}$	M	0.163	0.524	**0.173**	**0.412**	0.505	**0.289**	**0.045**	**0.146**	**0.182**	**0.231**	**0.592**	**0.572**
BiomedCLIP [20]	M, Pu	0.004	0.031	0.083	0.154	0.466	0.142	0.002	0.023	0.153	0.057	0.539	0.387
BC2C [44]	M,C	0.068	0.311	0.105	0.382	0.515	0.237	0.019	0.083	0.141	0.191	0.562	0.486
CXR-CLIP [13]	M,C	0.167	0.542	0.103	0.342	0.504	0.226	0.024	0.097	0.124	0.195	0.542	0.449
CLIP4CXR$_{\text{base}}$	M,C	0.107	0.429	0.099	0.354	0.508	0.242	0.014	0.077	0.146	0.172	0.575	0.495
LLM2CLIP4CXR$_{\text{base}}$	M,C	0.181	0.559	0.178	0.422	0.538	0.299	0.048	0.153	0.187	0.228	0.601	0.600
CLIP4CXR$_{\text{base}}$	M,C,P	0.094	0.389	0.084	0.347	0.499	0.239	0.012	0.083	0.153	0.166	0.559	0.481
LLM2CLIP4CXR$_{\text{base}}$	M,C,P	0.175	0.545	0.169	0.428	0.545	0.295	0.042	0.142	0.183	0.210	0.594	0.575
CLIP4CXR$_{\text{section}}$	M,C,P	0.091	0.385	0.081	0.350	0.497	0.232	0.010	0.085	0.159	0.168	0.545	0.479
LLM2CLIP4CXR$_{\text{section}}$	M,C,P	**0.182**	0.553	**0.179**	**0.430**	**0.547**	0.299	**0.052**	0.146	0.190	**0.231**	0.603	0.600
CLIP4CXR$_{\text{section}}$	M,C,P, Pr	0.072	0.352	0.058	0.322	0.483	0.227	0.006	0.052	0.144	0.152	0.523	0.452
LLM2CLIP4CXR$_{\text{section}}$	M,C,P, Pr	0.179	**0.567**	0.178	0.417	0.546	**0.308**	0.049	**0.155**	**0.196**	0.230	**0.610**	**0.618**

Dataset abbreviations: **M**=MIMIC, **C**=CheXpert-plus, **P**=PadChest, **U**=US-Mix, **B**=BimCV, **Pu**=other public datasets, **Pr**=private dataset.
Highlighting: bold indicates the best result among models trained only on MIMIC; **red bold** indicates the overall best result.
Metric: MF1 = CheXbert macro F1.

4) Summary of multimodal retrieval: Overall, *LLM2CLIP4CXR* scales robustly across diverse reporting styles: as stylistically varied sources are progressively added, its clinical retrieval performance is maintained or even improved, particularly on external Open-I evaluation (Table VI). BERT-based CLIP baselines instead degrade as report styles diverge, showing that a robust LLM-based text encoder, rather than data volume alone, is what enables clinically faithful and generalizable retrieval.

C. Qualitative evaluation of retrieved reports

As noted in Section IV, automated metrics may overlook clinically important aspects of report quality. To complement them, we performed a qualitative evaluation on 200 Open-I cases using three medical-student raters and four LLM judges. We included the generative model *MAIRA2* as a reference system and retrained four CLIP-based retrieval models (*CLIP4CXR*, *CXR-CLIP*, BC2C [44], and *LLM2CLIP4CXR*$_{\text{MC}}$) under identical MIMIC+CheXpert conditions for fair comparison. We also evaluated our large-scale model, *LLM2CLIP4CXR*$_{\text{all}}$, trained on 1.6M pairs.

Table VII shows that LLM2CLIP4CXR$_{\text{all}}$ achieves the best average rank among both student raters and LLM judges. Among models trained under the same MIMIC+CheXpert setting, *LLM2CLIP4CXR*$_{\text{MC}}$ achieves the best qualitative performance among CLIP-based retrieval models and ranks favorably relative to MAIRA2. These findings suggest that LLM-based encoders improve the clinical relevance of retrieved reports while preserving the controllability of retrieval-based output. The thoracic radiologist's rankings on the 72-case subset are summarized in Table VIII, alongside the corresponding student and LLM-judge rankings on the same cases.

TABLE VII
QUALITATIVE EVALUATION ON 200 OPEN-I CASES. STUDENT VALUES ARE AVERAGED ACROSS THREE MEDICAL-STUDENT RATERS, AND LLM VALUES ARE AVERAGED ACROSS FOUR LLM JUDGES.

Model	**Student Mean Rank ↓**	**LLM Mean Rank ↓**	**Overall Mean Rank ↓**
LLM2CLIP4CXR$_{\text{all}}$[a]	**1.85**	**2.46**	**2.19**
LLM2CLIP4CXR$_{\text{MC}}$	2.46	2.75	2.62
MAIRA2 [22]	2.50	3.55	3.08
CLIP4CXR	3.54	4.00	3.80
CXR-CLIP [13]	3.71	4.22	4.03
BC2C [44]	3.47	3.89	3.71

[a] *LLM2CLIP4CXR*$_{\text{all}}$ denotes the model trained on the full training corpus, including MIMIC-CXR, CheXpert-plus, PadChest, and SNUH reports. Lower rank indicates better qualitative performance. Ties were permitted.

1) Summary of qualitative evaluation: Across all rater types *LLM2CLIP4CXR*$_{\text{all}}$ was consistently judged to capture clinical semantics more faithfully than the baselines (Tables VII and VIII), with the expert ranking it first in 76% of cases. This agreement, holding even under specialist radiologist review, indicates that the LLM-based encoder encodes richer and more clinically accurate report semantics; notably, the retrieved reports were also judged comparable to or better than those from a report-generation baseline.

VI. DISCUSSION

This study demonstrates that incorporating LLM-based encoders into medical vision–language frameworks substantially improves clinical text understanding and cross-dataset generalization. By developing domain-adapted LLM encoders that handle the linguistic diversity of radiology reports, including abbreviated, templated, and impression-only styles, our approach enables more consistent multimodal learning across heterogeneous datasets.

Together, the text-only, multimodal, and dataset-composition experiments address the central question posed in

TABLE VIII
THORACIC-RADIOLOGIST-REFERENCE RANKING ON THE 72-CASE SUBSET. RANKS WERE COMPUTED WITHIN THE THREE REPRESENTATIVE SYSTEMS ONLY; LOWER MEAN RANK INDICATES BETTER CLINICAL MATCH TO THE GROUND-TRUTH REPORT.

Model	Expert mean rank ↓	Expert top-rank votes ↑	Student mean rank ↓	Student top-rank votes ↑	LLM mean rank ↓	LLM top-rank votes ↑
LLM2CLIP4CXR$_{\text{all}}$	**1.29**	**55/72 (76%)**	**1.14**	**193/216 (89%)**	**1.43**	**191/288 (66%)**
MAIRA2 [22]	2.21	11/72 (15%)	2.07	62/216 (29%)	2.39	48/288 (17%)
BC2C [44]	2.50	6/72 (8%)	2.35	35/216 (16%)	2.17	53/288 (18%)

the Introduction: whether stronger report representations can improve medical VLP under heterogeneous reporting styles. The results show that domain-adapted LLM encoders better capture CXR report semantics than BERT-based encoders and that this advantage transfers to CLIP-style retrieval, particularly when training data include impression-only or abbreviation-heavy reports.

The proposed *LLM2VEC4CXR* and *LLM2CLIP4CXR* frameworks further show that clinically robust medical VLP depends not only on dataset scale, but also on how report semantics are encoded. Combining LLM-based embeddings with masked token prediction and contrastive learning improves robustness to stylistic variation and supports clinically faithful retrieval, as reflected by gains on the error-discrimination and clinical-similarity evaluations. Importantly, our results indicate that large-scale integration of private, unstructured hospital reports is feasible when the text encoder is designed to handle report-style heterogeneity.

This study has several limitations. First, although we evaluated external generalization on Open-I and included heterogeneous public and private training data, the evaluation may not fully capture reporting variation across institutions, scanners, and clinical workflows. Second, expert validation was performed by a single thoracic radiologist, so inter-radiologist variability and consensus reliability could not be assessed. Third, our experiments focus on CXRs, and extending the framework to other imaging modalities will be necessary to assess broader generalizability. Finally, image-to-text retrieval serves as a proxy for multimodal understanding and does not directly evaluate downstream clinical decision-making or prospective workflow impact.

Clinically, LLM2CLIP4CXR could serve as a retrieval layer within PACS/RIS workflows by surfacing prior or similar CXR cases with semantically matched reports, supporting report drafting, discrepancy review, and image–report consistency checking. Because the system retrieves existing image–report pairs rather than generating unconstrained diagnostic text, its outputs remain traceable and can be reviewed by radiologists as decision support; however, prospective workflow integration, latency testing, and safety monitoring are required before clinical deployment.

Future work will address these limitations through multi-institutional, style-stratified, and temporal validation with multiple radiologists. The framework will be extended beyond CXRs to other imaging modalities and multi-view or longitudinal settings; in particular, the domain-adapted encoder (*LLM2VEC4CXR*) could be transferred to clinically adjacent thoracic report domains such as chest CT, with appropriate domain-specific adaptation, to probe its capability beyond CXR. To move beyond retrieval as a proxy task, we will evaluate downstream clinical applications enabled by retrieval-based representations, including report–image mismatch detection, similar-case retrieval, and clinician-in-the-loop decision support.

VII. CONCLUSION

We introduced *LLM2VEC4CXR*, a domain-adapted LLM encoder for chest X-ray reports, and its multimodal extension *LLM2CLIP4CXR* for image–text retrieval. Across diverse benchmarks, these models consistently outperformed BERT-based and prior medical CLIP variants, capturing clinical semantics more effectively and maintaining robustness across heterogeneous, abbreviation-rich, and impression-focused reporting styles. Trained on 1.6M CXRs from public and private sources, and supported by expert thoracic-radiologist validation, our models demonstrate that LLM-based encoders enable scalable, clinically aware, and generalizable multimodal learning, offering a foundation for more robust medical vision–language models in future research.

ETHICS APPROVAL

Use of public datasets (MIMIC-CXR, CheXpert-plus, PadChest, Open-I) complied with their data-use agreements and institutional policies; all data are fully de-identified. For the private hospital dataset, this retrospective study was approved by the Institutional Review Board of Seoul National University Hospital (IRB No.: 2405-070-1536), which granted a waiver of informed consent due to the use of de-identified data. No prospective data collection or intervention was performed.

CODE AVAILABILITY

We release model checkpoints and code for research use. The *LLM2VEC4CXR* model is available at `https://huggingface.co/lukeingawesome/llm2vec4cxr` and `https://github.com/lukeingawesome/llm2vec4cxr`, and training/inference code for *LLM2CLIP4CXR* is available at `https://github.com/lukeingawesome/llm2clip4cxr`. Please see the repositories for licenses and instructions.

TABLE IX
INSTRUCTION PROMPT TEMPLATES USED FOR LLM2CLIP4CXR TRAINING.

Task type	Prompt template
Semantic similarity	`Retrieve semantically similar sentences.`
Summarization	`Summarize the CXR report.`
Finding-level classification	`Determine the change or status of the {finding}.`

APPENDIX I
DATASETS

We summarize dataset usage and provide additional details of the preprocessing steps.

A. Use of Indication fields for abbreviation learning

To improve robustness to shorthand, we incorporate *Indication* fields from MIMIC, which frequently contain abbreviations and condensed expressions. These are paired with expanded versions created through our variation pipeline. This pairing enables the encoder to align abbreviated inputs with their clinically complete forms, strengthening its ability to handle style variation.

B. Prompt design for report variation generation

We used large language models to generate clinically faithful report variants from MIMIC-CXR and CheXpert data. For MIMIC-CXR, preprocessing was conducted with Gemini 2.0-Flash via Vertex AI, in accordance with the responsible-use guidelines published on PhysioNet[1]. The preprocessing was completed prior to September 24, 2025, before the release of updated guidelines. For CheXpert, we applied DeepSeek-R1-Distill-Qwen-14B locally to ensure controlled data handling and compliance with institutional data governance.

a) Splitting and omission: Sentence splitting and prior-omission were implemented as described by [44].

b) Anatomical partitioning: We used the prompts in RadExtract [47] that instructed the model to decompose each report into separate anatomical regions (e.g., pleura, lung zones, mediastinum). Recombination into structured variants was rule-based for each anatomy part.

c) Others: For error generation, creation of clinically similar reports, and evaluation we report the prompts in the github repository `https://github.com/lukeingawesome/llm2vec4cxr`

C. Prompt templates for LLM2CLIP4CXR training

Table IX summarizes the fixed instruction prompts used during LLM2CLIP4CXR training. The prompts were designed to cover complementary report-understanding objectives, including semantic similarity, summarization, and finding-level classification. The same templates were used throughout training and were not tuned on individual evaluation cases.

[1]`https://physionet.org/news/post/gpt-responsible-use`

APPENDIX II
ADDITIONAL EXPERIMENTS

A. Computational trade-off

TABLE X
COMPUTATIONAL COST OF LLM2VEC4CXR VARIANTS. PEAK GPU MEMORY AND STEP TIME ARE REPORTED UNDER THE SAME FOUR-A6000 TRAINING SETUP.

Configuration	Trainable parameters	Peak GPU memory	Step time
1B + LoRA ($r = 16$)	16M (1.3%)	∼10 GiB	∼6 s
1B full FT	1,236M (100%)	∼14 GiB	∼7 s
8B + LoRA ($r = 16$)	59M (0.7%)	∼31 GiB	∼44 s

Table X show that LoRA does not eliminate the computational burden of a large backbone. While 8B LoRA provides the strongest text-encoder capacity, the 1B fully fine-tuned model offers a more favorable accuracy–efficiency balance, requiring substantially lower memory and step time while retaining competitive downstream performance. We therefore use the 1B setting as the more practical configuration for scalable multimodal training.

B. Additional retrieval examples for private dataset

Table XI shows retrieval examples comparing *LLM2VEC4CXR* with the original *LLM2VEC* and BERT-based encoders, using private reports with frequent abbreviations as queries and MIMIC reports as candidates. BERT-based models and the general *LLM2VEC* often fail to interpret these abbreviations and to retrieve the correct clinical findings. In contrast, *LLM2VEC4CXR* accurately expands the abbreviations and retrieves clinically similar reports, demonstrating stronger robustness to abbreviated and noisy reporting styles.

C. Ablation

a) Ablation for dataset composition: We conducted ablation experiments to analyze the effect of dataset composition during the training of *LLM2VEC4CXR*. As shown in Table XII, each component contributed meaningfully to overall performance, indicating that exposure to diverse linguistic patterns enhances text encoder robustness. Among them, the inclusion of clinically similar reports produced the largest gains, reflecting their role in learning medical abbreviations and integrating stylistically varied report structures. In addition, we evaluated different text-pooling strategies (Table XIII) and found that latent attention pooling outperformed simple mean pooling, suggesting that selectively weighting token importance yields more informative clinical representations.

ACKNOWLEDGMENT

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT) (Grant No. RS-2024-00354666). This work was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-25442867, Development of a Generative AI-Supported System

TABLE XI
TOP-1 RETRIEVED REPORT FOR TASK 5. A REPORT FROM PRIVATE DATASET IS USED AS THE QUERY, AND THE MODEL RETRIEVES THE MOST SIMILAR REPORT FROM THE MIMIC POOL. BLUE = CORRECT FINDINGS; RED = INCORRECT/HALLUCINATED; ORANGE = UNCERTAIN/PARTIALLY WRONG STATEMENTS.

Model	**Text**
Ground Truth (Query)	A tiny calcified granuloma, RULF. No active lung disease.
Base BERT	In comparison with the study of _, the heart remains at the upper limits of normal in size and there is some tortuosity of the aorta. No vascular congestion. No evidence of acute focal pneumonia at this time.
ClinicalBERT	In comparison with the study of _, there is no evidence of acute cardiopulmonary disease at this time. No vascular congestion, pleural effusion, or acute pneumonia. No evidence of old granulomatous disease.
LLM2VEC(1B)	Mild cardiomegaly. Lung volumes are low. There is no focal consolidation. No pneumothorax.
LLM2VEC4CXR+	There is a small calcified granuloma in the right upper lung field. This is of no clinical significance. There is no lung consolidation, pleural effusion, or pneumothorax. Cardiomediastinal silhouette and hilar structures are normal.

TABLE XII
ABLATION OF DATASET COMPOSITION FOR TRAINING *LLM2VEC4CXR+*.

Excluded subset	**Task 1** (↑)		**Task 2** (↑)		**Task 3** (↑)	**Task 4** (↑)		**Task 5** (↑)		
	@1	@5	@1	@5	Acc	@1	@3	RadGraph	MF1	GREEN
Full (no exclusion)	0.918	0.984	0.203	0.290	0.826	0.607	0.854	0.373	0.715	0.686
– Summarization reports	0.912	0.978	0.115	0.182	0.812	0.611	0.854	0.353	0.698	0.668
– Prior-omitted reports	0.867	0.899	0.182	0.271	0.812	0.594	0.820	0.361	0.664	0.658
– Anatomical-partitioned reports	0.901	0.912	0.196	0.278	0.790	0.607	0.795	0.386	0.712	0.662
– Clinically similar reports	0.910	0.978	0.156	0.206	0.728	0.417	0.783	0.342	0.652	0.646

TABLE XIII
ABLATION OF POOLING STRATEGY FOR TRAINING *LLM2VEC4CXR+*.

Pooling method	**Task 1** (↑)		**Task 2** (↑)		**Task 3** (↑)	**Task 4** (↑)		**Task 5** (↑)		
	@1	@5	@1	@5	Acc	@1	@3	RadGraph	MF1	GREEN
Latent attention	0.918	0.984	0.203	0.290	0.826	0.607	0.854	0.373	0.715	0.686
Mean pooling	0.901	0.960	0.188	0.271	0.812	0.556	0.762	0.381	0.704	0.670

Software Framework for Optimal Execution of SDx Intelligent Services).

Generative AI was used in limited, disclosed ways: large language models generated synthetic training-report variants, served as judges in the qualitative evaluation (Methods and Experiments), and assisted with grammar refinement. All scientific content, analyses, and conclusions were produced and verified by the authors, who take full responsibility for the work.

REFERENCES

[1] W. Shi, L. Tong, Y. Zhu, and M. D. Wang, "Covid-19 automatic diagnosis with radiographic imaging: Explainable attention transfer deep neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2376–2387, 2021.

[2] M. Frid-Adar, R. Amer, O. Gozes, J. Nassar, and H. Greenspan, "Covid-19 in cxr: From detection and severity scoring to patient disease monitoring," *IEEE journal of biomedical and health informatics*, vol. 25, no. 6, pp. 1892–1903, 2021.

[3] S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J.-L. Suárez, J. Luengo, M. Valero-González *et al.*, "Covidgr dataset and covid-sdnet methodology for predicting covid-19 based on chest x-ray images," *IEEE journal of biomedical and health informatics*, vol. 24, no. 12, pp. 3595–3605, 2020.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, 2021, pp. 8748–8763.

[5] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, "Multi-modal understanding and generation for medical images and text via vision-language pre-training," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6070–6080, 2022.

[6] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar, "Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning," *Nat. Biomed. Eng.*, vol. 6, no. 12, pp. 1399–1406, 2022.

[7] Z. Zhao, Y. Liu, H. Wu, M. Wang, Y. Li, S. Wang, L. Teng, D. Liu, Z. Cui, Q. Wang, and D. Shen, "CLIP in medical imaging: A survey," *Med. Image Anal.*, vol. 102, p. 103551, 2025.

[8] S. Bannur, S. Hyland, Q. Liu, F. Perez-Garcia, M. Ilse, D. C. Castro, B. Boecking, H. Sharma, K. Bouzid, A. Thieme, A. Schwaighofer, M. Wetscherek, M. P. Lungren, A. Nori, J. Alvarez-Valle, and O. Oktay, "Learning to exploit temporal structure for biomedical vision–language processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 15 016–15 027.

[9] S. Agarwal, D. Wood, B. A. K. Murray, Y. Wei, A. A. Busaidi, S. Kafiabadi, E. Guilhem, J. Lynch, M. Townend, A. Mazumder, G. J. Barker, J. H. Cole, P. Sasieni, S. Ourselin, M. Modat, and T. C. Booth, "Impact of hospital-specific domain adaptation on BERT-based models to classify neuroradiology reports," *Eur. Radiol.*, vol. 35, no. 9, pp. 5299–5313, 2025.

[10] X. Liu, H. Liu, G. Yang, Z. Jiang, S. Cui, Z. Zhang, H. Wang, L. Tao, Y. Sun, Z. Song, T. Hong, J. Yang, T. Gao, J. Zhang, X. Li, J. Zhang,

Y. Sang, Z. Yang, K. Xue, S. Wu, P. Zhang, J. Yang, C. Song, and G. Wang, "A generalist medical language model for disease diagnosis assistance," *Nat. Med.*, vol. 31, no. 3, pp. 932–942, 2025.

[11] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proc. Clin. Nat. Lang. Process. Workshop (ClinicalNLP)*, 2019, pp. 72–78.

[12] K. Wang, N. Thakur, N. Reimers, and I. Gurevych, "GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. (NAACL-HLT)*, 2022, pp. 2345–2360.

[13] K. You, J. Gu, J. Ham, B. Park, J. Kim, E. K. Hong, W. Baek, and B. Roh, "CXR-CLIP: Toward large-scale chest X-ray language–image pre-training," in *Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, 2023, pp. 101–111.

[14] P. BehnamGhader, V. Adlakha, M. Mosbach, D. Bahdanau, N. Chapados, and S. Reddy, "LLM2VEC: Large language models are secretly powerful text encoders," in *Proc. Conf. Lang. Model. (COLM)*, 2024.

[15] W. Huang, A. Wu, Y. Yang, X. Luo, Y. Yang, U. Naseem, C. Wang, Q. Dai, X. Dai, D. Chen, C. Luo, L. Qiu, and L. Hu, "LLM2CLIP: Powerful language model unlocks richer cross-modality representation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 40, no. 7, 2026, pp. 5131–5139.

[16] C. Lee, R. Roy, M. Xu, J. Raiman, M. Shoeybi, B. Catanzaro, and W. Ping, "NV-embed: Improved techniques for training LLMs as generalist embedding models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2025.

[17] 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 *Mach. Learn. Healthc. Conf. (MLHC)*. PMLR, 2022, pp. 2–25.

[18] B. Boecking, N. Usuyama, S. Bannur, D. C. Castro, A. Schwaighofer, S. Hyland, M. Wetscherek, T. Naumann, A. Nori, J. Alvarez-Valle, H. Poon, and O. Oktay, "Making the most of text semantics to improve biomedical vision–language processing," in *Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–21.

[19] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2022, pp. 3876–3887.

[20] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang, and H. Poon, "A multimodal biomedical foundation model trained from fifteen million image–text pairs," *NEJM AI*, vol. 2, no. 1, p. AIoa2400640, 2025.

[21] W. Nimalsiri, M. Hennayake, K. Rathnayake, T. D. Ambegoda, and D. Meedeniya, "Automated radiology report generation using transformers," in *2023 3rd International Conference on Advanced Research in Computing (ICARC)*. IEEE, 2023, pp. 90–95.

[22] S. Bannur, K. Bouzid, D. C. Castro, A. Schwaighofer, S. Bond-Taylor, M. Ilse, F. Pérez-García, V. Salvatelli, H. Sharma, F. Meissen, M. Ranjit, S. Srivastav, J. Gong, N. C. Codella, F. Falck, O. Oktay, M. P. Lungren, M. T. Wetscherek, J. Alvarez-Valle, and S. L. Hyland, "MAIRA-2: Grounded radiology report generation," *arXiv preprint*, 2024, preprint.

[23] Z. Chen, M. Varma, M. Paschali, L. Blankemeier, D. Van Veen, J. M. J. Valanarasu, A. Youssef, E. P. Reis, J. P. Cohen, C. Olsen, T. M. Abraham, E. B. Tsai, C. F. Beaulieu, J. Jitsev, S. Gatidis, J.-B. Delbrouck, A. S. Chaudhari, and C. P. Langlotz, "CheXAgent: Towards a foundation model for chest X-ray interpretation," *arXiv preprint*, 2024, preprint.

[24] S. Lee, J. Youn, H. Kim, M. Kim, and S. H. Yoon, "CXR-LLAVA: A multimodal large language model for interpreting chest X-ray images," *Eur. Radiol.*, vol. 35, no. 7, pp. 4374–4386, 2025.

[25] D. Edirisinghe, W. Nimalsiri, M. Hennayake, D. Meedeniya, and G. Lim, "Chest X-ray report generation using abnormality guided vision language model," *IEEE Access*, vol. 13, pp. 157 651–157 673, 2025.

[26] X. Zhang, H.-Y. Zhou, X. Yang, O. Banerjee, J. N. Acosta, J. Miller, O. Huang, and P. Rajpurkar, "ReXrank: A public leaderboard for AI-powered radiology report generation," in *Proc. AAAI Bridge Program AI Med. Healthc.*, ser. PMLR, vol. 281, 2025, pp. 90–99.

[27] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2020, pp. 1500–1519.

[28] S. Jain, A. Agrawal, A. Saporta, S. Q. H. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng, C. P. Langlotz, and P. Rajpurkar, "RadGraph: Extracting clinical entities and relations from radiology reports," in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks (NeurIPS)*, vol. 1, 2021.

[29] S. Ostmeier, J. Xu, Z. Chen, M. Varma, L. Blankemeier, C. Bluethgen, A. E. Michalson, M. Moseley, C. Langlotz, A. S. Chaudhari, and J.-B. Delbrouck, "GREEN: Generative radiology report evaluation and error notation," in *Find. Assoc. Comput. Linguist. (EMNLP)*, 2024, pp. 374–390.

[30] B. Koloski, A. Margeloiu, X. Jiang, B. Škrlj, N. Simidjievski, and M. Jamnik, "LLM embeddings for deep learning on tabular data," *arXiv preprint*, 2025, preprint.

[31] C. Zhang, H. Zhang, S. Wu, D. Wu, T. Xu, X. Zhao, Y. Gao, Y. Hu, and E. Chen, "NoteLLM-2: Multimodal large representation models for recommendation," in *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Min. (KDD)*, 2025, pp. 2815–2826.

[32] R. Xu, W. Shi, Y. Yu, Y. Zhuang, Y. Zhu, M. D. Wang, J. C. Ho, C. Zhang, and C. Yang, "BMRetriever: Tuning large language models as better biomedical text retrievers," in *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, 2024, pp. 22 234–22 254.

[33] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[34] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[35] A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific data*, vol. 6, no. 1, p. 317, 2019.

[36] P. Chambon, J.-B. Delbrouck, T. Sounack, S.-C. Huang, Z. Chen, M. Varma, S. Q. H. Truong, C. T. Chuong, and C. P. Langlotz, "CheXpert plus: Hundreds of thousands of aligned radiology texts, images and patients," *arXiv preprint*, 2024, preprint.

[37] 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," *Med. Image Anal.*, vol. 66, p. 101797, 2020.

[38] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *J. Am. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, 2016.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. N. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. (NAACL-HLT)*, 2019, pp. 4171–4186.

[40] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, and A. Fan, "The Llama 3 herd of models," *arXiv preprint*, 2024, preprint.

[41] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[42] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[43] Z. Wang, X. Luo, X. Jiang, D. Li, and L. Qiu, "LLM-RadJudge: Achieving radiologist-level evaluation for X-ray report generation," *arXiv preprint*, 2024, preprint.

[44] H. Ko and C.-M. Park, "Bringing CLIP to the clinic: Dynamic soft labels and negation-aware learning for medical analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 25 897–25 906.

[45] V. M. Rao, S. Zhang, J. N. Acosta, S. Adithan, and P. Rajpurkar, "ReXErr: Synthesizing clinically meaningful errors in diagnostic radiology reports," in *Biocomputing 2025: Proc. Pacific Symp. Biocomput.*, 2024, pp. 70–81.

[46] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 3942–3951.

[47] Google Research, "Radextract: Radiology text extraction toolkit [software]," Hugging Face Spaces, 2024, accessed 28 Oct 2025. [Online]. Available: https://huggingface.co/spaces/google/radextract