FRAMING ABNORMALITIES IN CONTEXT: RECIPROCAL GROUNDING FOR RADIOGRAPH VISION–LANGUAGE PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical vision–language pre-training benefits from the rich supervision contained in radiology reports. However, existing methods often model focal clinical findings and their surrounding descriptions independently, without explicitly coordinating the complementary semantics they provide. We propose FRAME (**F**inding–context **R**eciprocal **A**lignment with **M**asked report reconstruction and **E**vidence-focused summary), a radiograph pre-training framework that couples finding-level and sentence-level semantics through shared visual evidence. Specifically, finding-conditioned visual representations are aligned with their corresponding sentence contexts, while context-conditioned visual representations are reciprocally aligned with the associated findings. This bidirectional constraint encourages the visual encoder to preserve both abnormality-specific evidence and its broader clinical context. Two complementary proxy tasks further strengthen the learned representation. Masked report reconstruction promotes the retention of dense diagnostic information, whereas evidence-focused summary emphasizes concise and clinically salient semantics. We evaluate FRAME across classification, segmentation, detection, visual question answering, and report generation under limited-label and cross-dataset settings. FRAME consistently outperforms strong Med-VLP baselines across these task families, especially with only 1% of the training labels. Ablation studies further demonstrate the advantage of reciprocal over unidirectional alignment and verify the complementary contributions of both proxy objectives. These results establish reciprocal finding–context grounding as an effective mechanism for learning transferable radiographic representations. Code will be available.

1 INTRODUCTION

Medical vision–language pre-training (Med-VLP) leverages paired radiographs and clinical reports to learn transferable visual representations. Unlike generic captions, radiology reports encode evidence at nested semantic levels. A focal clinical concept—such as an abnormality, disease, or anatomical term—points to relevant content, while its containing sentence qualifies it through anatomical location, attributes, and relations Li et al. (2025b). Effective radiograph pre-training should therefore preserve both concept-specific visual evidence and the clinical context in which it is described.

Existing Med-VLP methods exploit report supervision through complementary paradigms Wang et al. (2022b); Chen et al. (2022); Liu et al. (2023b). Contrastive approaches align radiographs and reports at global, local, or pathology-specific granularities Zhang et al. (2022); Huang et al. (2021); Wang et al. (2022a), while reconstructive methods recover corrupted multimodal content to retain detailed information Zhou et al. (2023); Wang et al. (2025). Autoregressive or hybrid frameworks further introduce language-prediction objectives Li et al. (2023); Wang et al. (2023b); Jiang et al. (2024); Kumar et al. (2026). However, aligning more granular units or adding proxy objectives does not specify how linked report semantics should jointly organize the visual representation. An isolated clinical concept is selective but provides limited contextual qualification, whereas its containing sentence is context-rich but may not uniquely isolate the associated visual evidence. The relationship between these concept-selective and context-rich views thus remains underexplored.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

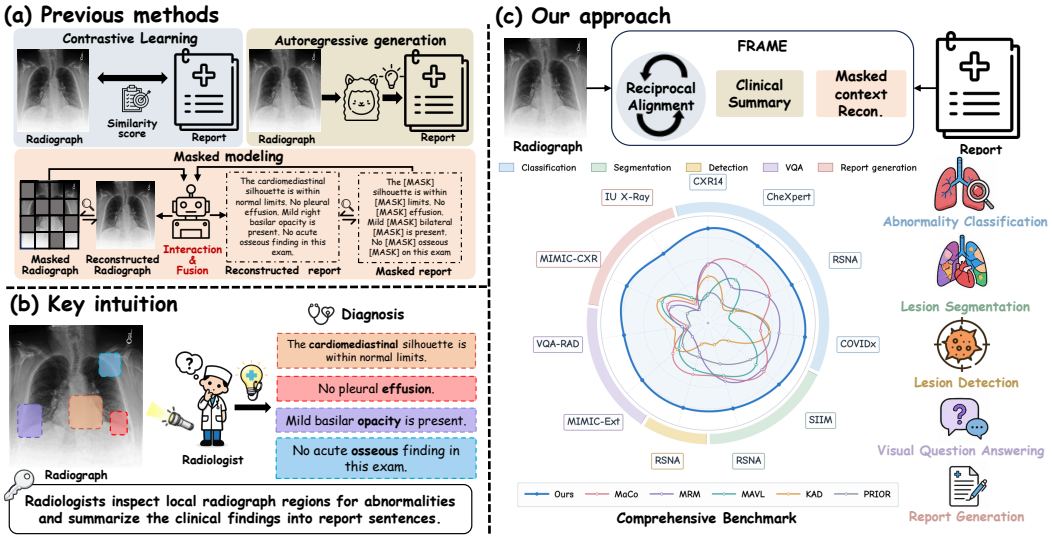

Figure 1: Motivation, design, and downstream evaluation of FRAME. FRAME couples local findings with sentence-level context through reciprocal alignment, masked report reconstruction, and evidence-focused summary; the resulting representation is evaluated across five downstream task families.

To address this gap, we propose FRAME (**F**inding–context **R**eciprocal **A**lignment with **M**asked report reconstruction and **E**vidence-focused summary), a radiograph vision–language pre-training framework centered on reciprocal finding–context alignment and supported by two complementary proxy tasks. Within FRAME, the finding view denotes a focal clinical concept extracted from the report, whereas the context view denotes its containing sentence. These views are coupled through two asymmetric image-mediated directions: the finding queries image patches and aligns the resulting visual response with its sentence context, while the context queries the image and aligns its response with the associated finding. Masked report reconstruction recovers corrupted report content using local and global radiographic evidence, whereas evidence-focused summary predicts a concise report-derived summary from the image. These proxy tasks provide detail-preserving and saliency-oriented supervision that complements the semantic organization established by reciprocal alignment.

We evaluate FRAME across diverse downstream tasks, including classification, segmentation, detection, visual question answering, and report generation, under both limited-label and cross-dataset settings. FRAME demonstrates strong and consistent transfer across global, spatial, question-conditioned, and language-conditioned tasks. Ablations show that the reciprocal formulation consistently outperforms either one-way alternative and that masked report reconstruction and evidence-focused summary each contribute across multiple task families. These results support both the reciprocal design and the complementary roles of the three FRAME components.

Our contributions are:

- We identify the insufficient coordination between focal clinical concepts and their sentence-level contexts as a key limitation of existing Med-VLP methods, and introduce finding–context reciprocal alignment to couple these complementary semantic views through two asymmetric grounding directions.
- We develop FRAME, a unified radiograph vision–language pre-training framework that combines reciprocal alignment with masked report reconstruction and evidence-focused summary. The three objectives provide context-organizing, detail-preserving, and saliency-oriented supervision over a shared visual encoder.
- We conduct extensive evaluations across classification, segmentation, detection, visual question answering, and report generation under limited-label and cross-dataset settings. FRAME achieves consistently strong transfer performance, while controlled ablations demonstrate the advantage of reciprocal over one-way alignment and the complementary effects of the two auxiliary objectives.

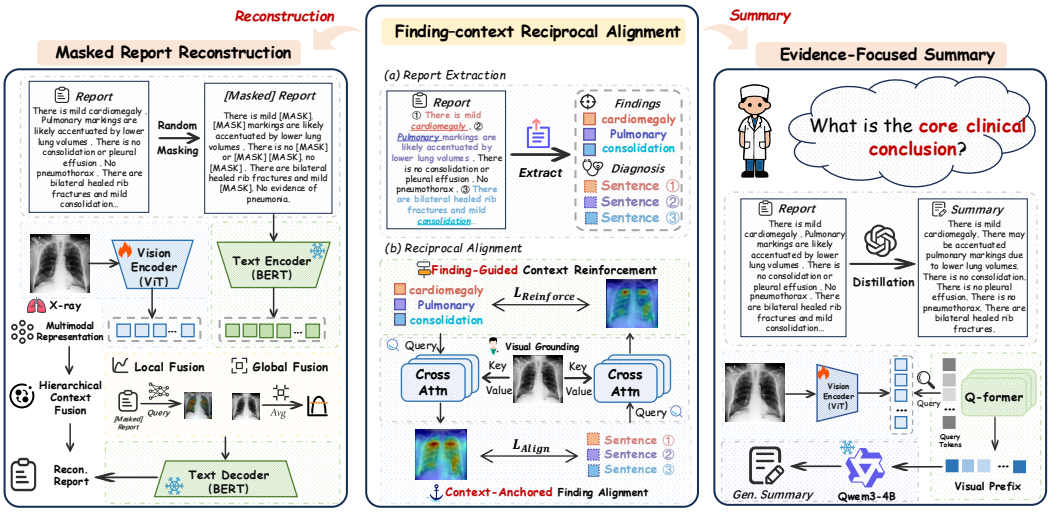

Figure 2: Overview of FRAME. The shared visual encoder is jointly optimized through finding–context reciprocal alignment (center), masked report reconstruction (left), and evidence-focused summary (right).

2 Related Work

Cross-Modal Alignment for Radiograph Pre-training. Radiograph–report pairs support transferable visual representation learning through cross-modal alignment. ConVIRT Zhang et al. (2022) performs global radiograph–report contrastive learning; GLoRIA Huang et al. (2021) introduces region–word correspondence; and MGCA Wang et al. (2022a) extends alignment to pathology-aware semantic granularities. Subsequent work refines alignment units and cross-sample relations. MLIP Li et al. (2024b) combines knowledge-guided objectives at global, token–patch, and category levels; MaCo Huang et al. (2024) applies correlation-weighted contrastive learning to masked radiographs; and SAHN Li et al. (2025a) uses clinical semantics to improve negative-sample construction. These advances improve the granularity and robustness of radiograph–report correspondence, yet multiscale alignment alone leaves constraints among linked levels of report semantics unspecified.

Clinically Structured Cross-Modal Integration. Medical knowledge and report structure make radiograph–text supervision more clinically specific. MedKLIP Wu et al. (2023) and KAD Zhang et al. (2023) guide visual representation learning with disease knowledge. ASG Li et al. (2024a) and ECAMP Wang et al. (2025) extract visual representation through report parsing and entity-centered modeling. GK-MVLP Deng et al. (2026) grounds knowledge prompts in anatomical regions, whereas CoGaze Liu et al. (2026) integrates clinical context and radiologist gaze into multi-level supervision. Our framework instead treats each finding mention and its containing sentence as linked but asymmetric report-internal semantic views. Reciprocal image grounding exchanges their conditioning and alignment roles.

Reconstructive and Autoregressive Proxy Objectives. Reconstruction and autoregressive prediction complement cross-modal alignment. M3AE Chen et al. (2022) reconstructs masked image and language inputs; MRM Zhou et al. (2023) jointly models masked radiographs and reports; and MedIM Xie et al. (2024) selects informative reconstruction regions using report-derived knowledge and sentence semantics. BLIP-2 Li et al. (2023) connects visual representations to frozen language models via a query-based adapter, which R2GenGPT Wang et al. (2023b) adapts for radiology report generation. HybridMED Jiang et al. (2024) combines multilevel alignment, captioning, and findings-to-impression summarization; MedUnifier Zhang et al. (2025) uses discrete visual representations for text and image generation; and CheXmix Kumar et al. (2026) couples early-fusion autoregressive pre-training with masked visual representation learning. We omit pixel reconstruction, pairing detail-preserving masked report reconstruction with saliency-oriented, image-conditioned prediction of a separately stored report-derived summary. Reciprocal grounding organizes both for visual understanding and downstream full-report generation.

3 METHOD

3.1 PRELIMINARIES

Each pre-training sample is a triplet (I_i, R_i, Y_i) comprising a radiograph, its original report $R_i = (r_{i,1}, \ldots, r_{i,L})$, and a separately stored report-derived evidence-focused summary $Y_i = (y_{i,1}, \ldots, y_{i,M_i})$. The report supplies the semantic pairs for reciprocal alignment and the targets for report reconstruction, whereas the summary is used only as a teacher-forced prediction target. A trainable visual encoder maps I_i to patch tokens $\mathbf{V}_i = E_v(I_i) \in \mathbb{R}^{P\times d}$, which are shared by all objectives. FRAME centers on Finding–context Reciprocal Alignment (FRA) and complements it with Masked Report Reconstruction (MRR) and Evidence-Focused Summary (EFS) as detail-preserving and evidence-compressing proxy tasks, respectively. Here, *finding* operationally denotes a focal clinical token matched by a predefined radiographic lexicon, and *context* denotes its containing sentence.

3.2 FINDING–CONTEXT RECIPROCAL ALIGNMENT

Each report finding represents a focal clinical concept associated with potentially relevant radiographic evidence, whereas its containing sentence supplies the attributes and relations that qualify that concept. Finding–context Reciprocal Alignment (FRA) connects these complementary report views through two reciprocal objectives, each grounded in visual features and defined on individual finding–sentence pairs.

A frozen CXR-BERT Boecking et al. (2022) encoder E_t represents report R_i as $\mathbf{T}_i = E_t(R_i) \in \mathbb{R}^{L\times d}$. We construct an abnormality vocabulary $\mathcal{V}_{\text{med}}$ from UMLS concepts Bodenreider (2004) and radiographic descriptors, such as *opacity*, *effusion*, and *pneumothorax*. Its curation is detailed in Appendix A.1. We classify report sentences as positive, possible, negative, or other and retain vocabulary occurrences only in positive or possible sentences. Negated mentions are excluded because they need not correspond to radiographic evidence of the named finding. Parsing details and the negation ablation appear in Appendices A.2 and C.1.

For study i, let $\mathbf{V}_i \in \mathbb{R}^{P\times d}$ denote its visual patch tokens (Sec. 3.1). Let $\mathbf{F}_i = [\mathbf{f}_{i,k}]_{k=1}^{K_i} \in \mathbb{R}^{K_i\times d}$ collect the contextualized features of retained finding occurrences, and let $\mathbf{C}_i = [\mathbf{c}_{i,j}]_{j=1}^{J_i} \in \mathbb{R}^{J_i\times d}$ collect the mean-pooled features of their distinct containing sentences. Repeated finding mentions remain separate occurrences. An incidence matrix $\mathbf{\Pi}_i \in \{0,1\}^{K_i\times J_i}$ records which sentence contains each occurrence: $[\mathbf{\Pi}_i]_{kj} = 1$ exactly when finding k occurs in sentence j. The relation set and the findings linked to sentence j are

$$\mathcal{E}_i = \{(k,j) : [\mathbf{\Pi}_i]_{kj} = 1\}, \qquad \mathcal{K}_{i,j} = \{k : (k,j) \in \mathcal{E}_i\}. \tag{1}$$

Each finding occurrence belongs to one sentence, whereas a sentence may contain multiple findings. Every edge in $\mathcal{E}_i$ supplies an exact positive relation for both alignment directions.

In both directions, textual queries attend over visual patch tokens:

$$\begin{aligned} \mathbf{u}^{\mathrm{F}}_{i,k} = \mathcal{G}_{\mathrm{F}}(\mathbf{f}_{i,k}, \mathbf{V}_i), \qquad \mathbf{u}^{\mathrm{C}}_{i,j} = \mathcal{G}_{\mathrm{C}}(\mathbf{c}_{i,j}, \mathbf{V}_i), \\ \mathcal{G}_b(\mathbf{q}, \mathbf{V}_i) = \mathrm{MHA}_b(g_t(\mathbf{q}), g_v(\mathbf{V}_i), g_v(\mathbf{V}_i)), \quad b \in \{\mathrm{F}, \mathrm{C}\}. \end{aligned} \tag{2}$$

Here $\mathrm{MHA}(Q, K, V)$ denotes multi-head attention: the projected textual features provide queries Q, and the projected visual patch tokens provide keys and values K, V. Each query attends over the patches to produce a query-conditioned visual representation. The projections g_t, g_v are shared across directions, whereas $\mathrm{MHA}_{\mathrm{F}}$ and $\mathrm{MHA}_{\mathrm{C}}$ have separate parameters.

Finding-Guided Context Reinforcement. In this direction, each finding independently attends over visual patch tokens to produce $\mathbf{u}^{\mathrm{F}}_{i,k}$. For $(k,j) \in \mathcal{E}_i$, its containing sentence $\mathbf{c}_{i,j}$ is the exact positive anchor. Thus FCR relates finding-conditioned visual evidence to the clinical context of that finding without merging the other relations in the study.

Context-Anchored Finding Alignment. The reciprocal direction uses each sentence context as a query over the visual patch tokens, producing $\mathbf{u}^{\mathrm{C}}_{i,j}$. For every $(k,j) \in \mathcal{E}_i$, the linked finding $\mathbf{f}_{i,k}$ is the exact positive anchor. A sentence with multiple retained findings therefore contributes a separate positive relation for each finding.

To contrast these local positive pairs with other studies, let $s(\mathbf{q}, \mathbf{a}) = \nu(\mathbf{q})^\top \nu(\mathbf{a})$, where $\nu(\mathbf{x}) = \mathbf{x}/\max(\|\mathbf{x}\|_2, \epsilon)$, and let τ be the contrastive temperature. We compare a local query with each candidate anchor in a nonmatching study, then aggregate the resulting similarities:

$$A_\tau(\mathbf{q}, \mathcal{X}) = \tau \log\left(\frac{1}{|\mathcal{X}|}\sum_{\mathbf{x}\in\mathcal{X}} \exp\left(\frac{s(\mathbf{q},\mathbf{x})}{\tau}\right)\right). \tag{3}$$

The cardinality-normalized log-mean-exp assigns more weight to confusable candidates without selecting a discrete hard negative. The factor $|\mathcal{X}|^{-1}$ prevents a larger candidate set from contributing a larger negative score merely because of its size. For a positive pair $(\mathbf{q}, \mathbf{a})$ from study $i \in \mathcal{B}_+$, let $\mathcal{X}_r$ be the candidate set from another valid study $r \neq i$. Here $\mathcal{B}_+$ contains studies with at least one retained relation, and the shorthand $r \neq i$ always ranges over $\mathcal{B}_+$. We define

$$p_\tau(\mathbf{q}, \mathbf{a}; \{\mathcal{X}_r\}_{r\neq i}) = \frac{\exp(s(\mathbf{q},\mathbf{a})/\tau)}{\exp(s(\mathbf{q},\mathbf{a})/\tau) + \sum_{r\in\mathcal{B}_+\setminus\{i\}} \exp(A_\tau(\mathbf{q}, \mathcal{X}_r)/\tau)}. \tag{4}$$

Each other study contributes one negative score, regardless of its number of anchors. Unlinked finding–sentence pairs from the same study are excluded from both positive and negative terms. Changing the within-study finding–sentence assignment therefore generally changes the positive terms, even when the finding and sentence sets are unchanged.

Let $\mathcal{C}_r = \{\mathbf{c}_{r,j}\}_{j=1}^{J_r}$ and $\mathcal{F}_r = \{\mathbf{f}_{r,k}\}_{k=1}^{K_r}$ denote the sentence and finding anchors of study r. Their visual counterparts are $\mathcal{U}_r^{\mathrm{F}} = \{\mathbf{u}_{r,k}^{\mathrm{F}}\}_{k=1}^{K_r}$ and $\mathcal{U}_r^{\mathrm{C}} = \{\mathbf{u}_{r,j}^{\mathrm{C}}\}_{j=1}^{J_r}$. The two directions retain symmetric InfoNCE at relation level:

$$\begin{aligned}
\ell_{i,k,j}^{\mathrm{F}} &= -\tfrac{1}{2}\big[\log p_\tau(\mathbf{u}_{i,k}^{\mathrm{F}}, \mathbf{c}_{i,j}; \{\mathcal{C}_r\}_{r\neq i}) + \log p_\tau(\mathbf{c}_{i,j}, \mathbf{u}_{i,k}^{\mathrm{F}}; \{\mathcal{U}_r^{\mathrm{F}}\}_{r\neq i})\big], \\
\ell_{i,k,j}^{\mathrm{C}} &= -\tfrac{1}{2}\big[\log p_\tau(\mathbf{u}_{i,j}^{\mathrm{C}}, \mathbf{f}_{i,k}; \{\mathcal{F}_r\}_{r\neq i}) + \log p_\tau(\mathbf{f}_{i,k}, \mathbf{u}_{i,j}^{\mathrm{C}}; \{\mathcal{U}_r^{\mathrm{C}}\}_{r\neq i})\big].
\end{aligned} \tag{5}$$

The forward terms compare each query-conditioned visual representation with textual anchors from other studies; the reverse terms compare each text anchor with visual representations from those studies. In both directions, similarities are computed before any relation-level loss is averaged.

We average local losses within each study so that reports with more retained relations do not receive greater weight. FCR averages over relation edges; CFA first averages over findings within each sentence and then over sentences:

$$\begin{aligned}
\mathcal{L}_{\mathrm{FCR}} &= \frac{1}{|\mathcal{B}_+|}\sum_{i\in\mathcal{B}_+}\frac{1}{|\mathcal{E}_i|}\sum_{(k,j)\in\mathcal{E}_i}\ell_{i,k,j}^{\mathrm{F}}, \\
\mathcal{L}_{\mathrm{CFA}} &= \frac{1}{|\mathcal{B}_+|}\sum_{i\in\mathcal{B}_+}\frac{1}{J_i}\sum_{j=1}^{J_i}\frac{1}{|\mathcal{K}_{i,j}|}\sum_{k\in\mathcal{K}_{i,j}}\ell_{i,k,j}^{\mathrm{C}}, \\
\mathcal{L}_{\mathrm{FRA}} &= \lambda_{\mathrm{FCR}}\mathcal{L}_{\mathrm{FCR}} + \lambda_{\mathrm{CFA}}\mathcal{L}_{\mathrm{CFA}}.
\end{aligned} \tag{6}$$

The outer averages give equal weight to each valid study, while the inner CFA average gives equal weight to each retained sentence. FRA thus exchanges the query and semantic-anchor roles of the two report views while retaining their finding–sentence correspondence in both directions.

3.3 Masked Report Reconstruction

We employ Masked Report Reconstruction (MRR) as an auxiliary task to preserve report-related visual details beyond the finding–guided and context–anchored visual representations learned by FRA. Given a randomly masked report $\widetilde{R}_i$, the frozen CXR-BERT Boecking et al. (2022) produces token features that attend to the visual patch tokens, yielding token-conditioned local visual evidence. We additionally mean-pool the patch tokens to obtain a global visual feature, which is broadcast across report positions. The local and global features are fused with the textual features and passed to the reconstruction Transformer. We denote these two visual conditioning sources as local (L) and global (G) in the ablation study. A vocabulary head predicts the original token at every report position, giving

$$\mathcal{L}_{\mathrm{MRR}} = -\frac{1}{|\mathcal{B}|L}\sum_{i\in\mathcal{B}}\sum_{n=1}^{L}\log p_\theta(r_{i,n} \mid I_i, \widetilde{R}_i, n). \tag{7}$$

Table 1: Linear-probe classification at various labeled-data fractions. † denotes locally reproduced results from officially released weights; ‡ denotes paper-reported results on matching splits. **Boldface** and underlining mark the best and second-best results; '–' denotes unavailable results. * marks improvement over the strongest locally evaluated baseline with reported variability in each column (two-sided Welch's t-test, $p < 0.05$).

Method	CXR14 (AUC)			CheXpert (AUC)			RSNA (AUC)			COVIDx (ACC)		
	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
MGCA † (2022a)	$75.1_{0.1}$	$79.5_{0.1}$	$81.7_{0.0}$	$84.0_{0.2}$	$88.0_{0.1}$	$88.1_{0.1}$	$90.8_{0.1}$	$91.6_{0.1}$	$91.9_{0.1}$	$71.8_{1.1}$	$77.3_{0.2}$	$81.0_{0.2}$
MRM (2023)	78.8	82.7	84.6	88.1	88.1	88.3	90.9	92.5	92.5	79.8	89.0	92.8
MLIP (2024b)	–	–	–	89.0	89.4	90.0	89.3	90.0	90.8	75.3	86.3	92.5
ASG‡ (2024a)	79.5	82.2	83.6	87.9	89.0	89.0	88.4	89.5	90.2	81.3	87.0	93.3
MaCo† (2024)	$78.7_{0.1}$	$82.5_{0.1}$	$83.6_{0.0}$	$89.5_{0.2}$	$\underline{90.6}_{0.1}$	$\underline{91.0}_{0.1}$	$91.3_{0.1}$	$92.2_{0.1}$	$93.1_{0.1}$	$72.0_{1.2}$	$75.5_{0.2}$	$78.8_{0.2}$
ECAMP (2025)	<u>81.0</u>	<u>83.8</u>	<u>85.1</u>	89.3	89.5	89.7	<u>91.4</u>	<u>92.9</u>	<u>93.5</u>	<u>86.3</u>	<u>95.0</u>	**96.0**
RadZero† (2025)	$77.5_{0.1}$	$82.2_{0.1}$	$84.0_{0.1}$	$87.4_{0.1}$	$89.5_{0.1}$	$90.1_{0.1}$	$90.2_{0.1}$	$91.7_{0.1}$	$92.8_{0.1}$	$85.5_{0.8}$	$\underline{95.0}_{0.2}$	$\mathbf{96.0_{0.1}}$
FaNe‡ (2026)	–	–	–	<u>89.7</u>	90.4	90.8	89.3	90.2	93.1	79.5	90.7	95.5
Ours	$\mathbf{81.2_{0.1}}^{*}$	$\mathbf{85.0_{0.1}}^{*}$	$\mathbf{85.7_{0.0}}^{*}$	$\mathbf{90.5_{0.2}}^{*}$	$\mathbf{91.2_{0.1}}^{*}$	$\mathbf{91.5_{0.1}}^{*}$	$\mathbf{92.2_{0.2}}^{*}$	$\mathbf{93.1_{0.1}}^{*}$	$\mathbf{93.9_{0.1}}^{*}$	$\mathbf{87.8_{1.2}}^{*}$	$\mathbf{95.5_{0.3}}^{*}$	$\underline{95.8}_{0.2}$

MRR therefore supplies cross–granularity visual supervision as a detail-preserving proxy task, complementing FRA without introducing another cross-modal alignment objective.

3.4 Evidence-Focused Summary

Evidence-Focused Summary (EFS) serves as an auxiliary generative task that encourages the visual encoder to preserve clinically salient evidence. Unlike downstream report generation, EFS does not aim to reproduce the complete radiology report. Instead, it predicts the separately prepared report-derived summary Y_i, which condenses the principal findings and their contextual qualifiers.

A trainable Q-Former queries the image features $\mathbf{V}_i$ to obtain a compact visual representation, which conditions a frozen language decoder to generate Y_i autoregressively. Under teacher forcing, the objective is

$$\mathcal{L}_{\text{EFS}} = -\frac{1}{\sum_{i\in\mathcal{B}} M_i} \sum_{i\in\mathcal{B}} \sum_{t=1}^{M_i} \log p(y_{i,t} \mid y_{i,<t}, F_q(\mathbf{V}_i)) \,. \tag{8}$$

The original report is not provided as input to this pathway. EFS therefore uses summary generation as a proxy task to emphasize concise and clinically relevant visual semantics.

Overall Objective. The three components are jointly optimized as

$$\mathcal{L}_{\text{FRAME}} = \mathcal{L}_{\text{FRA}} + \lambda_{\text{MRR}}\mathcal{L}_{\text{MRR}} + \lambda_{\text{EFS}}\mathcal{L}_{\text{EFS}}. \tag{9}$$

We fix $\lambda_{\text{FCR}} = \lambda_{\text{CFA}} = \lambda_{\text{MRR}} = \lambda_{\text{EFS}} = 1$ without downstream-task-specific weight tuning. The within-study reductions in Eq. equation 6 keep the reciprocal losses from growing merely with the number of extracted relations. FRA specifies the reciprocal grounding relation, whereas MRR and EFS supply complementary detail-preserving and evidence-compressing supervision through the shared visual encoder.

4 Experiments and Results

4.1 Experimental Setup

Pre-training and Evaluation Data. We pre-train FRAME on MIMIC-CXR v2.0.0 Johnson et al. (2019a;b). Downstream evaluation covers classification on CXR14 Wang et al. (2017), CheXpert Irvin et al. (2019), RSNA Pneumonia Shih et al. (2019), and COVIDx Wang et al. (2020); segmentation on SIIM Langer & Shih (2019) and RSNA; RSNA detection; medical VQA on MIMIC-Ext Bae et al. (2024) and VQA-RAD Lau et al. (2018); and report generation on MIMIC-CXR and IU X-Ray Demner-Fushman et al. (2016). CheXpert and VinDr-CXR Nguyen et al. (2022) are the cross-dataset validation targets. Dataset statistics, splits, and preprocessing are provided in Appendix B.1.

Table 2: Disease segmentation and detection at various labeled-data fractions. † denotes locally reproduced results from officially released weights. **Boldface** and underlining mark the best and second-best results, respectively; '–' denotes unavailable results. * marks improvement over the strongest locally evaluated baseline with reported variability in each column (two-sided Welch's t-test, $p < 0.05$).

Method	SIIM (Dice)			RSNA (Dice)			RSNA (mAP)		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
MRM (2023)	63.1	68.3	78.4	74.2	77.6	78.2	11.5	20.3	27.1
Med-UniC (ViT) (2023)	62.1	67.3	71.5	75.6	76.6	77.9	–	–	–
MedIM (2024)	–	64.2	82.0	–	–	–	–	–	–
MaCo† (2024)	$64.7_{0.7}$	$68.4_{0.4}$	$81.5_{0.1}$	$73.3_{0.5}$	$76.2_{0.5}$	$78.1_{0.1}$	$18.8_{0.7}$	$26.0_{0.4}$	$29.4_{0.2}$
ECAMP (2025)	67.3	70.8	84.5	77.1	78.6	79.5	19.8	28.0	31.6
RadZero† (2025)	$63.9_{0.7}$	$67.8_{0.5}$	$81.2_{0.3}$	$71.6_{0.6}$	$72.9_{0.3}$	$70.5_{0.1}$	$12.4_{0.7}$	$25.9_{0.4}$	$32.1_{0.1}$
Ours	$\mathbf{69.2_{0.9}}^*$	$\mathbf{73.8_{0.4}}^*$	$\mathbf{86.4_{0.2}}^*$	$\mathbf{80.0_{0.6}}^*$	$\mathbf{80.9_{0.4}}^*$	$\mathbf{81.6_{0.1}}^*$	$\mathbf{26.1_{0.8}}^*$	$\mathbf{28.5_{0.4}}^*$	$\mathbf{34.0_{0.2}}^*$

Implementation Details. We use ViT-B/16 Dosovitskiy et al. (2021) as the visual encoder and frozen CXR-BERT Boecking et al. (2022) as the text encoder. For EFS, a trainable Q-Former and projector condition frozen Qwen3-4B Qwen Team (2025) on visual features; downstream VQA and report generation use separate task-specific adapters with the same frozen decoder. GPT-4o OpenAI (2024) distills each report into a concise clinical summary used solely as the EFS target. Summary construction and pre-training settings are detailed in Appendices A.3 and B.2, respectively. For the repeated evaluations in Tables 1 and 2, we use random seeds 42, 123, 3407, 5678, and 9223. Entries with reported variability show the mean, with the standard deviation as a subscript.

4.2 Downstream Tasks

We evaluate FRAME across five task families spanning global recognition, spatial perception, question-conditioned reasoning, and clinical text generation. Complete protocols and extended results are provided in Appendices B.3, C.2–C.4.

Classification. Under linear probing, FRAME consistently achieves the best performance on CXR14, CheXpert, and RSNA across all evaluated label fractions (Table 1). With only 1% labeled data, it obtains AUCs of 81.2%, 90.5%, and 92.2% on the three datasets, respectively, together with an accuracy of 87.8% on COVIDx. This advantage extends to external validation: when transferred from CXR14 to CheXpert and VinDr-CXR without target-domain training, FRAME achieves the highest mean AUROC on both datasets (Figure 3). In particular, it reaches 82.8% mean AUROC on VinDr-CXR, outperforming the next-best method by 6.8 percentage points. These consistent in-domain and cross-dataset results support the robustness and transferability of the learned radiographic representations. Compared with the strongest locally evaluated baseline with reported variability, FRAME shows statistically significant gains in 11 of the 12 classification settings ($p < 0.05$).

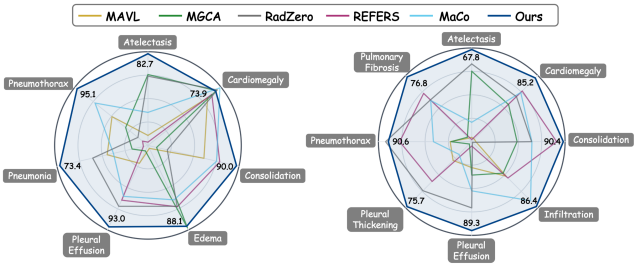

Figure 3: Cross-dataset classification from CXR14 (1% labels) to CheXpert (left) and VinDr-CXR (right). Per-label AUROC is plotted on label-specific axes.

Segmentation and Detection. FRAME ranks first across all settings (Table 2). With 1% labeled data, it attains 69.2% and 80.0% Dice on SIIM and RSNA segmentation, respectively, and 26.1% mAP on RSNA detection, supporting label-efficient spatial transfer. The gains

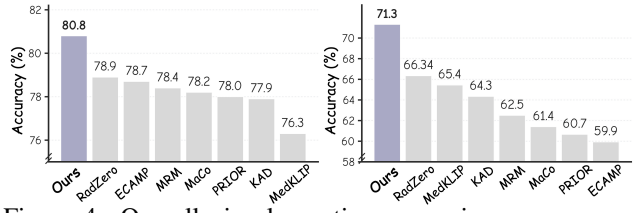

Figure 4: Overall visual question answering accuracy on MIMIC-Ext (left) and VQA-RAD (right).

Table 3: Natural-language generation (NLG) and clinical-efficacy metrics on report generation tasks (all values in %). BL-1/BL-4, MTR, and RG-L denote BLEU-1/BLEU-4, METEOR, and ROUGE-L; F1, RG-P, and GRN denote 14-label CheXbert Micro-F1, RadGraph Partial-F1, and GREEN. **Boldface** and underlining mark the best and second-best image–text pre-training results; gray cells mark column maxima across all reported methods. Direct VLMs and task-specific methods (†) are shown for reference and excluded from the primary ranking; '–' denotes unreported results.

Method	MIMIC-CXR							IU X-Ray						
	NLG				*Clinical Efficacy*			*NLG*				*Clinical Efficacy*		
	BL-1	BL-4	MTR	RG-L	F1	RG-P	GRN	BL-1	BL-4	MTR	RG-L	F1	RG-P	GRN
VLMs														
Qwen3-VL-4B (2025)	28.8	2.4	13.3	19.4	33.4	15.7	19.6	15.6	2.1	15.3	18.7	24.5	18.3	69.8
Lingshu-7B (2025)	31.5	8.5	13.6	23.5	42.4	22.3	26.7	41.5	12.4	17.3	32.5	57.8	33.4	57.3
Task-specific methods														
R2Gen† (2020)	35.3	10.3	14.2	27.7	–	–	–	47.0	16.5	18.7	37.1	–	–	–
METransformer† (2023a)	38.6	12.4	15.2	29.1	–	–	–	48.3	17.2	19.2	38.0	–	–	–
R2GenGPT† (2023b)	41.1	13.4	16.0	29.7	–	–	–	48.8	17.3	21.1	37.7	–	–	–
REFERS (2022)	24.1	8.2	12.1	28.7	33.6	23.2	28.1	44.8	15.4	18.7	37.8	51.5	38.9	62.4
MGCA (2022a)	23.8	8.0	11.6	28.4	30.9	20.1	26.8	44.3	17.0	18.9	38.1	54.3	37.9	65.6
MRM (2023)	20.6	7.6	11.7	28.6	31.5	23.1	28.7	38.8	14.3	17.9	38.1	54.0	42.1	60.0
MaCo (2024)	24.7	8.7	12.1	28.8	34.5	22.5	27.0	42.6	15.2	18.6	36.1	55.1	39.0	60.1
ECAMP (2025)	27.3	9.4	13.1	29.5	38.7	21.3	29.3	47.7	17.0	19.7	36.3	56.2	32.1	66.4
RadZero (2025)	25.9	9.3	12.7	29.1	33.3	23.2	28.8	44.9	16.8	19.3	38.6	55.9	41.8	63.7
Ours	**35.5**	**11.5**	**14.5**	**29.7**	**42.2**	**24.8**	**30.4**	**49.8**	**18.3**	**21.2**	**39.8**	**58.3**	**43.3**	**68.0**

over the strongest locally evaluated baseline with reported variability are statistically significant in all nine settings ($p < 0.05$).

Medical Visual Question Answering. Under a matched adaptation protocol that freezes the pre-trained visual encoder and Qwen3-4B Qwen Team (2025) while training only the Q-Former and projector, FRAME achieves 80.8% overall accuracy on MIMIC-Ext, outperforming the runner-up by 1.9 percentage points. It obtains 71.3% closed-ended accuracy on VQA-RAD, outperforming the next-best method by 5.0 percentage points (Figure 4). These results show consistent transfer to question-conditioned clinical prediction across both benchmarks.

Radiology Report Generation. The visual encoder and LLM remain frozen; only a task-specific Q-Former and projector are trained. FRAME leads Med-VLP methods on all seven metrics for MIMIC-CXR and IU X-Ray (Table 3); specialized generators and zero-shot VLMs are shown for reference. Compared with the strongest Med-VLP baselines, CheXbert Micro-F1 improves by 3.5 points on MIMIC-CXR (42.2% vs. 38.7%) and 2.1 points on IU X-Ray (58.3% vs. 56.2%).

4.3 Ablation Study

Table 4 evaluates the two FRA paths, the global and local MRR branches, and EFS across five downstream task families. All configurations are trained and evaluated under the same settings as the main experiments, enabling direct component-wise comparisons.

Effectiveness of Finding–context Reciprocal Alignment. Among configurations with MRR and EFS, 1% RSNA detection mAP is 9.8% without FRA, 11.1% with CFA, 19.5% with FCR, and 26.1% with both paths. The joint configuration also exceeds both one-way variants in every other column, consistent with complementary finding selectivity and contextual coherence.

Contribution of Masked Report Reconstruction. Among configurations with FRA and EFS, the local-only variant records a 1% RSNA detection mAP of 20.7% and an IU X-Ray Micro-F1 of 56.9%, compared with 15.2% and 55.2% without MRR. The global-only variant exceeds the local-only variant on MIMIC-Ext accuracy (79.0% vs. 77.6%) and IU X-Ray ROUGE-L (37.8% vs. 36.2%). The joint configuration ranks first in every column, consistent with complementary fine-grained patch evidence and study-level context.

Contribution of Evidence-Focused Summary. Compared with the full model, the no-EFS configuration is lower on every reported metric, including 1% CheXpert AUC (88.2% vs. 90.5%), 1% RSNA Dice (74.5% vs. 80.0%), and IU X-Ray Micro-F1 (54.2% vs. 58.3%). These cross-task differ-

Table 4: Component ablation across five downstream task families (all values in %). FCR/CFA, G/L, and EFS denote reciprocal FRA directions, global/local MRR conditioning, and evidence-focused summary supervision, respectively. MIMIC, BL-4/RG-L/MTR, and F1 denote MIMIC-Ext overall accuracy, BLEU-4/ROUGE-L/METEOR, and IU X-Ray 14-label CheXbert Micro-F1, respectively. Boldface marks column maxima. * on the final row indicates improvement over the best-performing ablated configuration in each column, assessed by a paired Wilcoxon signed-rank test ($p < 0.05$).

FRA		MRR		EFS	CheXpert (AUC)			RSNA (Dice)			RSNA (mAP)			MIMIC	IU X-Ray			
FCR	CFA	G	L		1%	10%	100%	1%	10%	100%	1%	10%	100%	Acc.	BL-4	RG-L	MTR	F1
		✓	✓	✓	88.4	89.8	90.5	73.4	74.9	71.7	9.8	23.4	31.9	77.5	15.7	35.6	19.0	54.6
	✓	✓	✓	✓	89.9	90.9	90.7	74.5	76.5	72.3	11.1	25.0	31.8	78.8	17.6	36.4	19.9	56.8
✓		✓	✓	✓	90.1	90.7	91.0	74.7	76.4	72.4	19.5	26.2	30.3	78.9	16.9	37.5	20.2	57.1
✓	✓			✓	89.1	90.2	90.3	73.2	76.3	70.8	15.2	22.8	27.1	77.2	13.0	35.1	16.9	55.2
✓	✓		✓	✓	89.4	90.3	90.4	74.0	76.8	72.4	20.7	25.4	26.8	77.6	15.9	36.2	18.8	56.9
✓	✓	✓		✓	89.7	90.5	90.7	74.3	77.1	71.9	17.6	23.4	27.7	79.0	15.0	37.8	18.1	57.4
✓	✓	✓	✓		88.2	89.8	90.0	74.5	77.1	72.0	25.3	27.8	28.2	78.9	13.8	37.1	17.9	54.2
✓	✓	✓	✓	✓	**90.5***	**91.2***	**91.5***	**80.0***	**80.9***	**81.6***	**26.1***	**28.5***	**34.0***	**80.8***	**18.3***	**39.8***	**21.2***	**58.3***

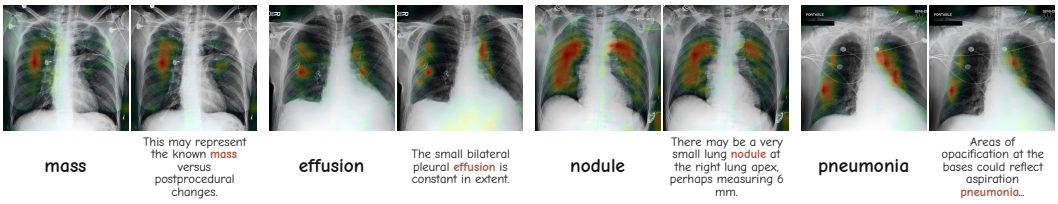

Figure 5: Finding–context reciprocal alignment. Each pair contrasts finding-guided context reinforcement (FCR; left) and context-anchored finding alignment (CFA; right) attention, with the corresponding queries below. Only the linked finding is highlighted; ellipses indicate omitted trailing context.

ences suggest that EFS supports both clinical understanding and generation. Together, FRA, MRR, and EFS provide complementary supervision.

4.4 Qualitative Analysis

Figure 5 contrasts finding-guided context reinforcement (FCR) and context-anchored finding alignment (CFA) attention for each finding–context pair. The paired maps qualitatively illustrate differences in the spatial attention patterns produced by the two reciprocal paths. Additional qualitative analyses are provided in Appendix D.

5 Conclusion

We introduced FRAME, a radiograph vision–language pre-training framework centered on Finding–context Reciprocal Alignment (FRA). Its two asymmetric paths align finding-guided visual evidence with sentence context and context-guided evidence with linked findings. Masked Report Reconstruction (MRR) and Evidence-Focused Summary (EFS) complement FRA with detail-preserving and saliency-oriented supervision. Results across five downstream task families support global, spatial, question-conditioned, and generative transfer. Descriptive ablation trends are consistent with complementary roles for FRA, MRR, and EFS. Matched-budget comparisons, visual-dependence controls, and broader localization analyses could further clarify the sources of the observed gains. One limitation is that FRAME does not exploit longitudinal information across examinations, preventing explicit modeling of temporal changes in clinical findings. Future work could incorporate prior radiographs and reports to model such changes.

AI USE STATEMENT

GPT-4o is used to generate concise, report-derived summaries from the original radiology reports for the Evidence-Focused Summary (EFS) pre-training objective (Appendix A.3). These summaries serve as training targets and are not used as test annotations. Generative AI tools also assist with the preliminary survey and organization of related literature during method development, and with editing the manuscript for clarity, grammar, and LaTeX presentation. The authors review AI-assisted research and writing suggestions and take full responsibility for the final method, experiments, citations, results, claims, and text.

REFERENCES

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and Edward Choi. MIMIC-Ext-MIMIC-CXR-VQA: A complex, diverse, and large-scale visual question answering dataset for chest x-ray images. PhysioNet, 2024. Version 1.0.0.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Teodora Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1439–1449. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.112. URL https://aclanthology.org/2020.emnlp-main.112/.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 5904–5914. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.459.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *MICCAI*, pp. 679–689. Springer, 2022.

Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. Prior: Prototype representation joint learning from medical images and reports. In *Int. Conf. Comput. Vis.*, pp. 21361–21371, 2023.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Qiao Deng, Zhongzhen Huang, Yunqi Wang, Zhichuan Wang, Zhao Wang, Xiaofan Zhang, Qi Dou, Yeung Yu Hui, and Edward S. Hui. Grounded knowledge-enhanced medical vision-language pre-training for chest x-ray. *Biomedical Signal Processing and Control*, 112:108416, 2026. doi: 10.1016/j.bspc.2025.108416.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Int. Conf. Comput. Vis.*, pp. 3942–3951, 2021.

Weijian Huang, Cheng Li, Hong-Yu Zhou, Hao Yang, Jiarun Liu, Yong Liang, Hairong Zheng, Shaoting Zhang, and Shanshan Wang. Enhancing representation in radiography-reports foundation model: A granular alignment algorithm using masked contrastive learning. *Nature Communications*, 15(1):7620, 2024.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Hanqi Jiang, Xixuan Hao, Yuzhou Huang, Chong Ma, Jiaxun Zhang, Yi Pan, and Ruimao Zhang. Advancing medical radiograph representation learning: A hybrid pre-training paradigm with multilevel semantic granularity. *arXiv preprint arXiv:2410.00448*, 2024.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019a.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019b.

Ashwin Kumar, Robbie Holland, Corey Barrett, Jangwon Kim, Maya Varma, Zhihong Chen, Yunhe Gao, Greg Zaharchuk, Tara Taghavi, Krishnaram Kenthapadi, and Akshay Chaudhari. CheXmix: Unified generative pretraining for vision language models in medical imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 9466–9476, 2026.

Steve G. Langer and George Shih. SIIM-ACR pneumothorax segmentation. Kaggle competition, 2019. URL `https://www.kaggle.com/competitions/siim-acr-pneumothorax-segmentation`.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. doi: 10.1038/sdata.2018.251.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.

Qingqiu Li, Xiaohan Yan, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, and Shujun Wang. Anatomical structure-guided medical vision-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 80–90. Springer, 2024a.

Yongxin Li, Ying Cheng, Yaning Pan, Wen He, Qing Wang, Rui Feng, and Xiaobo Zhang. Semantic-aware hard negative mining for medical vision-language contrastive pretraining. In *The 33rd ACM International Conference on Multimedia*, pp. 3133–3142, 2025a.

Zhe Li, Laurence T Yang, Bocheng Ren, et al. Mlip: Enhancing medical visual representation with divergence encoder and knowledge-guided contrastive learning. In *CVPR*, pp. 11704–11714, 2024b.

Zihan Li et al. Visionunite: A vision-language foundation model for ophthalmology enhanced with clinical knowledge. *TPAMI*, 2025b.

Bo Liu, Donghuan Lu, Dong Wei, Xian Wu, Yan Wang, Yu Zhang, and Yefeng Zheng. Improving medical vision-language contrastive pretraining with semantics-aware triage. *IEEE Transactions on Medical Imaging*, 2023a.

Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226, 2023b.

Kang Liu, Zhuoqi Ma, Siyu Liang, Yunan Li, Xiyue Gao, Chao Liang, Kun Xie, and Qiguang Miao. Seeing like radiologists: Context- and gaze-guided vision-language pretraining for chest x-rays. *arXiv preprint arXiv:2603.26049*, 2026.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

OpenAI. GPT-4o system card, 2024. URL `https://openai.com/index/gpt-4o-system-card/`.

Jonggwon Park, Soobum Kim, Byungmu Yoon, and Kyoyun Choi. RadZero: Similarity-based cross-attention for explainable vision-language alignment in radiology with zero-shot multi-task capability. *arXiv preprint arXiv:2504.07416*, 2025.

Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang, Zhibin Liao, Qi Wu, Minh-Son To, and Johan W. Verjans. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11492–11501, 2024.

Qwen Team. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36:56186–56197, 2023.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in neural information processing systems*, 35:33536–33549, 2022a.

Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 10(1): 1–12, 2020.

Rongsheng Wang, Qingsong Yao, Zihang Jiang, Haoran Lai, Zhiyang He, Xiaodong Tao, and S Kevin Zhou. Ecamp: entity-centered context-aware medical vision language pre-training. *Medical Image Analysis*, 105:103690, 2025. doi: 10.1016/j.media.2025.103690.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald 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*, pp. 2097–2106, 2017.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. METransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11558–11567, 2023a.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023b. doi: 10.1016/j.metrad.2023.100033.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, December 2022b.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. MedKLIP: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Int. Conf. Comput. Vis.*, pp. 21372–21383, 2023.

Yutong Xie, Lin Gu, Tatsuya Harada, Jianpeng Zhang, Yong Xia, and Qi Wu. Rethinking masked image modelling for medical image representation. *Medical Image Analysis*, 98:103304, 2024.

Peng Zhang, Zhihui Lai, Wenting Chen, Xu Wu, and Heng Kong. FaNe: Towards fine-grained cross-modal contrast with false-negative reduction and text-conditioned sparse attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(15):12681–12689, 2026. doi: 10.1609/aaai.v40i15.38264.

Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542, 2023.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.

Ziyang Zhang, Yang Yu, Yucheng Chen, Xulei Yang, and Si Yong Yeo. MedUnifier: Unifying vision-and-language pre-training on medical data with vision generation task using discrete visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29744–29755, 2025.

Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40, 2022.

Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

A Additional Method Details

We first define the model-facing finding anchors (Sec. A.1), then describe their polarity-aware pairing with report sentences (Sec. A.2), and finally document the independently stored report-derived target used by EFS (Sec. A.3). Figure A.1 keeps the corresponding counting units explicit.

A.1 Radiographic Finding Vocabulary

Within FRAME, a *finding* is a token-level lexical anchor that names or characterizes radiographically relevant evidence. This operational definition is narrower than a report-level diagnosis or an ontological concept. Starting from Unified Medical Language System (UMLS) terminology Bodenreider (2004), we manually retained 39 surface forms denoting thoracic findings, disease entities, morphology, or anatomical context (Table A.1). Each retained form maps to one WordPiece token under the saved tokenizer, enabling deterministic matching. A frozen, pretrained CXR-BERT encoder supplies its contextualized feature. Only these 39 tokens form Finding–context Reciprocal Alignment (FRA) queries; the broader parser lexicon in Sec. A.2 determines sentence polarity but cannot create a finding query.

At the study-report level, the vocabulary matches 96.3% of 216,141 reports and 55.0% of their period-delimited sentences. The distinct report- and image-row-level counts are summarized in Figure A.1.

A.2 Report Parsing and Reciprocal Pair Construction

The report is segmented into sentences, after which a rule-based parser assigns each sentence to {positive, possible, negative, other}. It uses 59 parser expressions plus uncertainty, negation, and generic-negative cues. In the default configuration, only occurrences of the 39 anchors within positive or possible sentences are retained. Repeated finding occurrences remain distinct in both branches. Finding-Guided Context Reinforcement (FCR) pairs each occurrence with its containing sentence; Context-Anchored Finding Alignment (CFA) uses that sentence as a visual query and keeps each linked finding as a separate positive anchor. Algorithm A.1 gives the conceptual construction procedure. Negative and other sentences do not form default FRA pairs; Sec. C.1 reports a separate negation-augmented variant. This filter applies only to reciprocal-pair construction, so a separately derived EFS target may still preserve clinically salient negative statements.

The incidence graph in Eq. equation 1 of the main paper supplies the exact positive edges for both directions. FCR aligns each finding-conditioned visual response with its own sentence; CFA aligns each sentence-conditioned response separately with all findings it contains. The representations are compared at relation level. For each other study, similarities to its candidate anchors are aggregated into one negative score as in Eq. equation 3; local losses are then reduced within studies as in Eq. equation 6. Row-level retention and positive/possible sentence composition are shown in Figure A.1**(b–c)**.

A.3 Evidence-Focused Summary Target Construction

EFS is a report-level target intended to retain salient findings and their supported qualifiers; unlike Masked Report Reconstruction (MRR), it does not reproduce every report sentence. Targets are generated offline through direct calls to the GPT-4o API, following the report-level prompting protocol of Wang et al. (2025). No language model is trained, fine-tuned, or distilled to construct these targets. For each source report, GPT-4o receives the fixed one-shot message template shown in Figure A.2, with the report text substituted for `{report}` and the temperature set to zero. The returned summary is stored as the study's EFS target and shared by all radiographs from that study. The formal FRAME loader subsequently reads this precomputed target from the local `llm_output` field without making any API call.

One target is shared by all images from a study. The joined corpus contains 360,091 image–report–target rows from 216,141 studies and 64,731 patients. Mean target and source-report lengths are 16.07 and 57.25 whitespace tokens (28.1%). At study level, 99.4% of target concepts occur in the source, 58.5% of source concepts are retained, and shared concepts have 89.3% polarity agreement.

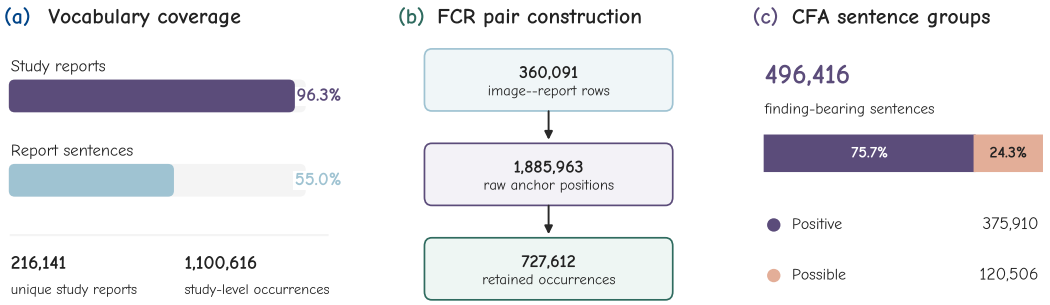

Figure A.1: Corpus coverage and reciprocal-pair construction. **(a)** Vocabulary coverage is measured over 216,141 unique study reports and their period-delimited sentences; 1,100,616 is the number of matched study-report occurrences. **(b–c)** Counts are measured after expansion to 360,091 image–report rows: raw anchor positions are polarity-filtered into finding–sentence edges for FCR and distinct sentence queries for CFA. Each CFA query retains every linked finding as a separate positive anchor. The counts describe within-report supervision units, not independent studies.

Table A.1: Complete 39-token radiographic finding vocabulary. Entries are arranged alphabetically down columns; each maps to one WordPiece token.

Finding anchors (39)		
abnormality	*copd*	*infiltrate*
abnormalities	*deformity*	*mass*
abscess	*distention*	*nodule*
atelectasis	*edema*	*opacity*
bronchiectasis	*effusion*	*opacities*
calcification	*embolism*	*perihilar*
cardiomediastinal	*emphysema*	*pneumonia*
cardiomegaly	*engorgement*	*pneumothorax*
chf	*fibrosis*	*sarcoidosis*
collapse	*granuloma*	*silhouette*
congestion	*hernia*	*thickening*
consolidation	*hilar*	*tuberculosis*
contour	*hemidiaphragm*	*vasculature*

Algorithm A.1: Polarity-aware reciprocal-pair construction.

Input: report R; anchor vocabulary $\mathcal{V}$; parser lexicon $\mathcal{M}$; polarity cues.
Output: edge-level positive pairs $\mathcal{P}_{\mathrm{F}}$ and $\mathcal{P}_{\mathrm{C}}$ for FCR and CFA.

$\mathcal{S} \leftarrow \mathrm{Split}(R)$
$\mathcal{P}_{\mathrm{F}}, \mathcal{P}_{\mathrm{C}} \leftarrow \varnothing$
for each sentence $S \in \mathcal{S}$ **do**
  $\mathcal{U} \leftarrow \mathrm{MatchParser}(S, \mathcal{M})$
  $p \leftarrow \mathrm{ResolvePolarity}(\mathcal{U}, \text{cues})$
  $\mathcal{A} \leftarrow \mathrm{MatchAnchors}(S, \mathcal{V})$
  if $p \in \{\mathrm{POS}, \mathrm{POSS}\}$ and $\mathcal{A} \neq \varnothing$ **then**
    $\mathbf{c} \leftarrow \mathrm{EncodeSentence}(S)$
    for each occurrence $a \in \mathcal{A}$ **do**
      $\mathbf{f} \leftarrow \mathrm{EncodeFinding}(a)$
      $\mathcal{P}_{\mathrm{F}} \leftarrow \mathcal{P}_{\mathrm{F}} \cup \{(\mathbf{f}, \mathbf{c})\}$
      $\mathcal{P}_{\mathrm{C}} \leftarrow \mathcal{P}_{\mathrm{C}} \cup \{(\mathbf{c}, \mathbf{f})\}$
    end for
  end if
end for
return $\mathcal{P}_{\mathrm{F}}, \mathcal{P}_{\mathrm{C}}$

These corpus-level lexical measurements characterize the stored API-generated targets; they are not clinical evaluation metrics.

During pre-training, a trainable 32-query Q-Former and projector condition a frozen Qwen3-4B decoder to predict the stored target from the image alone, with a maximum target length of 128 tokens. The source report is not an input to this pathway. Downstream report generation instead trains a separate adapter to generate the complete report.

```
messages = [
 {
   "role":  "system",
   "content":  f"""
   You are a highly experienced physician with extensive clinical knowledge.
   """
 },
 {
   "role":  "user",  "content":  f"""
   Please analyze the medical report and generate a concise clinical summary.
   Based on the information provided in the report, summarize the mentioned
   findings using the following format:

   "There is [a/an] [adjective] [entity]." when the presence of the finding
   is clearly supported by the report, or
   "There may be [a/an] [adjective] [entity]." when the finding is uncertain
   or only suspected.

   Here, [adjective] describes the severity, extent, or certainty of the finding,
   and [entity] represents the medical abnormality or observation.

   Generate each conclusion as a separate sentence. Keep each sentence
   focused on a single finding only. Do not combine multiple findings using
   "and" or "or". Do not include any explanations or additional text. Output
   only the concise clinical summary in the specified format.

   ---Example Start---
   Report:
   As compared to the prior examination dated there has been no significant
   interval change. low lung volumes resultant crowding of the bronchovascular
   structures. there is no lobar consolidation pleural effusion or pneumothorax.
   the heart size is within normal limits. a large hiatal hernia is again seen.
   multiple known osseous metastases are poorly visualized on todays examination.
   stable appearance of the chest with low lung volumes and a large hiatal hernia.
   no evidence for superimposed acute cardiopulmonary process.
   Summary:
   There is no effusion.
   There is no pneumothorax.
   There is no consolidation.
   There is a large hiatal hernia.
   ---Example End---

   Now follow the same format for the given input.

   Report: {report}
   Summary:
   """
 }
]
```

Figure A.2: Fixed one-shot prompt used to construct EFS targets through direct GPT-4o API calls. For each source report, `{report}` is replaced by the report text and the returned summary is stored as the corresponding target. The displayed report–summary pair is an in-prompt demonstration. No language-model training, fine-tuning, or distillation is performed in this target-construction step.

B Experimental Details

B.1 Datasets, Splits, and Preprocessing

Datasets and prediction targets. MIMIC-CXR v2.0.0 provides the image–report pairs used to pre-train FRAME Johnson et al. (2019a;b). Downstream evaluation covers classification on CXR14 Wang et al. (2017), CheXpert Irvin et al. (2019), RSNA Pneumonia Shih et al. (2019), and COVIDx

Table B.1: Dataset roles, evaluation partitions, labeled-data budgets, and prediction targets used by FRAME. Counts denote images unless another unit is stated; dashes indicate that a partition is not used.

Dataset	Task	Train	Val.	Test	Label budget	Target space
Representation learning						
MIMIC-CXR	Pre-training	360,091 image–report pairs	–	–	Full corpus	FCR, CFA, MRR, and EFS
Classification						
CXR14	Multilabel	78,468	11,219	22,433	785 / 7,847 / 78,468	14 findings
CheXpert	Multilabel	218,414	234	668 (official)	2,184 / 21,841 / 218,414	5 findings
RSNA	Binary	25,184	1,500	3,000	252 / 2,518 / 25,184	Pneumonia
COVIDx	Three-class	27,117	3,013	400	271 / 2,712 / 27,117	Normal, pneumonia, or COVID-19
Spatial prediction and direct transfer						
SIIM	Segmentation	7,969	1,898	1,904	80 / 797 / 7,969	Pneumothorax mask
RSNA	Segmentation and detection	16,010	5,337	5,337	160 / 1,601 / 16,010	Pneumonia mask or box
VinDr-CXR	Direct transfer	–	–	3,000	Not applicable	8 shared findings
Clinical language						
MIMIC-Ext	VQA	217,828 QA	52,840 QA	9,377 QA	217,828 (100%)	Candidate answer
IU X-Ray	Report generation	2,069 studies	296 studies	590 studies	Not applicable	Full report
MIMIC-CXR	Report generation	270,374 images	2,127 images	3,849 images	Not applicable	Full report

Wang et al. (2020); segmentation on SIIM Langer & Shih (2019) and RSNA; and detection on RSNA. The classification targets are 14 findings for CXR14, five findings for CheXpert, pneumonia for RSNA, and normal, pneumonia, or COVID-19 for COVIDx. SIIM predicts pneumothorax masks, whereas the RSNA spatial tasks predict pneumonia masks or boxes.

Medical VQA is evaluated on MIMIC-Ext Bae et al. (2024) and VQA-RAD Lau et al. (2018), using their single-answer and closed-question targets, respectively. Report generation uses MIMIC-CXR and IU X-Ray Demner-Fushman et al. (2016) and predicts complete reports. CheXpert and VinDr-CXR Nguyen et al. (2022) are additionally used for direct cross-dataset transfer. Table B.1 lists the retained training, validation, and test partitions and the corresponding output spaces.

Data splits. The 1%, 10%, and 100% regimes use pre-defined, fixed training lists, with the validation and test sets held constant across label fractions. The exact image, study, or question counts used by FRAME are reported in Table B.1.

Pre-training inputs. The pre-training corpus contains 360,091 radiographs from 216,141 studies and 64,731 patients. Images undergo bicubic random-resized cropping at the native 224×224 resolution, random horizontal flipping, grayscale replication to three channels, and channel-wise normalization. ViT-B/16 consequently produces a 14×14 patch grid. Reports are padded or truncated to 256 WordPiece tokens and EFS targets to 128 tokens. For MRR, 75% of eligible report tokens are replaced by `[MASK]`, and the reconstruction target contains all non-padding tokens from the original report.

Downstream preprocessing. Classification uses 224×224 random-resized crops and horizontal flips for training and a deterministic resize–center-crop view for evaluation. Segmentation applies matched geometric transforms to each image–mask pair; detection uses a deterministic 224×224 resize; and VQA and report generation use grayscale inputs with task-specific training augmentation and deterministic evaluation views at the native encoder resolution.

Table B.2: Shared-label interface for direct transfer. CheXpert normalization uses mean 0.4722 and standard deviation 0.3028. VinDr-CXR preprocessing is selected on validation and then fixed for held-out test inference.

Target	Shared labels	Evaluation input
CheXpert	Atelectasis; Cardiomegaly; Consolidation; Edema; Pleural Effusion ($\leftrightarrow$ Effusion); Pneumonia; Pneumothorax	Bilinear resize and center crop to 224×224; grayscale-to-RGB; source normalization
VinDr-CXR	Atelectasis; Cardiomegaly; Consolidation; Infiltration; Pleural Effusion; Pleural Thickening; Pneumothorax; Pulmonary Fibrosis	Min–max intensity restoration with gamma 2.2; horizontal-flip test-time augmentation selected on the target validation split

Table B.3: Pre-training architecture, initialization, and optimization status. The visual encoder and objective-specific modules are optimized jointly; the pretrained CXR-BERT encoder and Qwen3-4B remain frozen.

Module	Configuration and initialization	Status
Visual encoder	ViT-B/16 Dosovitskiy et al. (2021): 224×224 input, patch size 16, width 768, 12 layers and 12 heads; initialized from MAE-based pretrained weights He et al. (2022)	Trainable
Text encoder	CXR-BERT Boecking et al. (2022): pretrained weights; vocabulary 30,522, width 768, 12 layers and 12 heads	Frozen
FRA branches	Two reciprocal cross-attention branches with separate parameters and shared 768-dimensional projections	Trainable
MRR pathway	Six-layer reconstruction Transformer with multimodal fusion and maximum report length 256	Trainable
Q-Former and projector	32 queries, width 768, six layers, 12 heads, and visual-to-language projector	Trainable
Language model	Qwen3-4B Qwen Team (2025), pretrained and loaded in bf16	Frozen

Direct cross-dataset transfer. The CXR14 1% linear-probe model is transferred without updating the visual encoder or classifier. CheXpert uses seven shared labels: Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion, Pneumonia, and Pneumothorax. VinDr-CXR uses eight: Atelectasis, Cardiomegaly, Consolidation, Infiltration, Pleural Effusion, Pleural Thickening, Pneumothorax, and Pulmonary Fibrosis. Table B.2 gives the exact label mapping and target-specific evaluation preprocessing.

B.2 Pre-training Implementation

Architecture and initialization. The visual encoder is a ViT-B/16 Dosovitskiy et al. (2021) initialized from MAE-based pretrained weights He et al. (2022) and jointly optimized with the FRAME objectives. The text pathway uses a pretrained CXR-BERT encoder whose parameters remain frozen throughout pre-training Boecking et al. (2022). The remaining modules comprise two reciprocal cross-attention branches for Finding–context Reciprocal Alignment (FRA), a six-layer reconstruction Transformer for MRR, a 32-query Q-Former, a visual-to-language projector, and frozen Qwen3-4B Qwen Team (2025). Table B.3 summarizes their configurations and optimization states.

Training objectives. FRAME assigns equal weight to Finding-Guided Context Reinforcement (FCR), Context-Anchored Finding Alignment (CFA), Masked Report Reconstruction (MRR), and Evidence-Focused Summary (EFS). The two reciprocal alignment losses use relation-level, study-grouped symmetric InfoNCE with temperature 0.07; MRR reconstructs the original report tokens, while EFS optimizes the visual Q-Former and projector to predict the stored target through the frozen language decoder.

Optimization and checkpoint. We train FRAME for 30 epochs with AdamW, betas $(0.9, 0.95)$, learning rate 1.5×10^{-4}, and weight decay 0.05. The learning rate warms up linearly for six epochs and then follows cosine decay. Training uses bf16 on four NVIDIA H20 GPUs, a per-GPU batch

Table B.4: Downstream adaptation and evaluation protocols. The visual encoder is frozen for every task; Qwen3-4B is also frozen for VQA and report generation. Only the listed task-specific interface is trained, where applicable.

Task	Dataset	Trainable interface	Optimization and model selection	Test metric
Classification	CXR14, CheXpert, RSNA, COVIDx	Linear classifier	SGD, warmup–cosine; validation AUROC or accuracy	Mean AUROC or accuracy
Segmentation	SIIM, RSNA	Four-stage decoder (512, 256, 128, and 64 channels)	AdamW, focal and Dice-based losses; validation Dice	Dice
Detection	RSNA	Detector projection and YOLO head	AdamW, YOLO losses; validation-selected checkpoint	mAP@IoU 0.40–0.75
Medical VQA	MIMIC-Ext, VQA-RAD	32-query task-specific Q-Former and projector	Validation-selected adapter; normalized predictions	Accuracy
Report generation	MIMIC-CXR, IU X-Ray	Task-specific Q-Former, query embeddings, and projector	Validation-selected full-report adapter and decoding	NLG and clinical metrics
Direct transfer	CXR14 (1% source labels); CheXpert and VinDr-CXR targets	None	Direct shared-label evaluation	Macro and per-label AUROC

size of 64, 12-step gradient accumulation, and an effective batch size of 3,072. The base seed is 42, with rank-specific offsets. The common epoch-10 checkpoint initializes the downstream evaluations reported here.

B.3 Downstream Adaptation and Evaluation Protocols

Across downstream tasks, the pretrained visual encoder is frozen, and a task-specific prediction head or adapter is trained. The shared settings are summarized in Table B.4, and the task-specific optimization settings are consolidated in Tables D.1 and D.2.

For the classification, segmentation, and detection rows that report variability in Tables 1, 2, D.4, and D.5, we use random seeds 42, 123, 3407, 5678, and 9223. Means and standard deviations are computed over these five runs.

Classification. All four datasets use linear probing: a linear classifier is trained on top of the frozen visual encoder. SGD uses momentum 0.9, zero weight decay, and a warmup–cosine learning-rate schedule; multilabel tasks use binary cross-entropy and COVIDx uses cross-entropy. Fraction-specific learning rates and step budgets appear in Table D.1. Checkpoints are selected by validation mean AUROC or accuracy, and test performance is reported as mean AUROC for CXR14, CheXpert, and RSNA and accuracy for COVIDx.

Segmentation. A four-stage convolutional decoder is trained above the frozen visual encoder with a focal and Dice-based objective. AdamW uses learning rate 5×10^{-4}, weight decay 0.05, and cosine decay. SIIM runs for 3,000, 5,000, and 30,000 steps at 1%, 10%, and 100%, respectively; RSNA uses 3,000 steps for each fraction. The checkpoint with the highest validation Dice is evaluated on the test set using Dice.

Detection. The ViT backbone is frozen while the detector projection and head are trained with YOLO losses. The 1%, 10%, and 100% settings use 3,000, 3,000, and 20,000 steps, respectively. AdamW uses learning rate 5×10^{-4}, weight decay 0.05, cosine decay, and maximum gradient norm 1.0. RSNA detection is evaluated by mAP over IoU thresholds 0.40–0.75; this is the RSNA mAP reported in the main paper.

Medical VQA. The visual encoder and Qwen3-4B are frozen, while a task-specific Q-Former and projector with 32 queries are trained for MIMIC-Ext and VQA-RAD. The MIMIC-Ext model trains for five epochs with batch size 16, gradient accumulation 2, learning rate 3×10^{-5}, weight decay 0.01, warmup ratio 0.05, bf16, and seed 42; its checkpoint is selected on validation accuracy. Both

datasets are evaluated by accuracy, with predictions normalized following the standard evaluation protocol.

Report generation. For MIMIC-CXR and IU X-Ray, the visual encoder and Qwen3-4B are frozen, while the task-specific Q-Former, query tokens, and projector are trained. This full-report generation adapter is separate from the pre-training EFS adapter. MIMIC-CXR uses seed 9223, batch size 4, gradient accumulation 8, Q-Former learning rate 5×10^{-5}, and query/projector learning rate 10^{-4}; IU X-Ray uses gradient accumulation 4 and dropout 0.15. Validation selects the checkpoint and decoding configuration. Test reports are generated autoregressively; the IU X-Ray configuration uses beam size 7, no-repeat n-gram size 2, a maximum of 159 new tokens, and length penalty 1.1. We report BLEU, METEOR, ROUGE-L, CheXbert, RadGraph, and GREEN metrics.

Direct transfer. The source encoder and linear classifier from the CXR14 1% linear-probe model are directly transferred to each target dataset without parameter updates. Predictions are restricted to the seven CheXpert or eight VinDr-CXR shared labels listed in Section B.1. Macro AUROC is the unweighted mean of the per-label AUROCs.

C Additional Quantitative Results

The main paper emphasizes ViT-based comparators. This section broadens the backbone coverage with representative CNN-based radiograph–text methods, while also reporting two textual-supervision variants, structural VQA results, and the complete report-generation metric suite. These descriptive extensions improve comparison coverage; they are not causal tests of backbone family.

C.1 Additional Ablations on Textual Supervision

Table D.3 compares FRAME with two epoch-10 variants. *Negation-Augmented Alignment* adds contrastive supervision for negated findings while preserving the default positive/possible pairs. *Full-Report Auxiliary Target* replaces the concise EFS target with the original report under the same 128-token target limit; downstream full-report generation remains a separate task.

The table reports absolute performance because the available runs do not form a perfectly controlled one-variable comparison. Neither variant exceeds FRAME in any displayed column. The widest descriptive separation occurs in the RSNA 1% detection setting: FRAME obtains 26.10 mAP, versus 17.99 for Negation-Augmented Alignment and 8.35 for the Full-Report Auxiliary Target. These single-run results motivate the default design but do not isolate a causal contribution.

C.2 CNN-based Comparisons across Visual Tasks

We additionally compare FRAME with representative CNN-based Med-VLP methods on classification, segmentation, and detection Huang et al. (2021); Wang et al. (2022a); Liu et al. (2023a); Cheng et al. (2023); Wu et al. (2023); Zhang et al. (2023); Wan et al. (2023); Xie et al. (2024); Boecking et al. (2022); Li et al. (2024b); Phan et al. (2024). Tables D.4 and D.5 combine paper-reported entries with explicitly marked local evaluations from released weights; unavailable results are denoted by "–."

Among the displayed values, FRAME is highest in all 12 classification and nine spatial-task columns. Relative to the strongest displayed CNN entry per column, its margins are 1.6–2.5 AUROC points on CXR14, 1.2–1.6 on CheXpert, 1.9–3.1 on RSNA classification, and 2.2–10.6 accuracy points on COVIDx. The corresponding displayed margins are 3.9–8.6 Dice points for segmentation and 2.9–7.1 points for RSNA detection mAP, with the largest generally occurring under limited labels. Because source provenance and training details differ, these are table-level comparisons rather than a controlled backbone ablation.

C.3 Medical VQA by Question Type

The MIMIC-Ext test set contains 6,921 verification, 741 choice, and 1,715 query questions; *Closed* combines verification and choice. With the complete 217,828-question training set, FRAME obtains

80.80% overall accuracy and the highest displayed values for Overall, Closed, Verify, and Choose. MaCo leads Query at 75.90%, followed by FRAME at 74.90% (Table D.6). The structural columns are reported as provided and should not be interpreted as an additive partition reconstructed from a common released prediction file.

Figure D.1 compares the displayed CNN- and ViT-based methods on 272 author-labeled closed questions in VQA-RAD. FRAME correctly answers 194 questions (71.32%). Figure D.2 further shows the 11 official question types. Five questions carry two type labels, so the memberships sum to 277 while overall accuracy remains based on 272 unique questions; small categories should be read with their displayed counts.

C.4 Complete Report-generation Metric Suite

Tables D.7, D.8, and D.9 report the complete report-generation evaluation. The NLG suite comprises BLEU-1–4, METEOR, ROUGE-L, and CIDEr; the clinical suite includes 14- and five-label CheXbert agreement, CheXpert precision, recall, and F1, three RadGraph F1 metrics, and GREEN. R2Gen and R2GenCMN Chen et al. (2020; 2021) are task-specific references and are excluded from ranking among image–text pre-training methods.

Among the displayed image–text pre-training methods, FRAME is highest on all seven NLG metrics for both MIMIC-CXR and IU X-Ray.

On MIMIC-CXR, FRAME leads or ties every displayed clinical metric except five-label exact agreement, where RadZero obtains 0.307 and FRAME and ECAMP both obtain 0.299.

On IU X-Ray, FRAME leads the five CheXbert measures, CheXpert precision, RadGraph partial-relation F1, and GREEN. ECAMP leads CheXpert recall and F1, whereas MRM leads RadGraph entity and complete-relation F1.

D Additional Qualitative Analyses

D.1 Finding-Guided Attention with Reference Boxes

Figure D.3 presents six qualitative cases selected from the predeclared balanced-v3 box-grounded review pool. In each column, the top row shows the radiograph with its source lesion annotation, the bottom row shows the finding-query FCR attention overlay with the same annotation, and the queried finding is named below. The official annotation label is used as the query, and every attention map is produced by the epoch-10 checkpoint.

The green dashed boxes are external RSNA or VinDr-CXR annotations used only for post-hoc comparison; they are neither model inputs nor predicted boxes. These examples passed the review pool's predefined overlap gates and are therefore a curated success-case gallery rather than a representative localization estimate, detector evaluation, or clinical-faithfulness assessment.

D.2 Evidence-Focused Summary Examples

Figure D.4 presents four image-conditioned EFS cases. Each panel places the radiograph beside its original report, the separately stored report-derived reference summary, and the generated summary. Colored highlights connect corresponding findings within each panel; the colors are panel-local visual guides rather than a shared diagnostic-class legend.

The epoch-10 model uses greedy decoding with an 80-new-token cap and a token-boundary sentence-completion stopping check. These outputs belong to the pre-training summary pathway, not the downstream full-report generator. The displayed radiographs occur in the pre-training corpus, so the curated examples provide an in-corpus qualitative view of the pre-training summary pathway rather than held-out or clinician-adjudicated performance.

Table D.1: Image-task optimization settings for classification, segmentation, and detection. Stacked entries follow the 1%, 10%, and 100% labeled-data regimes; warmup and total duration are measured in optimizer steps. LR and WD denote learning rate and weight decay.

(a) Image-task optimization

Task	Dataset	Optimizer / objective	Peak LR (1/10/100%)	Updates (1/10/100%)	Warmup (1/10/100%)	Other explicit settings
Classification	CXR14	SGD; BCE	3×10^{-2} 3×10^{-2} 3×10^{-2}	3,000 3,000 3,000	50 50 200	Momentum 0.9; WD 0; warmup–cosine; validation mean AUROC
Classification	CheXpert	SGD; BCE	3×10^{-3} 3×10^{-3} 1×10^{-3}	9,000 9,000 30,000	150 150 1,500	Momentum 0.9; WD 0; warmup–cosine; validation mean AUROC
Classification	RSNA	SGD; BCE	1×10^{-2} 3×10^{-2} 3×10^{-3}	1,000 5,000 9,000	50 150 1,500	Momentum 0.9; WD 0; warmup–cosine; validation mean AUROC
Classification	COVIDx	SGD; cross-entropy	1×10^{-1} 1×10^{-1} 1×10^{-1}	500 5,000 9,000	30 350 1,750	Momentum 0.9; WD 0; warmup–cosine; validation accuracy
Segmentation	SIIM	AdamW; focal + Dice	5×10^{-4} (all)	3,000 5,000 30,000	– – –	WD 0.05; cosine; validation Dice
Segmentation	RSNA	AdamW; focal + Dice	5×10^{-4} (all)	3,000 3,000 3,000	– – –	WD 0.05; cosine; validation Dice
Detection	RSNA	AdamW; YOLO losses	5×10^{-4} (all)	3,000 3,000 20,000	– – –	WD 0.05; cosine; gradient clip 1.0

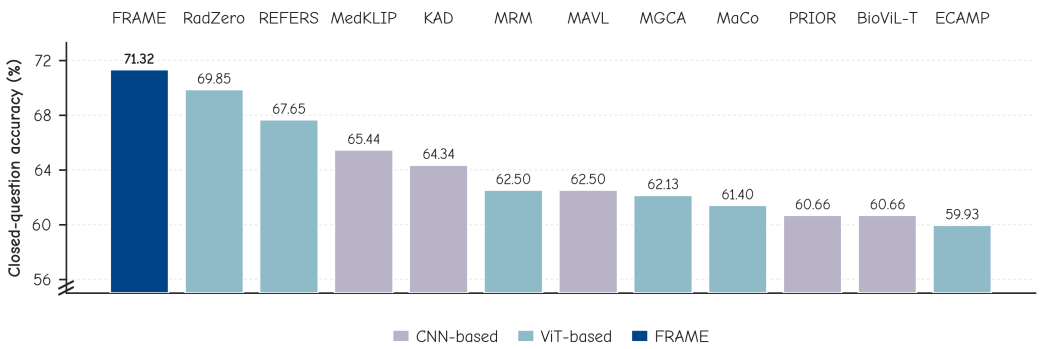

Figure D.1: VQA-RAD exact-match accuracy on 272 author-labeled closed questions. Colors distinguish CNN-based, ViT-based, and FRAME systems; the explicit axis break indicates that the displayed range begins at 55%, while labels retain the exact percentages. FRAME answers 194 questions correctly.

Table D.2: Clinical-language and direct-transfer settings. Dashes denote values not separately specified in Sec. B.3; "not applicable" denotes the absence of target-domain training. The visual encoder is frozen for every task, and Qwen3-4B is additionally frozen for VQA and report generation.

(b) Clinical-language and direct-transfer settings

Task	Dataset	Trainable interface	Duration / batch	Optimization	Selection / decoding
Medical VQA	MIMIC-Ext	32-query Q-Former and projector	5 epochs; batch 16; accumulation 2	LR 3×10^{-5}; WD 0.01; warmup ratio 0.05; bf16; seed 42	Validation accuracy; normalized predictions
Medical VQA	VQA-RAD	32-query Q-Former and projector	–	–	Validation-selected adapter; normalized predictions
Report generation	MIMIC-CXR	Q-Former, query tokens, and projector	Batch 4; accumulation 8	Q-Former LR 5×10^{-5}; query/projector LR 1×10^{-4}; seed 9223	Validation-selected checkpoint and decoding
Report generation	IU X-Ray	Q-Former, query tokens, and projector	Accumulation 4	Dropout 0.15	Validation-selected; beam 7; no-repeat n-gram 2; max 159; length penalty 1.1
Direct transfer	CheXpert, VinDr-CXR	None	No target-domain updates	Not applicable	Seven/eight shared labels; macro and per-label AUROC

Table D.3: Epoch-10 results for alternative textual-supervision designs (all values in %). FRAME denotes the configuration selected for the main paper. Each entry is a single-run measurement on the indicated split. Because task-specific optimization settings differ across designs, the comparison is descriptive and reports absolute values only. Boldface and underlining mark the highest and second-highest displayed values.

(a) Visual and VQA Transfer

	CheXpert AUROC			**RSNA Dice**			**RSNA mAP**			**MIMIC-Ext VQA**
Pre-training design	1%	10%	100%	1%	10%	100%	1%	10%	100%	Acc.
FRAME	**90.50**	**91.20**	**91.50**	**80.00**	**80.90**	**81.60**	**26.10**	**28.50**	**34.00**	**80.80**
Negation-Augmented Alignment	<u>89.20</u>	<u>90.35</u>	<u>90.52</u>	<u>75.37</u>	76.24	<u>72.38</u>	<u>17.99</u>	25.11	<u>29.81</u>	77.19
Full-Report Auxiliary Target	88.27	90.26	90.41	75.24	<u>76.49</u>	71.10	8.35	<u>26.13</u>	26.92	<u>78.24</u>

(b) Report Generation

Pre-training design	**IU X-Ray report generation**			
	BLEU-4	ROUGE-L	METEOR	CheXbert F1
FRAME	**18.30**	**39.80**	**21.20**	**58.30**
Negation-Augmented Alignment	14.94	37.17	18.08	<u>55.94</u>
Full-Report Auxiliary Target	<u>16.70</u>	<u>38.54</u>	<u>19.38</u>	54.51

Table D.4: Frozen-encoder classification with CNN-based radiograph–text pre-training methods at 1%, 10%, and 100% labeled-data fractions. CXR14, CheXpert, and RSNA report AUROC; COVIDx reports accuracy (all in %). Rows marked † were evaluated locally from released weights; other baseline values are taken from the corresponding publications. FRAME follows Sec. B. Published baselines may use different implementation details. Boldface and underlining mark the best and second-best displayed values; "–" denotes an unavailable result. Values retain the precision reported by their respective sources. Subscripts denote standard deviations for the rows where shown.

Method	CXR14			CheXpert			RSNA			COVIDx		
	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
GLoRIA	77.0	81.9	83.8	84.6	85.8	86.2	87.2	88.1	88.9	73.3	87.7	92.1
MGCA (ResNet-50)	78.7	82.7	84.1	88.8	89.1	89.7	89.1	89.9	90.8	74.8	84.8	92.3
SAT	74.4	79.2	81.8	86.9	88.3	88.6	87.4	89.2	90.2	74.5	84.8	89.3
PRIOR	74.9	79.5	82.1	86.3	86.7	87.3	88.2	90.4	90.7	74.8	85.0	89.8
MedKLIP	78.5	80.9	83.2	86.2	86.5	87.7	88.1	90.8	92.0	74.5	85.3	90.3
KAD	78.1	79.8	81.6	86.4	86.9	87.8	87.9	90.0	91.3	75.0	84.3	90.5
Med-UniC	–	–	–	88.2	89.2	89.5	89.1	90.4	90.8	76.5	89.0	92.8
MedIM-C	–	–	–	88.9	89.3	89.7	–	–	–	77.2	90.3	93.6
BioViL-T†	$47.84_{0.1}$	$48.53_{0.1}$	$51.65_{0.0}$	$65.06_{0.1}$	$68.75_{0.1}$	$70.73_{0.1}$	$62.65_{0.1}$	$75.61_{0.1}$	$76.92_{0.1}$	$34.50_{0.9}$	$44.00_{0.4}$	$45.25_{0.1}$
MAVL†	$74.8_{0.1}$	$79.0_{0.1}$	$79.0_{0.0}$	$88.1_{0.1}$	$90.0_{0.1}$	$90.1_{0.1}$	$87.4_{0.1}$	$90.0_{0.1}$	$90.2_{0.1}$	$56.8_{1.0}$	$63.0_{0.2}$	$64.5_{0.1}$
FRAME	$\mathbf{81.2_{0.1}}$	$\mathbf{85.0_{0.1}}$	$\mathbf{85.7_{0.0}}$	$\mathbf{90.5_{0.2}}$	$\mathbf{91.2_{0.1}}$	$\mathbf{91.5_{0.1}}$	$\mathbf{92.2_{0.2}}$	$\mathbf{93.1_{0.1}}$	$\mathbf{93.9_{0.1}}$	$\mathbf{87.8_{1.2}}$	$\mathbf{95.5_{0.3}}$	$\mathbf{95.8_{0.2}}$

Table D.5: Semantic segmentation and object detection with CNN-based radiograph–text pre-training methods at 1%, 10%, and 100% labeled-data fractions (all in %). Rows marked † were evaluated locally from released weights; other baseline values are taken from the corresponding publications. FRAME follows Sec. B. Published baselines may use different implementation details. The detection columns report RSNA mAP. Boldface and underlining mark the best and second-best displayed values; "–" denotes an unavailable result. Values retain the precision reported by their respective sources. Subscripts denote standard deviations for the rows where shown.

Method	SIIM Dice			RSNA Dice			RSNA mAP		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
MGCA (ResNet-50)	49.7	59.3	64.2	63.0	68.3	69.8	12.9	16.8	24.9
MedKLIP	50.2	60.8	63.9	66.2	69.4	71.9	8.9	16.3	24.5
KAD	35.8	46.9	63.4	59.3	67.5	67.8	9.8	14.8	18.8
SAT	59.2	68.2	74.7	67.8	74.8	75.3	10.6	19.1	22.0
PRIOR	54.6	63.7	76.8	68.7	71.9	74.4	13.8	20.5	25.6
Med-UniC	56.7	62.2	64.4	72.6	74.4	76.7	16.6	22.3	31.1
MedIM-C	–	63.5	81.3	–	–	–	–	–	–
BioViL-T†	$56.45_{0.6}$	$58.61_{0.3}$	$66.37_{0.1}$	$66.81_{0.7}$	$67.89_{0.4}$	$75.39_{0.1}$	$6.54_{0.7}$	$16.90_{0.4}$	$19.38_{0.2}$
MLIP	51.6	60.8	68.1	67.7	68.8	73.5	17.2	19.1	25.8
MAVL†	$60.6_{0.8}$	$63.1_{0.5}$	$79.4_{0.1}$	$72.8_{0.7}$	$75.7_{0.5}$	$77.7_{0.1}$	$19.0_{0.7}$	$21.5_{0.5}$	$23.5_{0.1}$
FRAME	$\mathbf{69.2_{0.9}}$	$\mathbf{73.8_{0.4}}$	$\mathbf{86.4_{0.2}}$	$\mathbf{80.0_{0.6}}$	$\mathbf{80.9_{0.4}}$	$\mathbf{81.6_{0.1}}$	$\mathbf{26.1_{0.8}}$	$\mathbf{28.5_{0.4}}$	$\mathbf{34.0_{0.2}}$

Table D.6: Full-data MIMIC-Ext accuracy (%) across five structural question types: Overall, Closed, Verify, Choose, and Query. Closed denotes the dataset's combined verification-and-choice category. Boldface and underlining mark the highest and second-highest displayed values in each column. All rows use the same local normalized exact-match evaluator on the released test split.

Family	Method	Overall	Closed	Verify	Choose	Query
Number of questions		9,377	7,662	6,921	741	1,715
CNN	MedKLIP	76.30	78.14	81.10	50.30	68.30
	PRIOR	78.00	79.33	81.10	<u>60.40</u>	72.10
	BioViL-T	65.88	67.80	70.31	44.40	57.32
	KAD	77.90	79.54	81.70	59.00	70.40
	MAVL	76.40	77.75	80.38	53.17	70.38
ViT	REFERS	75.78	77.63	80.23	53.31	67.52
	MGCA	74.98	77.15	79.69	53.44	65.31
	MRM	78.40	79.30	82.00	54.00	74.50
	MaCo	78.20	78.78	81.80	50.60	**75.90**
	ECAMP	78.70	79.71	<u>82.40</u>	50.20	74.30
	RadZero	<u>78.90</u>	<u>80.25</u>	82.10	52.60	74.10
FRAME		**80.80**	**80.38**	**82.80**	**61.70**	<u>74.90</u>

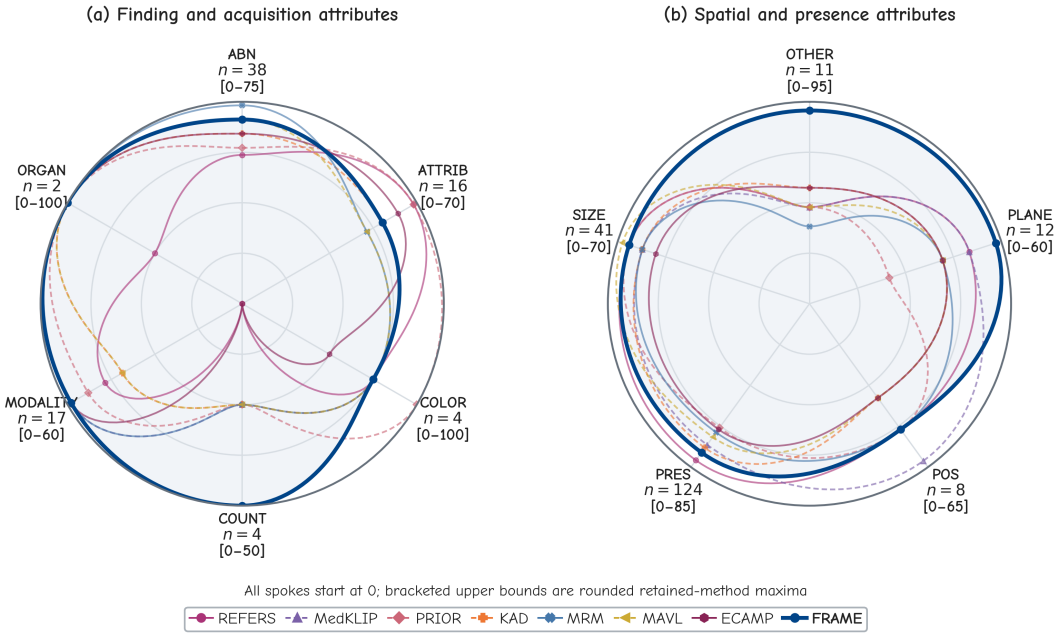

Figure D.2: VQA-RAD accuracy by the 11 official question types, grouped for readability into **(a)** finding and acquisition attributes and **(b)** spatial and presence attributes. Both panels use the same circular radar construction. Every spoke starts at 0; its bracketed upper bound is the maximum among the displayed methods rounded upward to the next five percentage points, and one mapping is shared by all methods on that spoke. Five dual-labeled questions enter two axes, producing 277 memberships from 272 unique questions. Results in small categories are descriptive. Dashed and solid baseline traces denote CNN- and ViT-based systems, respectively; FRAME is emphasized in dark blue.

Table D.7: Complete natural-language generation results on MIMIC-CXR and IU X-Ray. BL-1–4 denote BLEU-1–4; all values are fractions. Boldface and underlining mark the best and second-best results, respectively, among the CNN- and ViT-based image–text pre-training methods. These rows are local evaluations under the common downstream framework; task-specific generators are included only as published references.

(a) MIMIC-CXR

Method	BL-1	BL-2	BL-3	BL-4	METEOR	ROUGE-L	CIDEr
Task-specific references							
R2Gen†	0.353	0.218	0.145	0.103	0.142	0.277	–
R2GenCMN†	0.353	0.218	0.148	0.106	0.142	0.278	–
CNN-based image–text pre-training							
MedKLIP	0.223	0.147	0.100	0.069	0.112	0.268	0.050
PRIOR	0.218	0.147	0.104	0.076	0.115	0.283	0.073
BioViL-T	0.154	0.103	0.072	0.052	0.097	0.265	0.049
KAD	<u>0.303</u>	<u>0.201</u>	<u>0.140</u>	<u>0.102</u>	<u>0.134</u>	<u>0.295</u>	0.100
MAVL	0.252	0.171	0.121	0.088	0.125	0.292	0.092
ViT-based image–text pre-training							
REFERS	0.241	0.162	0.114	0.082	0.121	0.287	0.081
MGCA (ViT-B/16)	0.238	0.158	0.111	0.080	0.116	0.284	0.085
MRM	0.206	0.142	0.102	0.076	0.117	0.286	0.090
MaCo	0.247	0.168	0.119	0.087	0.121	0.288	0.084
ECAMP	0.273	0.184	0.130	0.094	0.131	<u>0.295</u>	<u>0.107</u>
RadZero	0.259	0.176	0.126	0.093	0.127	0.291	0.097
FRAME	**0.355**	**0.234**	**0.161**	**0.115**	**0.145**	**0.297**	**0.114**

(b) IU X-Ray

Method	BL-1	BL-2	BL-3	BL-4	METEOR	ROUGE-L	CIDEr
Task-specific references							
R2Gen†	0.470	0.304	0.219	0.165	0.187	0.371	–
R2GenCMN†	0.475	0.309	0.222	0.170	0.191	0.375	–
CNN-based image–text pre-training							
MedKLIP	0.468	0.298	0.218	0.171	0.190	0.355	<u>0.614</u>
PRIOR	<u>0.496</u>	<u>0.322</u>	0.231	0.168	<u>0.204</u>	0.379	0.382
BioViL-T	0.436	0.295	0.219	0.166	0.186	0.389	0.348
KAD	0.472	0.318	<u>0.235</u>	<u>0.179</u>	0.195	<u>0.392</u>	0.388
MAVL	0.476	0.309	0.221	0.164	0.200	0.370	0.522
ViT-based image–text pre-training							
REFERS	0.448	0.293	0.208	0.154	0.187	0.378	0.375
MGCA (ViT-B/16)	0.443	0.300	0.223	0.170	0.189	0.381	0.352
MRM	0.388	0.256	0.188	0.143	0.179	0.381	0.405
MaCo	0.426	0.278	0.201	0.152	0.186	0.361	0.467
ECAMP	0.477	0.307	0.222	0.170	0.197	0.363	0.579
RadZero	0.449	0.300	0.220	0.168	0.193	0.386	0.541
FRAME	**0.498**	**0.329**	**0.236**	**0.183**	**0.212**	**0.398**	**0.649**

†Published task-specific report generators, included as references and excluded from the image–text pre-training ranking.

Table D.8: Complete clinical agreement on the 3,849-image MIMIC-CXR test set. Micro-14, Macro-14, Micro-5, and Macro-5 are F1 scores; Exact-5 is five-label exact agreement. P, R, and F1 are 14-label micro CheXpert scores, while Entity, Partial, and Complete are the three RadGraph F1 variants. All values are fractions. Boldface and underlining mark the best and second-best image–text pre-training results; ties share the same formatting. CNN- and ViT-based image–text pre-training rows are scored with common local metric implementations.

(a) CheXbert label agreement

Method	Micro-14	Macro-14	Micro-5	Macro-5	Exact-5
CNN-based image–text pre-training					
MedKLIP	0.295	0.156	0.430	0.289	0.242
PRIOR	0.240	0.153	0.291	0.226	0.286
BioViL-T	0.030	0.009	0.004	0.003	0.264
KAD	0.347	0.201	0.437	0.335	0.284
MAVL	0.356	0.241	0.427	0.341	0.287
ViT-based image–text pre-training					
REFERS	0.336	0.210	0.399	0.315	0.294
MGCA (ViT-B/16)	0.309	0.182	0.347	0.256	0.281
MRM	0.315	0.198	0.357	0.277	0.289
MaCo	0.345	0.208	0.437	0.332	0.287
ECAMP	<u>0.387</u>	<u>0.267</u>	<u>0.444</u>	<u>0.362</u>	<u>0.299</u>
RadZero	0.333	0.201	0.401	0.306	**0.307**
FRAME	**0.422**	**0.286**	**0.489**	**0.414**	<u>0.299</u>

(b) CheXpert, RadGraph, and GREEN

	CheXpert			RadGraph			
Method	P	R	F1	Entity	Partial	Complete	GREEN
Task-specific reference							
R2GenCMN[†]	0.334	0.275	0.278	–	–	–	–
CNN-based image–text pre-training							
MedKLIP	0.435	0.238	0.308	0.238	0.222	0.165	0.248
PRIOR	0.392	0.163	0.230	0.227	0.209	0.160	0.267
BioViL-T	0.065	0.020	0.030	0.202	0.194	0.152	0.195
KAD	0.528	0.262	0.350	0.255	<u>0.234</u>	0.172	0.283
MAVL	0.525	0.270	0.357	<u>0.256</u>	0.233	0.171	0.288
ViT-based image–text pre-training							
REFERS	0.519	0.240	0.329	0.254	0.232	0.175	0.281
MGCA (ViT-B/16)	0.485	0.247	0.327	0.223	0.201	0.145	0.268
MRM	0.509	0.232	0.319	0.251	0.231	0.175	0.287
MaCo	0.503	0.256	0.340	0.241	0.225	0.166	0.270
ECAMP	0.534	<u>0.304</u>	<u>0.389</u>	**0.269**	0.213	<u>0.181</u>	<u>0.293</u>
RadZero	<u>0.540</u>	0.247	0.339	0.252	0.232	0.175	0.288
FRAME	**0.542**	**0.364**	**0.435**	**0.269**	**0.248**	**0.185**	**0.304**

[†]Published task-specific result, included as a reference and excluded from the image–text pre-training ranking.

Table D.9: Complete clinical agreement on the 590-study IU X-Ray test set. Column definitions and units match Table D.8. Boldface and underlining mark the best and second-best results, respectively, among the CNN- and ViT-based image–text pre-training methods. All image–text pre-training rows are scored with common local metric implementations.

(a) CheXbert label agreement

Method	Micro-14	Macro-14	Micro-5	Macro-5	Exact-5
CNN-based image–text pre-training					
MedKLIP	<u>0.563</u>	0.077	<u>0.077</u>	<u>0.047</u>	<u>0.827</u>
PRIOR	0.482	0.049	0.000	0.000	0.824
BioViL-T	0.540	0.050	0.000	0.000	0.824
KAD	0.549	0.056	0.000	0.000	0.824
MAVL	0.551	<u>0.084</u>	0.042	0.038	0.803
ViT-based image–text pre-training					
REFERS	0.515	0.048	0.000	0.000	0.824
MGCA (ViT-B/16)	0.543	0.052	0.016	0.006	0.824
MRM	0.540	0.050	0.000	0.000	0.824
MaCo	0.551	0.054	0.016	0.006	0.825
ECAMP	0.562	0.058	0.016	0.006	0.825
RadZero	0.559	0.059	0.016	0.006	0.824
FRAME	**0.583**	**0.112**	**0.209**	**0.130**	**0.829**

(b) CheXpert, RadGraph, and GREEN

	CheXpert			RadGraph			
Method	P	R	F1	Entity	Partial	Complete	GREEN
CNN-based image–text pre-training							
MedKLIP	0.597	0.492	0.539	0.368	0.335	0.274	0.661
PRIOR	<u>0.701</u>	0.319	0.439	0.390	0.365	0.261	0.650
BioViL-T	0.544	0.486	0.513	0.392	0.371	0.286	0.647
KAD	0.568	0.475	0.517	0.405	0.376	0.276	0.658
MAVL	0.573	0.455	0.508	0.385	0.359	0.272	0.643
ViT-based image–text pre-training							
REFERS	0.547	0.381	0.449	0.412	0.389	0.293	0.624
MGCA (ViT-B/16)	0.547	0.487	0.515	0.405	0.379	0.295	0.656
MRM	0.544	0.486	0.513	**0.443**	<u>0.421</u>	**0.359**	0.600
MaCo	0.562	0.489	0.523	0.419	0.390	0.299	0.601
ECAMP	0.599	**0.529**	**0.562**	0.348	0.321	0.247	<u>0.664</u>
RadZero	0.581	0.483	0.527	<u>0.440</u>	0.418	<u>0.335</u>	0.637
FRAME	**0.703**	<u>0.495</u>	<u>0.542</u>	0.430	**0.433**	0.300	**0.680**

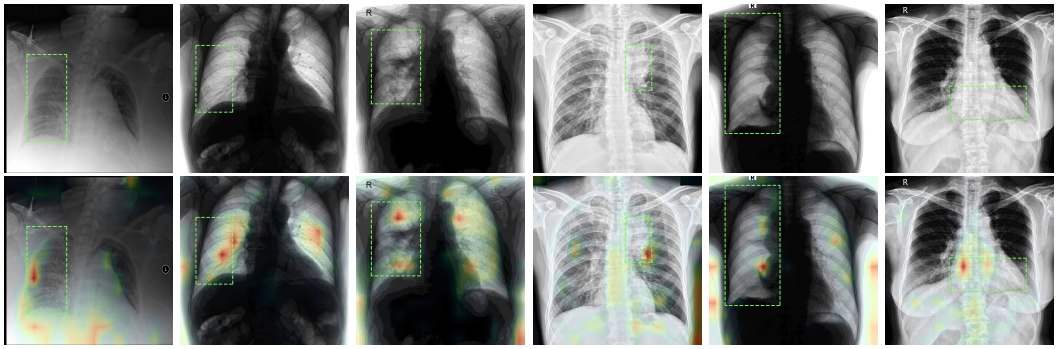

Figure D.3: Finding-guided attention with reference lesion boxes. Each column shows the source radiograph with its dataset annotation (top), the FCR attention overlay with the same annotation (bottom), and the queried finding below. Green dashed boxes are external ground-truth references, not model inputs, predicted boxes, or detection outputs. The six pathology-diverse cases are curated high-overlap examples from the balanced-v3 manual-review pool and do not estimate population-level localization performance.

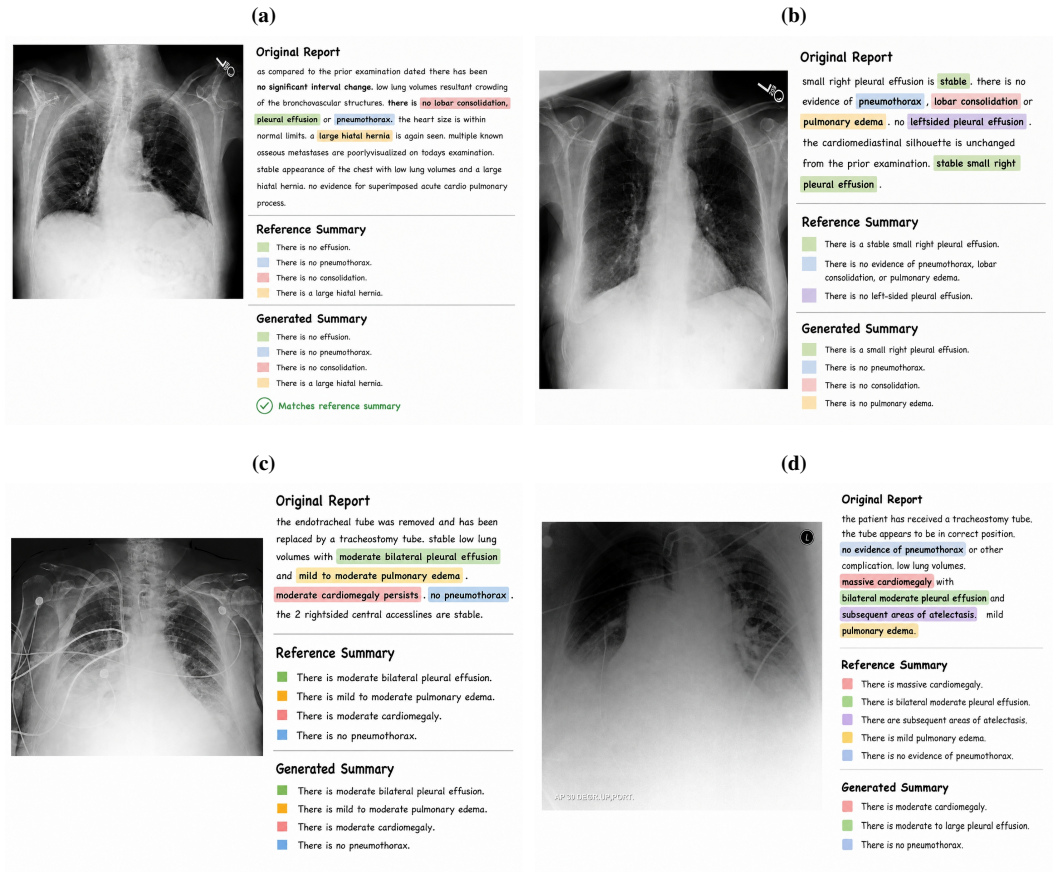

Figure D.4: Four curated image-conditioned EFS examples. Each panel pairs the radiograph with its original report, report-derived reference summary, and generated summary. Panel-local colors link corresponding findings and do not define classes across panels. The examples include exact and partial matches with omitted or altered target details; they illustrate the pre-training summary pathway rather than downstream full-report, held-out, or clinician-adjudicated performance.