RadJEPA: Radiology Encoder for Chest X-Rays via Joint Embedding Predictive Architecture

Anas Khan[1,4*] **Mariam Husain**[2†] **Pratik Jalan**[1,3†] **Kshitij Jadhav**[3]
[1]Department of Computer Science (CSE), Indian Institute of Technology Bombay, India
[2]Department of Biomedical Engineering, Johns Hopkins University, USA
[3]Koita Centre for Digital Health, Indian Institute of Technology Bombay, India
[4]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
{anaskhan, pratik}@cse.iitb.ac.in anas.khan@mbzuai.ac.ae
mhusai10@jh.edu kshitij.jadhav@iitb.ac.in

Abstract

Vision-language pretraining has driven progress in medical image representation learning, but it depends on paired image-text data and can inherit reporting bias from clinical narratives. We study whether language-free predictive pretraining can produce an image encoder that transfers effectively to radiology report generation. RadJEPA is a chest-X-ray adaptation of I-JEPA, pretrained on approximately 840K unlabeled radiographs using latent context-to-target prediction. Our primary contribution is an extensive empirical evaluation of this language-free encoder for report generation: the frozen image encoder is coupled to a trainable two-layer projector and language decoder, and is also substituted into four established vision-language backbones. Across MIMIC-CXR and IU-Xray, RadJEPA matches or exceeds the evaluated image-only and image-text baselines on lexical, entity-relation, and clinical-label metrics. Controlled MIMIC-only comparisons provide evidence that the predictive objective contributes beyond domain-specific pretraining, while broader comparisons also reflect differences in pretraining data, model capacity, and input resolution. Complementary classification and segmentation experiments assess transfer beyond report generation. Code and pretrained weights are available at [1].

1 Introduction

Vision–language modeling has largely been driven by the use of textual supervision (Radford et al., 2021). In biomedical imaging (Zhou et al., 2023), representations are often reused as fixed visual tokens for downstream reasoning, such as report generation, image classification, and segmentation. This paradigm implicitly assumes that textual descriptions provide a sufficiently complete and unbiased account of visual content (Dehghani et al., 2023). More broadly, multimodal generation studies (Khan et al., 2024) have shown that multimodal interpretation can be text-dominant. Radiology reports emphasize selective findings, frequently omitting subtle variations or absent observations. As a result, visual encoders trained to align images with text may inherit bias (Jones et al., 2024) and fail to preserve the full semantic.

Recent self-supervised alternatives such as DINO-style self-distillation have emerged as powerful label-free approaches (Caron et al., 2021; Oquab et al., 2023). These methods learn invariances by aligning global representations across multiple augmented views of the same image using a teacher–student formulation in chest X-rays (Pérez-García et al., 2025). Self-distillation methods (Dong et al., 2023) emphasize invariance by aligning representations across augmented views.

Chest X-rays are fundamentally *semantic* rather than purely view-centric (Çallı et al., 2021): clinical interpretation depends on global anatomical context, spatial relationships, and subtle deviations from normal structure. Consequently, representation learning strategies that rely primarily on image–text alignment or view-level invariance may not explicitly model conditional dependencies within the image. Instead, we adopt a *predictive latent modeling* perspective.

Unlike contrastive alignment or self-distillation, this objective explicitly models conditional structure in representation space, encouraging abstraction of context-dependent semantic information without pixel reconstruction or textual supervision. JEPA formalize by predicting abstract latent representations of masked regions (Assran et al., 2023). Latent predictive architectures have shown promise in neuroimaging (Dong et al., 2024), but they remain relatively underexplored in medical imaging.

[*]Work done during Master's at IIT Bombay.
[†]**Mariam Husain and Pratik Jalan contributed equally.**
[1]RadJEPA Model Page

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 30238–30259
October 24-29, 2026 ©2026 Association for Computational Linguistics

Our main contributions are as follows:

- We adapt the I-JEPA latent predictive pretraining framework to chest X-rays and pretrain a ViT-B/14 encoder on 839,364 images from five open-source datasets, without report or label supervision.
- We provide an extensive empirical study of language-free visual pretraining for radiology report generation. We evaluate a frozen RadJEPA encoder with a trainable projector and Vicuna-7B decoder and replace the native vision encoder in four established VLM pipelines, covering five language decoders in total.
- We compare against image-text, DINO-style, JEPA-style baselines; include data-matched control comparisons to separate objective effects from domain-pretraining effects; and report complementary classification and segmentation results under common downstream protocols.

2 Related Works

Chest X-ray representation learning methods can be broadly divided into image-text alignment and image-only self-supervised approaches. Vision-language models such as CLIP, BioViL, BiomedCLIP, CheXzero, and MRM learn joint image-text embeddings using paired radiographs and reports. CLIP (Radford et al., 2021) learns shared image-text embeddings using paired radiographs and reports. BioViL (Bannur et al., 2023) incorporates localized visual grounding, BiomedCLIP (Zhang et al., 2023) scales biomedical contrastive pretraining, CheXzero (Tiu et al., 2022) enables zero-shot diagnosis generation, and MRM (Zhou et al., 2023) improves multimodal representation matching. Medical image-conditioned language generation has also been explored beyond chest radiography. For example (Khan and Khan, 2024), combines visual representations with class-enriched textual information for otoscopic images.

Despite strong downstream performance, image-text contrastive frameworks have several limitations. They heavily rely on paired radiology reports, which often contain noise, reporting bias, and inconsistent terminology (Jones et al., 2024). Furthermore, contrastive objectives mainly optimize global image-text alignment and may fail to capture fine-grained anatomical structure while also introducing shortcut correlations from language supervision (Sun et al., 2024).

To reduce dependence on language supervision, recent works explored image-only self-supervised learning methods. DINO-v2 (Oquab et al., 2023) learns representations through self-distillation between teacher and student networks using augmented views, while RAD-DINO (Pérez-García et al., 2025) adapts this framework specifically for chest X-ray representation learning using large-scale radiology datasets. Although these methods achieve strong downstream performance, they still rely heavily on augmentation consistency and explicit embedding alignment, which may encourage invariant shortcut features rather than contextual predictive understanding.

I-JEPA (Assran et al., 2023) introduces predictive representation learning in latent space instead of reconstruction or contrastive alignment. Recent work such as CheXWorld (Yue et al., 2025) explored I-JEPA for chest X-ray representation learning; however, its evaluation was limited and did not include report generation or comprehensive AUPRC-based classification analysis. We compare it extensively in our work, with additional details provided in Section 13.

In contrast, our proposed RadJEPA adopts a JEPA-style predictive learning framework for radiology representation learning. Rather than aligning image-text embeddings or enforcing augmented-view consistency, RadJEPA learns contextual latent prediction directly from chest X-ray images without language supervision. This enables the model to focus more on structural and semantic understanding of radiological patterns while avoiding dependence on paired reports during pretraining.

3 Pretraining Setup and Baselines

RadJEPA pretraining follows the I-JEPA protocol (Assran et al., 2023), as described in Section 4 and Figure 1 (Parameters details c.f. Table 19). We adopt a data configuration aligned with RAD-DINO (Pérez-García et al., 2025), leveraging large-scale chest X-ray datasets summarised in Table 1. In publicly available datasets, frontal views substantially outnumber lateral views (approximately 6:1). Unlike RAD-DINO, which incorporates approximately 90K proprietary images, we rely exclusively on open-source data. To mitigate viewpoint imbalance, we augment the lateral subset with approximately 90K additional lateral chest X-rays

Dataset	View	Patient cohort	#Subjects	#Images	Frontal	Lateral
BRAX (Reis et al., 2022)	frontal, lateral	institutional PACS	19,351	41,620	24,959	16,661
CheXpert (Irvin et al., 2019)	frontal, lateral	inpatient, outpatient	65,240	224,316	191,229	33,087
MIMIC-CXR (Johnson et al., 2019)	frontal, lateral	ICU	188,546	300,491	210,491	90,000
ChestX-ray14 (Wang et al., 2017)	frontal	not specified	32,717	112,120	112,120	0
PadChest (Bustos et al., 2020)	frontal, lateral	all available	67,000	160,817	96,287	64,530
Total			**372,854**	**839,364**	**635,086**	**204,278**

Table 1: Chest X-ray datasets used for RadJEPA pretraining. For MIMIC-CXR, only a subset of subjects is included to avoid overlap with the evaluation sets.

Model type	Model	Arch.	# Params.	Training dataset	# Images	# Text	Resolution
Image & Text	CLIP@224 (Radford et al., 2021)	ViT-L/14	304 M	WebImageText	400 M	400 M	224^2
Image & Text	CLIP@336 (Radford et al., 2021)	ViT-L/14	304 M	WebImageText	400 M	400 M	336^2
Image & Text	BioViL-T (Bannur et al., 2023)	ResNet50	27 M	MIMIC-CXR	197 k	174 k	512^2
Image & Text	BioMedCLIP (Zhang et al., 2023)	ViT-B/16	86 M	PMC-15M	15 M	15 M	224^2
Image & Text	CheXzero (Tiu et al., 2022)	ViT-B/32	151 M	MIMIC-CXR	377 k	227 k	224^2
Image & Text	MRM (Zhou et al., 2023)	ViT-B/16	86 M	MIMIC-CXR	377 k	227 k	448^2
Image Only	CheXWorld (Yue et al., 2025)	ViT-B/14	86M	multiple datasets	448K	–	224^2
Image Only	DINO-v2 (Oquab et al., 2023)	ViT-G/14	1.1 B	LVD	142 M	–	518^2
Image Only	DINO-v3 (Siméoni et al., 2025)	ViT-B/16	86 M	LVD-1689M	1.689 B	–	224^2
Image Only	RAD-DINO$_{\text{control}}$	ViT-B/14	87 M	MIMIC-CXR	197 k	–	518^2
Image Only	RAD-DINO (Pérez-García et al., 2025)	ViT-B/14	87 M	Multi-CXR	838 k	–	518^2
Image Only	I-JEPA (Assran et al., 2023)	ViT-H/14	0.6 B	IN1K	1.28 M	–	448^2
Image Only	RadJEPA$_{\text{control}}$	ViT-B/14	86 M	MIMIC-CXR	197 k	–	224^2
Image Only	RadJEPA	ViT-B/14	86 M	Multi-CXR	839 k	–	224^2

Table 2: Overview of image backbones and their training dataset characteristics employed in experimental analysis.

from MIMIC-CXR, giving a final frontal-to-lateral ratio of approximately 3:1. In total, RadJEPA is pretrained on 839,364 chest X-ray images. A set of baseline approaches is considered for experimental analysis (Table 2), with RAD-DINO serving as the key image-only baseline. Comparison with other generic image-only self-supervised methods is outside the scope of this study, as it has been extensively explored in prior work (Oquab et al., 2023; Caron et al., 2021). As summarised in Table 2, RadJEPA uses a ViT-B/14 backbone; despite its smaller capacity, this encoder consistently outperforms substantially larger image-only backbones in the downstream evaluations.

4 Methodology

Let $x \in \mathbb{R}^{H\times W}$ denote a chest X-ray. RadJEPA pretraining operates by sampling a context region c and multiple target regions t. Visual encoder f extracts latent representations $z_c = f(c)$ and $z_t = f(t)$, a predictor g estimates the target embedding as $\hat{z}_t = g(z_c)$. Model is trained using a latent prediction objective $\mathcal{L}_{\text{JEPA}} = \mathbb{E}_{(c,t)}\left[\text{SmoothL1}\left(\text{stopgrad}(z_t), \hat{z}_t\right)\right]$. The target encoder parameters θ' are updated via EMA $\theta' \leftarrow \tau\theta' + (1-\tau)\theta$, where $\tau \in [0, 1)$. After pretraining, the encoder f is frozen and reused across downstream tasks (cf. Figure 1).

Disease Classification. Given an image x, frozen features $h = f(x)$ are passed to a linear classifier producing probabilities $p = \text{softmax}(Wh)$, optimized using cross-entropy $\mathcal{L}_{\text{cls}} = -\sum_{k=1}^{K} y_k \log p_k$. Only the classification head is trained.

Semantic Segmentation. Multi-scale features $\{h_l\}_{l=1}^{L} = f^{(l)}(x)$ are extracted and aggregated by a UperNet decoder $\mathcal{D}$ to produce pixel-wise logits $S = \mathcal{D}(\{h_l\})$. The segmentation loss is $\mathcal{L}_{\text{seg}} = -\sum_{i,j}\sum_{c=1}^{C} y_{ij}^{c} \log \text{softmax}(S_{ij}^{c})$. Only decoder parameters are optimized.

Report Generation. Frozen visual embeddings $v = f(x)$ are projected into the language space through a two-layer MLP adapter $\tilde{v} = A(v) = v + \lambda W_2\, \sigma(W_1 v)$, where σ is the GeLU nonlinearity and λ is a learnable residual scalar initialised to 1. The projected embeddings are concatenated with an instruction prompt and passed to a language model $\mathcal{M}$ (Vicuna-7B in the main experiments), which generates reports autoregressively with $p(y_t \mid y_{<t}, x) = \mathcal{M}(y_t \mid y_{<t}, \tilde{v})$ and is optimised via negative log-likelihood $\mathcal{L}_{\text{txt}} = -\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, x)$. The encoder f remains frozen; only A and $\mathcal{M}$ are updated.

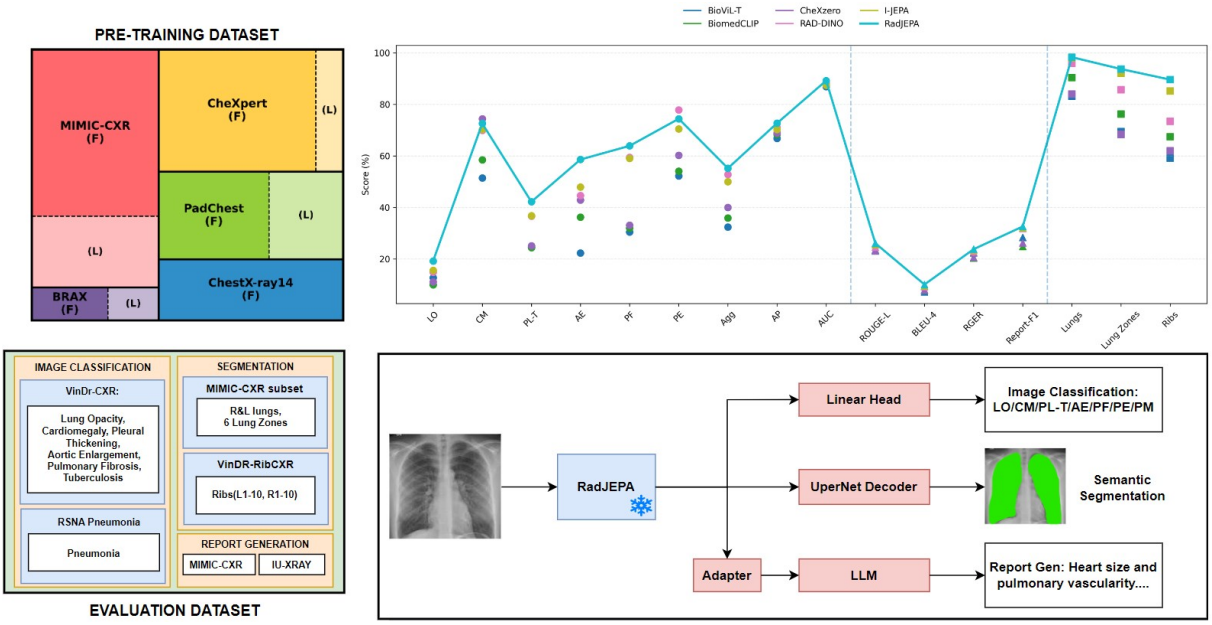


Figure 1: Overview of the RadJEPA framework and evaluation. **Top-left:** Proportional visualization of pre-training datasets, where area reflects dataset size; (F) and (L) denote frontal and lateral views, respectively. **Bottom-left:** Summary of evaluation datasets across image classification, segmentation, and report generation. **Bottom-right:** Model pipeline during downstream evaluation, RadJEPA is frozen, while task-specific heads (linear head, UPerNet decoder, adapter + LLM) are trained. **Top-right:** Unified performance comparison across three downstream tasks, reporting five baselines consistently evaluated across all tasks.

5 Evaluation

Classification and Segmentation: We also evaluated RadJEPA on classification and segmentation tasks. Although these tasks are not directly related to computational linguistics, as they do not involve textual modalities, they are important for understanding the potential and generalization capability of RadJEPA on foundational vision tasks. Note that we could not evaluate performance on the CANDID-PTX dataset (Feng et al., 2021), as, upon contacting the authors, the dataset is no longer publicly available as it was previously. A detailed comparison is provided in Section 10.

5.1 Compare with Existing Vision Encoders

Datasets and Implementation Details

We evaluate report generation on **MIMIC-CXR** and **IU-Xray**. For MIMIC-CXR we follow the official splits used in RAD-DINO (Pérez-García et al., 2025), retaining only frontal chest X-rays with a *Findings* section, yielding 146,909/7,250/2,461 train/validation/test image-text pairs. These splits are used to fine-tune the language decoder; the image encoder remains frozen. To probe out-of-distribution generalisation we additionally report results on IU-Xray (3,306 test studies), which is not used during the training of either the encoder or the decoder. The pipeline follows the LLaVA-1.5 recipe (Liu et al., 2024): patch embeddings from the frozen image encoder are projected through a two-layer MLP and concatenated with the instruction prompt "⟨image_tokens⟩ Provide a description of the findings in the radiology image.". We use Vicuna-7B (v1.5) as the language model $\mathcal{M}$. Training uses AdamW with cosine scheduling at peak learning rate 2×10^{-5}, batch size 128 (32 per GPU) with 3% warmup, on four NVIDIA A100 nodes for three epochs. During inference, we use 32-bit precision, up to 150 tokens, and a batch size of 1.

We report standard report-generation metrics. **ROUGE-L** (Lin, 2004) and **BLEU-4** (Papineni et al., 2002) measure n-gram overlap with the reference *Findings*. $\mathbf{RG_{ER}}$ denotes the entity-relation F1 score from the RadGraph parser of Yu et al. (2023); Jain et al. (2021) and captures the alignment of clinical entities and their relations. **Macro-F1-14** is the macro-averaged F1 of the 14 CheXpert-style labels extracted from the generated and reference reports by the CheXbert labeller (Smit et al., 2020).

Results analysis

Table 3 reports report generation results on MIMIC-CXR and IU-Xray using a fixed Vicuna-7B de-

Dataset	Image encoder	Input res.	# Tokens	ROUGE-L	BLEU-4	RG_{ER}	Macro-F1-14
MIMIC	CLIP@224(Radford et al., 2021)	224×224	256	23.0 [22.7, 23.4]	8.3 [7.9, 8.6]	20.3 [19.8, 20.7]	24.7 [23.6, 26.0]
	CLIP@336(Radford et al., 2021)	336×336	576	23.3 [22.9, 23.7]	8.4 [8.0, 8.7]	20.4 [19.9, 20.9]	25.3 [24.2, 26.5]
	DINO-v2 (Oquab et al., 2023)	518×518	261	22.7 [22.4, 23.2]	7.6 [7.3, 7.9]	18.5 [18.1, 19.1]	18.6 [17.8, 19.5]
	DINOv3 (Siméoni et al., 2025)	224×224	201	21.4 [21.0, 21.8]	7.3 [7.0, 7.7]	18.8 [18.3, 19.2]	18.1 [17.2, 19.0]
	BiomedCLIP (Zhang et al., 2023)	224×224	256	23.1 [22.8, 23.5]	7.9 [7.5, 8.2]	20.4 [19.9, 20.8]	24.9 [23.8, 26.1]
	CheXzero (Tiu et al., 2022)	224×224	49	23.2 [22.9, 23.6]	8.0 [7.7, 8.4]	20.6 [20.2, 21.1]	26.2 [25.0, 27.5]
	BioViL-T(Bannur et al., 2023)	512×512	196	23.5 [23.2, 23.9]	7.3 [7.0, 7.6]	22.4 [21.9, 22.8]	28.4 [27.2, 29.8]
	$RAD\text{-}DINO_{control}$	518×518	1369	24.2 [23.8, 24.6]	9.0 [8.7, 9.4]	22.4 [21.9, 22.9]	31.5 [30.1, 32.9]
	RAD-DINO (Pérez-García et al., 2025)	518×518	1369	25.1 [24.7, 25.4]	9.4 [9.0, 9.8]	23.2 [22.7, 23.6]	32.0 [30.5, 33.5]
	CheXWorld (Yue et al., 2025)	224×224	256	24.1 [23.6, 24.5]	8.9 [8.4, 9.3]	20.1 [19.7, 20.4]	30.2 [28.8, 31.7]
	I-JEPA (Assran et al., 2023)	224×224	256	23.2 [22.8, 23.6]	8.1 [7.7, 8.5]	21.6 [21.2, 22.0]	29.4 [27.9, 31.0]
	$RadJEPA_{control}$	224×224	256	25.5 [25.1, 25.8]	9.7 [9.3, 10.0]	23.3 [22.9, 23.7]	31.9 [30.7, 33.2]
	RadJEPA	224×224	256	**26.1 [25.7, 26.4]**	**10.1 [9.7, 10.4]**	**23.8 [23.4, 24.2]**	**32.6 [31.2, 33.8]**
IU	CLIP@224(Radford et al., 2021)	224×224	256	25.4 [25.1, 25.7]	9.2 [8.9, 9.5]	25.8 [25.3, 26.2]	18.1 [16.1, 20.8]
	CLIP@336(Radford et al., 2021)	336×336	576	25.3 [24.9, 25.6]	8.0 [7.8, 8.3]	25.3 [24.8, 25.6]	18.5 [16.7, 20.8]
	DINO-v2 (Oquab et al., 2023)	518×518	261	25.4 [25.1, 25.7]	8.0 [7.7, 8.2]	23.6 [23.2, 24.0]	12.3 [10.6, 14.1]
	DINOv3 (Siméoni et al., 2025)	224×224	201	25.3 [24.9, 25.7]	7.8 [7.5, 8.1]	23.9 [23.6, 24.0]	12.5 [11.7, 13.4]
	BiomedCLIP (Zhang et al., 2023)	224×224	256	20.2 [19.9, 20.4]	6.3 [6.1, 6.5]	20.0 [19.7, 20.4]	7.1 [5.9, 8.5]
	CheXzero (Tiu et al., 2022)	224×224	49	25.6 [25.2, 25.9]	8.5 [8.2, 8.8]	25.7 [25.2, 26.1]	18.1 [16.3, 20.1]
	BioViL-T(Bannur et al., 2023)	512×512	196	26.3 [25.9, 26.6]	8.2 [7.9, 8.4]	25.3 [24.9, 25.7]	20.2 [18.0, 23.0]
	$RAD\text{-}DINO_{control}$	518×518	1369	25.5 [25.2, 25.9]	9.2 [8.9, 9.4]	26.2 [25.8, 26.6]	23.8 [21.4, 26.3]
	RAD-DINO (Pérez-García et al., 2025)	518×518	1369	26.3 [25.9, 26.7]	9.4 [9.1, 9.8]	26.6 [26.1, 27.0]	26.4 [24.0, 28.7]
	CheXWorld (Yue et al., 2025)	224×224	256	26.1 [25.8, 26.5]	9.1 [8.6, 9.6]	26.3 [25.8, 26.9]	24.9 [23.4, 26.5]
	I-JEPA (Assran et al., 2023)	224×224	256	25.8 [25.3, 26.2]	8.9 [8.5, 9.2]	25.9 [25.4, 26.4]	25.4 [23.1, 27.8]
	$RadJEPA_{control}$	224×224	256	27.1 [26.7, 27.4]	9.6 [9.3, 9.9]	27.0 [26.5, 27.4]	26.8 [24.8, 29.1]
	RadJEPA	224×224	256	**28.4 [28.0, 28.7]**	**9.9 [9.6, 10.2]**	**27.5 [27.1, 27.9]**	**27.6 [25.3, 29.8]**

Table 3: Downstream radiology report generation results on MIMIC-CXR and IU-Xray. In all models, the image encoder remains frozen, while the two-layer MLP projector and the language decoder are fully fine-tuned. We report median and 95% confidence intervals from 500 bootstrap samples. $RadJEPA_{control}$ denotes the variant pretrained only on MIMIC-CXR for an architecture-matched comparison with $RAD\text{-}DINO_{control}$ (cf. Table 2). Numbers in bold are the best in each column. The statistical protocol is described in Section 12; complementary visual analysis is in Section 17.

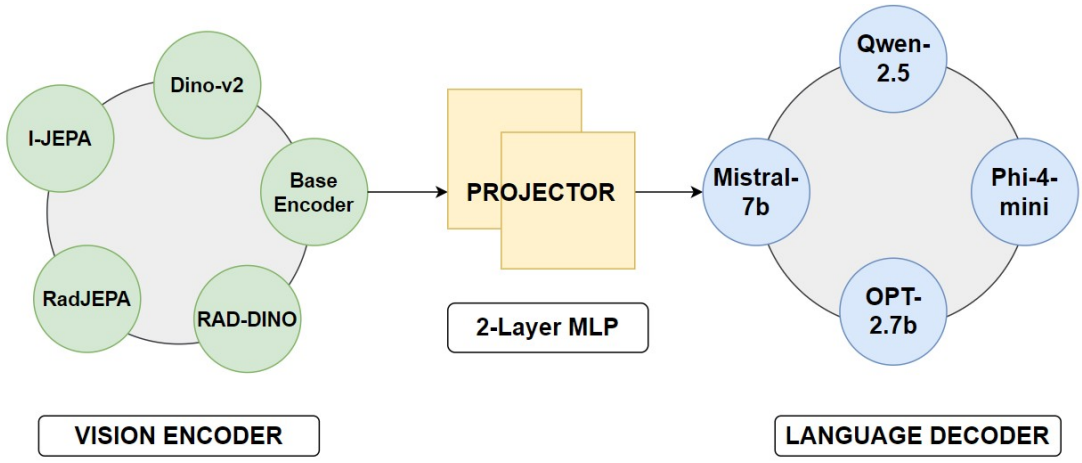


Figure 2: Encoder-swap protocol used in Table 4. The native vision encoder of each VLM is replaced by DINO-v2, I-JEPA, RAD-DINO, or RadJEPA while the projector and language decoder are kept unchanged. With four VLM backbones and four encoder swaps (plus the four base baselines), this gives 20 configurations per dataset and 40 runs in total across MIMIC-CXR and IU-Xray.

coder with frozen image encoders. RadJEPA achieves the best performance across all metrics and datasets among image-only, predictive, and vision-language baselines. On MIMIC-CXR, it attains the highest ROUGE-L (**26.1**), BLEU-4 (**10.1**), RG_{ER} (**23.8**), and Macro-F1-14 (**32.6**), outperforming RAD-DINO, I-JEPA, and prior vision-language models. On IU-Xray, RadJEPA further improves ROUGE-L by **+1.7**, RG_{ER} by **+0.9**, and Macro-F1-14 by **+1.3** over the strongest non-JEPA baseline, indicating robust out-of-distribution generalization. These gains are achieved despite operating at lower input resolution than contrastive baselines such as RAD-DINO. Under a controlled MIMIC-only pretraining setting, $RadJEPA_{control}$ matches or exceeds architecture-matched baselines, while

Dataset	VLMs	Encoder	ROUGE-L	BLEU-4	RG_{ER}	Macro-F1-14
MIMIC-CXR	MedLLaVA	Base Encoder	21.4 [21.0, 21.8]	5.3 [5.0, 5.6]	16.8 [16.4, 17.2]	18.9 [17.1, 20.7]
		I-JEPA	24.9 [24.4, 25.4]	9.1 [8.7, 9.5]	19.9 [19.4, 20.4]	27.4 [25.5, 29.5]
		DINO-v2	22.9 [22.5, 23.3]	7.8 [7.5, 8.1]	18.3 [17.9, 18.8]	14.7 [13.0, 16.5]
		DINO-v3	22.3 [21.9, 22.7]	7.6 [7.3, 8.0]	18.0 [17.6, 18.4]	14.6 [12.9, 16.4]
		RAD-DINO	25.2 [24.8, 25.6]	9.0 [8.7, 9.4]	21.9 [21.4, 22.3]	29.2 [27.7, 30.9]
		RadJEPA	**25.9 [25.6, 26.2]**	**9.6 [9.3, 9.9]**	**22.4 [22.1, 22.8]**	**29.8 [28.5, 31.1]**
	Qwen-2.5	Base Encoder	22.9 [22.5, 23.3]	5.1 [4.8, 5.4]	18.5 [18.0, 18.9]	16.6 [14.8, 18.5]
		I-JEPA	27.2 [26.7, 27.7]	10.2 [9.8, 10.6]	22.5 [22.0, 23.0]	29.8 [27.9, 31.8]
		DINO-v2	24.0 [23.6, 24.4]	8.4 [8.1, 8.7]	18.7 [18.2, 19.2]	16.8 [14.9, 18.7]
		DINO-v3	23.7 [23.3, 24.1]	8.5 [8.1, 8.8]	18.8 [18.3, 19.3]	16.2 [14.4, 18.1]
		RAD-DINO	26.8 [26.4, 27.2]	9.9 [9.5, 10.3]	22.8 [22.3, 23.3]	31.3 [29.6, 33.1]
		RadJEPA	**28.1 [27.8, 28.4]**	**10.7 [10.4, 11.0]**	**23.5 [23.2, 23.9]**	**32.9 [31.4, 34.2]**
	BLIP-2	Base Encoder	23.0 [22.6, 23.4]	5.6 [5.3, 5.9]	17.8 [17.4, 18.2]	18.5 [16.6, 20.4]
		I-JEPA	25.7 [25.2, 26.2]	9.1 [8.7, 9.5]	23.9 [23.4, 24.4]	24.7 [22.9, 26.7]
		DINO-v2	24.3 [23.9, 24.7]	7.9 [7.6, 8.2]	19.8 [19.3, 20.3]	15.3 [13.5, 17.1]
		DINO-v3	24.0 [23.6, 24.4]	7.6 [7.3, 7.8]	19.7 [19.2, 20.3]	14.9 [13.1, 16.8]
		RAD-DINO	26.5 [26.1, 26.9]	9.4 [9.0, 9.8]	23.5 [23.0, 24.0]	29.9 [28.2, 31.6]
		RadJEPA	**27.9 [27.6, 28.2]**	**10.3 [10.0, 10.6]**	**24.1 [23.7, 24.4]**	**31.7 [30.2, 33.0]**
	Phi-4	Base Encoder	18.2 [17.8, 18.6]	4.8 [4.5, 5.1]	16.7 [16.3, 17.1]	11.8 [10.2, 13.6]
		I-JEPA	20.7 [20.2, 21.2]	7.5 [7.1, 7.9]	18.4 [17.9, 18.9]	27.9 [26.0, 29.9]
		DINO-v2	19.8 [19.4, 20.2]	6.2 [5.9, 6.5]	17.1 [16.7, 17.5]	13.9 [12.1, 15.8]
		DINO-v3	19.6 [19.2, 20.0]	5.8 [5.5, 6.2]	17.2 [16.8, 17.6]	13.0 [11.3, 14.8]
		RAD-DINO	21.7 [21.3, 22.1]	7.7 [7.3, 8.0]	20.3 [19.9, 20.7]	**29.5 [28.0, 31.2]**
		RadJEPA	**23.5 [23.2, 23.8]**	**8.6 [8.3, 8.9]**	**20.7 [20.4, 21.0]**	29.3 [27.9, 30.8]
IU-Xray	MedLLaVA	Base Encoder	22.3 [21.9, 22.8]	4.8 [4.5, 5.2]	21.1 [20.6, 21.6]	13.7 [11.8, 15.9]
		I-JEPA	26.2 [25.6, 26.8]	8.4 [8.0, 8.8]	25.2 [24.6, 25.8]	22.5 [20.3, 24.8]
		DINO-v2	25.4 [24.9, 25.9]	7.8 [7.4, 8.2]	21.9 [21.3, 22.5]	11.7 [9.8, 13.8]
		DINO-v3	24.8 [24.3, 25.3]	7.5 [7.1, 7.9]	22.2 [21.6, 22.9]	11.3 [9.4, 13.4]
		RAD-DINO	26.7 [26.2, 27.1]	8.6 [8.2, 9.0]	25.8 [25.2, 26.3]	23.9 [21.7, 26.2]
		RadJEPA	**27.8 [27.4, 28.1]**	**9.1 [8.8, 9.4]**	**26.3 [25.9, 26.7]**	**25.8 [23.8, 27.8]**
	Qwen-2.5	Base Encoder	23.6 [23.1, 24.1]	4.7 [4.3, 5.1]	23.3 [22.8, 23.8]	14.9 [12.8, 17.1]
		I-JEPA	29.0 [28.4, 29.7]	9.3 [8.8, 9.8]	25.9 [25.3, 26.6]	22.5 [20.0, 25.1]
		DINO-v2	27.1 [26.6, 27.7]	8.3 [7.9, 8.7]	22.6 [22.0, 23.2]	12.0 [10.1, 14.0]
		DINO-v3	26.7 [26.2, 27.2]	7.9 [7.5, 8.4]	22.3 [21.8, 22.9]	11.7 [9.8, 13.8]
		RAD-DINO	28.4 [27.9, 28.9]	9.6 [9.2, 10.0]	26.1 [25.5, 26.7]	25.8 [23.5, 28.2]
		RadJEPA	**30.2 [29.8, 30.6]**	**10.1 [9.8, 10.4]**	**27.7 [27.2, 28.2]**	**26.9 [24.8, 29.0]**
	BLIP-2	Base Encoder	26.8 [26.3, 27.4]	5.1 [4.7, 5.5]	20.1 [19.5, 20.7]	12.3 [10.4, 14.5]
		I-JEPA	27.5 [26.9, 28.1]	9.1 [8.6, 9.6]	24.9 [24.3, 25.5]	23.6 [21.3, 26.0]
		DINO-v2	27.3 [26.8, 27.8]	7.9 [7.5, 8.3]	21.8 [21.2, 22.4]	10.8 [8.9, 12.8]
		DINO-v3	27.0 [26.5, 27.5]	7.8 [7.4, 8.3]	22.0 [21.4, 22.6]	10.2 [8.4, 12.2]
		RAD-DINO	28.2 [27.7, 28.7]	9.0 [8.6, 9.4]	26.6 [26.0, 27.2]	25.3 [23.1, 27.6]
		RadJEPA	**29.7 [29.3, 30.1]**	**9.4 [9.1, 9.7]**	**28.2 [27.7, 28.7]**	**25.9 [23.8, 28.0]**
	Phi-4	Base Encoder	20.2 [19.7, 20.7]	4.2 [3.9, 4.6]	18.4 [17.9, 18.9]	9.6 [7.8, 11.5]
		I-JEPA	24.1 [23.5, 24.7]	6.9 [6.5, 7.3]	22.8 [22.2, 23.5]	23.9 [21.6, 26.3]
		DINO-v2	22.9 [22.4, 23.4]	6.1 [5.7, 6.5]	20.7 [20.1, 21.2]	10.0 [8.1, 12.0]
		DINO-v3	22.6 [22.1, 23.1]	5.8 [5.4, 6.2]	20.9 [20.3, 21.5]	9.6 [7.8, 11.5]
		RAD-DINO	23.8 [23.3, 24.3]	7.2 [6.8, 7.6]	24.9 [24.3, 25.5]	**24.5 [22.2, 26.9]**
		RadJEPA	**25.0 [24.6, 25.4]**	**8.3 [8.0, 8.6]**	**25.3 [24.8, 25.8]**	24.2 [22.0, 26.5]

Table 4: Report-generation performance when the native vision encoder of each VLM is replaced by DINO-v2, I-JEPA, RAD-DINO, or RadJEPA (cf. Figure 2). *Base Encoder* denotes the unmodified VLM. All experiments use the identical training recipe described in Section 5.1. We report median and 95% confidence intervals from 500 bootstrap samples. For each VLM-metric cell, bold marks the best result among the encoder swaps; underline marks the second-best when the gap to the best is within the bootstrap interval (used only when RadJEPA is the close runner-up to highlight that the gap is not statistically meaningful). The full statistical protocol is in Section 12; complementary visual analysis is in Section 17.2.

full RadJEPA yields additional improvements. The same ranking holds under the LLM-based GREEN clinical-correctness metric (Ostmeier et al., 2024) (Appendix 15).

5.2 Extended Comparison with RAD-DINO

The objective of our work is to compare the performance of image-text paired vision encoders and DINO-style self-distillation methods against Joint Embedding Predictive Architectures (JEPA). In Table 3, we compare both image-text paired vision encoders and DINO-style self-distillation encoders (DINO-v2 and RAD-DINO). In this subsection, we further extend the comparison of RadJEPA with a particular focus on RAD-DINO and its foundational model DINO-v2 to support our claim that, under similar settings, Joint Embedding Predictive Architectures demonstrate superior performance

over DINO-style view-augmented self-supervised models for radiology chest X-ray images. We do not include comparisons with other SSL methods such as SimCLR, as no large-scale chest X-ray pretrained SimCLR (Chen et al., 2020) models exist, falling outside the scope of this study.

Baseline Models: We evaluate four widely used Vision-Language Models (VLMs): MedLLaVA (Li et al., 2023a), Phi-4 (Abdin et al., 2024), Qwen-2.5 (Yang et al., 2025a), and BLIP-2 (Li et al., 2023b). We choose MedLLaVA instead of the original LLaVA, as LLaVA employs the Vicuna-7B (v1.5) decoder, extensively evaluated in Table 3. Public checkpoint URLs for every backbone, decoder, and image encoder used in this paper are listed in Appendix 11.

For all experiments, each VLM is trained on the MIMIC-CXR dataset under the same experimental setting described in Section 5.1, including identical dataset splits and prompting strategies. We first evaluate each VLM without modifying its original architecture, and report these results in Table 4 under the *Base Encoder* setting. The corresponding language decoders are Mistral-7B for MedLLaVA, Qwen2.5-7B for Qwen-2.5, OPT-2.7B (Zhang et al., 2022) for BLIP-2, and Phi-4-mini for Phi-4. We then replace the original vision encoder of each VLM with DINO-v2, I-JEPA, RAD-DINO, and RadJEPA (cf. Figure 2), while keeping the remaining architecture unchanged.

Experimental Setup: We follow the standard LLaVA-style architecture, where embeddings from the frozen vision encoder are projected through a two-layer MLP to match the embedding dimension required by the language decoder. Since both RadJEPA and RAD-DINO are based on the ViT-B architecture, we restrict our experiments to language models with up to 7B parameters. Except for the vision encoder, all remaining components, including the projection layer and the LLM decoder, are fully trainable during fine-tuning. The complete results are reported in Table 4.

Result Analysis: RadJEPA achieves the best overall performance in almost every cell of Table 4, with RAD-DINO as the typical runner-up. On MIMIC-CXR, the relative gains over RAD-DINO are approximately **2.8%**, **8.1%**, **2.7%**, and **4.6%** on the four metrics, averaged across the four VLM backbones; on IU-Xray the same gains are **4.0%**, **5.7%**, **4.8%**, and **3.2%**. The only exception is Macro-F1-14 with the Phi-4 backbone, where RAD-DINO is ahead by **0.7–1.2%** across the two datasets, likely reflecting the smaller ($\approx$3.5B) Phi-4-mini decoder versus the $\approx$7B Mistral/Qwen/Vicuna decoders used by the other VLMs. DINO-v3 underperforms both RadJEPA across all downstream tasks and DINO-v2 across many downstream tasks. This is likely influenced by the publicly available DINO-v3 model using a ViT-B/16 backbone[2] with 86M parameters, comparable in scale to RAD-DINO and RadJEPA, whereas DINO-v2 uses a substantially larger ViT-G/14 backbone with 1.1B parameters.

	$\mathcal{N}^+$ 7,667	$\mathcal{N}^-$ 195,425
$\mathcal{B}^+$	3,552 (0.463)	30,988 (0.159)
$\mathcal{B}^-$	1,345 (0.175)	153,822 (0.787)

Table 5: Data distribution bias of Pleural Effusion ($\mathcal{B}$) conditioned on Pneumothorax ($\mathcal{N}$), i.e., $P(\mathcal{B}^*|\mathcal{N}^*)$, in the MIMIC-CXR dataset. '+' and '-' denote abnormal and normal conditions, respectively. Integer values represent the number of samples, while the decimal values in parentheses denote conditional probabilities.

6 Error Analysis

The largest dataset used during RadJEPA pretraining is MIMIC-CXR-JPG, contributing approximately **36%** of the total pretraining data. Chest X-ray datasets often suffer from dataset-specific biases, as they are collected under constrained clinical settings (Jones et al., 2024). For example, MIMIC-CXR was collected from Beth Israel Deaconess Medical Center, primarily from ICU patients, making the dataset vulnerable to strong correlations between findings that may not necessarily be causally related. One such example is the correlation between *Pleural Effusion* and *Pneumothorax* (Song et al., 2024); cf. Table 5.

During qualitative analysis, we observed multiple bias behaviours in RadJEPA-generated reports where the model simultaneously predicted the absence of both pleural effusion and pneumothorax, even when one of these findings was clearly present in the ground-truth report. To quantify this, we manually re-examined the 200 cases in our human-annotated sample (Section 14.1) whose ground-truth reports mention pleural effusion or pneumothorax. Within this subset, RadJEPA produced 33 such joint-absence errors against 37 for

[2] facebook/dinov3-vitb16-pretrain-lvd1689m

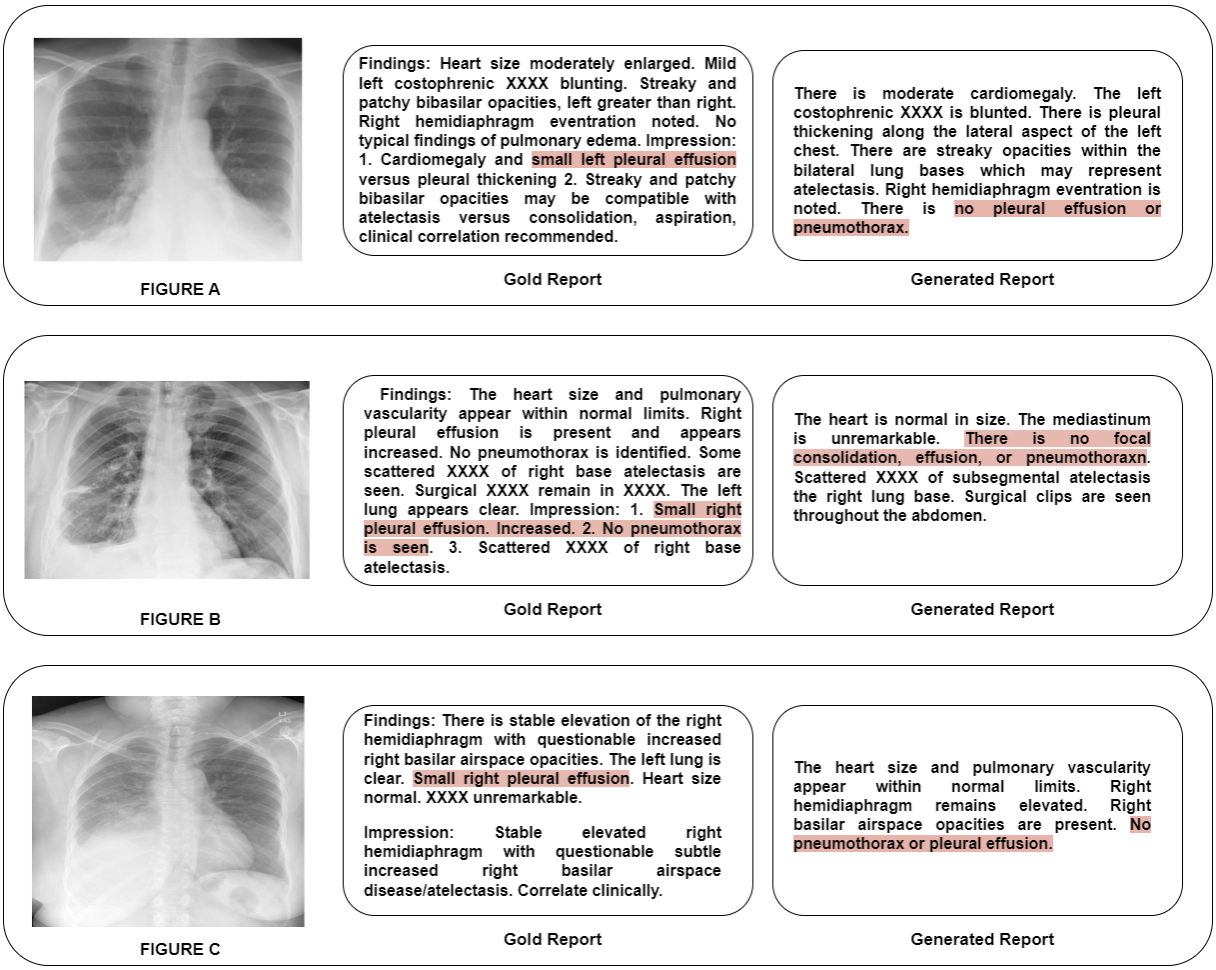


Figure 3: Examples of error in RadJEPA-generated reports, where the model predicts the absence of pleural effusion and pneumothorax despite positive findings in the ground-truth report.

RAD-DINO under the same decoder. The reduction is modest and we surface it here so that future work is aware of the failure mode rather than treat it as solved. All examples shown in Figure 3 belong to the IU-Xray dataset and use the Vicuna-7B decoder. In the first example (Figure A), the original report mentions a small left pleural effusion, whereas the RadJEPA-generated report incorrectly states that neither pleural effusion nor pneumothorax is present. Similarly, in the second example (Figure B), a right pleural effusion is present in the ground-truth report, but the generated report again predicts the absence of both findings. In the third example (Figure C), a similar scenario was observed. Beyond the representative failures in Figure 3, Appendix 14.1 provides a broader qualitative analysis of the human-reviewed reports, including false-positive findings, hallucinated clinical content, missed subtle abnormalities, and examples from low-frequency disease categories.

Inter-Annotator Agreement: Two clinical annotators and three LLM judges (aggregated by per-metric majority vote) independently graded a stratified 500-report sample drawn equally from IU-Xray and MIMIC-CXR. Inter-human quadratic-weighted Cohen's κ is 0.91 on the 0–5 score and 0.83 on the binary acceptability flag (Landis and Koch, 1977); the LLM vote agrees with the humans at κ=0.84 and 0.81, and the five-rater Krippendorff's α is 0.77. All values fall in the substantial-to-almost-perfect range on the Landis-Koch scale. Full statistics are in Appendix 14.1.

7 Conclusion & Future Works

We introduced RadJEPA, a language-supervision-free Joint Embedding Predictive encoder for chest X-rays, pretrained on ≈840K unlabeled images. On report generation, RadJEPA matches or exceeds the strongest image-only and vision-language baselines across four VLM backbones, two datasets, and four metrics, and shows consistent improve-

ments on classification and segmentation. The radiologist study demonstrates high agreement in judgments of RadJEPA generated reports, but it does not constitute a paired clinical comparison with baselines. Future work should evaluate broader institutions and views, perform paired radiologist comparisons, and study bias and collapse mitigation in JEPA-style radiology encoders.

8 Limitation

In this section, we briefly discuss the limitations of our work arising from both the characteristics of the datasets and the JEPA-style architecture used in this study.

8.1 Dataset Limitations

Although RadJEPA is pretrained on approximately 840K chest X-ray images collected from five large open-source datasets, these datasets still contain several inherent limitations. Most datasets, such as MIMIC-CXR and CheXpert, are collected from specific hospitals and clinical environments, introducing demographic, acquisition, and disease-distribution biases that may not generalize across institutions. Additionally, several instances contain strong label co-occurrence correlations (e.g., pleural effusion and pneumothorax) (Song et al., 2024), which may encourage shortcut learning rather than true causal understanding (Jones et al., 2024). The datasets also vary considerably in image quality, acquisition protocols, and reporting styles, introducing domain shifts during pretraining. Similarly, the downstream evaluation datasets also contain limitations. Chest X-ray benchmarking datasets are relatively small and may not fully reflect the diversity and complexity of real-world clinical settings, a limitation also commonly observed in multimodal evaluation benchmarks outside the medical domain (Khan et al., 2025). Furthermore, evaluations are restricted to frontal chest X-rays and may not fully capture the diversity of real-world radiology settings, thereby limiting broader clinical generalization.

Similarly, the downstream evaluation datasets also contain limitations. Chest X-ray evaluation benchmarks are relatively small and may not fully reflect the diversity and complexity of real-world clinical settings, a limitation also observed in other multimodal evaluation benchmarks (Khan et al., 2025). This make performance estimates noisy and restrict conclusions about broader generalization.

8.2 Architectural Limitations

One limitation of JEPA-style architectures is that collapse prevention is not explicitly enforced at the objective level, but instead relies on architectural design choices such as teacher-student asymmetry and EMA-based updates. As discussed in prior theoretical analyses comparing PMAX (similar to JEPA-style architectures) with PMIN, PMAX type of architecture (JEPA) does not directly constrain internal redundancy or enforce entropy balancing within representations, allowing correlated or redundant latent features to persist. In practice, this may lead to occasional representation collapse. Moreover, JEPA-style models remain sensitive to factors such as masking strategy, teacher momentum, and dataset bias, all of which can affect training stability and representation quality.

9 Ethical Considerations

All datasets used in this work are publicly available open-source datasets obtained from their respective official sources and appropriately cited throughout the paper. The datasets are de-identified and do not contain personally identifiable patient information. Furthermore, all pretrained models, VLM backbones, and baseline methods used in this study are open-source and used in accordance with their respective licenses and research guidelines. As discussed in the Limitation section (cf. Section 8), both dataset bias and architectural bias may affect the learned representations of RadJEPA. Despite efforts to mitigate such biases, it remains challenging to eliminate biased or discriminatory patterns from the learned representations and generated outputs. AI assistants were used for writing support and sentence polishing.

Acknowledgement

We gratefully acknowledge the Computation for Indian Language Technology (CFILT) - IIT Bombay, Koita Centre for Digital Health - IIT Bombay, IIT Bombay Trust Lab, and the Department of Computer Science and Engineering - IIT Bombay, for their support, resources, and funding that enabled this work. We are sincerely grateful for their generous support. We also remember with deep respect and gratitude Prof. Pushpak Bhattacharyya, whose mentorship, encouragement, and invaluable contributions continue to inspire this work.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. 2023. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629.

Shayantan Banerjee, Pratik Jalan, Rudrapratap Chinhara, Chinmayee Kalle, Pramod P Wangikar, and Kshitij Jadhav. 2026. Ontology-guided harmonization enables unified discovery of public metabolomics studies within and across repositories. *bioRxiv*.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, and 1 others. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15016–15027.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Erdi Çallı, Ecem Sogancioglu, Bram Van Ginneken, Kicky G Van Leeuwen, and Keelin Murphy. 2021. Deep learning for chest x-ray analysis: A survey. *Medical image analysis*, 72:102125.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, and 1 others. 2023. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pages 7480–7512. PMLR.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, and 1 others. 2023. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10995–11005.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan T Nguyen, Joanna S Chong, Fang Ji, Nathanael R Tong, Christopher L Chen, and Juan H Zhou. 2024. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37:86048–86073.

Sijing Feng, Damian Azzollini, Ji Soo Kim, Cheng-Kai Jin, Simon P Gordon, Jason Yeoh, Eve Kim, Mina Han, Andrew Lee, Aakash Patel, and 1 others. 2021. Curation of the candid-ptx dataset with free-text reports. *Radiology: Artificial Intelligence*, 3(6):e210136.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and 1 others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and 1 others. 2018. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *Preprint*, arXiv:2106.14463.

Pratik Jalan, Mukul Joshi, Akhilesh Magotra, and Kshitij Jadhav. 2026. MedBench: Deliberative evaluation of medical language models. In *BioNLP 2026*, pages 981–991, San Diego, California. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Charles Jones, Daniel C Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. 2024. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, 6(2):138–146.

Anas Anwarul Haq Khan, Tanik Saikh, Arpan Phukan, and Asif Ekbal. 2024. Hope 'the paragraph guy'explains the rest: Introducing mesum, the meme summarizer. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6654–6668.

Anas Anwarul Haq Khan, Utkarsh Verma, and Ganesh Ramakrishnan. 2025. Early exit and multi stage knowledge distillation in vlms for video summarization. *arXiv preprint arXiv:2504.21831*.

Eram Anwarul Khan and Anas Anwarul Haq Khan. 2024. Sumotosima: A framework and dataset for classifying and summarizing otoscopic images. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 1–11.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. 2022. Exploring plain vision transformer backbones for object detection. In *European conference on computer vision*, pages 280–296. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

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, and 1 others. 2022. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429.

Hoang C Nguyen, Tung T Le, Hieu H Pham, and Ha Q Nguyen. 2021. Vindr-ribcxr: A benchmark dataset for automatic segmentation and labeling of individual ribs on chest x-rays. *arXiv preprint arXiv:2107.01327*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and 1 others. 2024. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 374–390.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, and 1 others. 2025. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Eduardo P Reis, Joselisa PQ De Paiva, Maria CB Da Silva, Guilherme AS Ribeiro, Victor F Paiva, Lucas Bulgarelli, Henrique MH Lee, Paulo V Santos, Vanessa M Brito, Lucas TW Amaral, and 1 others. 2022. Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, and 1 others. 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Xiao Song, Jiafan Liu, Yan Liu, Yun Li, Wenbin Lei, and Ruxin Wang. 2024. Rethinking radiology report generation via causal inspired counterfactual augmentation. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10.

Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 2024. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 6883–6893.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. 2022. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406.

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

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, and 1 others. 2021. Chest imagenome dataset. *Physio Net*.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. 2018. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zefan Yang, Ge Wang, James Hendler, Mannudeep K Kalra, and Pingkun Yan. 2025b. X-win: Building chest radiograph world model via predictive sensing. *arXiv preprint arXiv:2511.14918*.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, and 1 others. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Yang Yue, Yulin Wang, Chenxin Tao, Pan Liu, Shiji Song, and Gao Huang. 2025. Chexworld: Exploring image world modeling for radiograph representation learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20778–20788.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. 2023. Advancing radiograph representation learning with masked record modeling. *arXiv preprint arXiv:2301.13155*.

10 Evaluation on Downstream Tasks

10.1 Image Classification

10.1.1 Datasets and Implementation Details

We evaluate on VinDr-CXR (Nguyen et al., 2022) and RSNA-Pneumonia [3]. For VinDr-CXR (diverse thoracic findings), we use six findings with a 15,000/3,000 train/test split by subject. For RSNA-Pneumonia (26,684 images; acute lung opacities), we adopt a 60/20/20 train/validation/test split by subject. All splits are performed by subject identifier. Images are converted to single-channel grayscale, center-cropped, and resized. Training augmentations include random horizontal flipping, affine transformations, random cropping, color jittering, and additive Gaussian noise. Classification is performed via linear probing, where a sigmoid-activated linear classifier is trained on frozen 768-dimensional global embeddings using binary cross-entropy loss. Optimization uses AdamW with a learning rate of 5×10^{-5} and cosine scheduling. Experiments are conducted on 3 NVIDIA A6000 GPUs with batch size 32 per GPU (total 96) for 100 epochs. Results are reported as mean of AUPRC over 5-fold cross-validation.

10.1.2 Results analysis

Table 6 reports results on VinDr-CXR and RSNA-Pneumonia. RadJEPA outperforms all image-only and vision-language baselines across both datasets. On VinDr-CXR, it achieves the highest mean AUPRC (Agg.) of **55.2**, surpassing RAD-DINO (52.8), I-JEPA (47.0), and MRM (51.3). On RSNA-Pneumonia, it attains the best AUPRC (**72.7**) and AUROC (**89.2**). Notably, RadJEPA yields strong

[3] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge

Model	VinDr-CXR (AUPRC)							RSNA	
	LO	CM	PL-T	AE	PF	PE	Agg.	AUPRC	AUC
CLIP@224	9.7 ± 0.4	42.6 ± 0.2	18.8 ± 0.4	30.0 ± 0.5	24.1 ± 0.4	21.8 ± 0.4	23.8	60.1 ± 2.0	83.7 ± 0.7
CLIP@336	9.1 ± 0.1	46.1 ± 0.2	18.5 ± 0.2	29.0 ± 0.3	22.8 ± 0.3	18.6 ± 0.3	23.4	60.0 ± 1.7	84.2 ± 0.4
BioViL-T	12.7 ± 0.1	51.4 ± 0.5	24.6 ± 0.2	22.3 ± 0.1	30.5 ± 0.1	52.2 ± 0.4	32.4	66.8 ± 1.5	86.9 ± 0.5
BiomedCLIP	10.0 ± 0.3	58.5 ± 0.8	24.4 ± 0.5	36.2 ± 0.2	32.0 ± 0.6	54.1 ± 0.6	35.9	68.4 ± 1.7	87.5 ± 0.4
CheXzero	11.1 ± 0.6	74.4 ± 0.2	25.1 ± 0.3	42.9 ± 0.2	33.1 ± 0.4	60.2 ± 0.5	40.0	68.9 ± 1.9	87.9 ± 0.4
MRM	12.2 ± 0.3	**79.7 ± 0.4**	35.8 ± 0.3	47.7 ± 0.2	47.1 ± 0.4	77.2 ± 0.3	51.3	71.4 ± 1.5	89.0 ± 0.5
DINO-v2	12.6 ± 0.4	67.3 ± 0.5	31.7 ± 0.4	38.9 ± 0.3	53.5 ± 0.5	69.8 ± 0.4	45.6	67.6 ± 1.8	85.9 ± 0.6
DINO-v3	11.8 ± 0.5	66.9 ± 0.4	29.7 ± 0.3	35.4 ± 0.4	52.0 ± 0.5	66.5 ± 0.5	43.7	65.2 ± 1.6	82.8 ± 0.4
RAD-DINO (Pérez-García et al., 2025)	14.2 ± 0.3	68.4 ± 0.4	35.8 ± 0.5	48.5 ± 0.3	57.2 ± 0.4	74.9 ± 0.4	49.8	69.1 ± 1.7	86.8 ± 0.5
CheXWorld (Yue et al., 2025)	14.5 ± 0.3	69.1 ± 0.4	36.1 ± 0.5	44.1 ± 0.3	58.7 ± 0.4	**76.9 ± 0.4**	49.9	70.4 ± 1.7	88.0 ± 0.5
I-JEPA	13.6 ± 0.3	68.3 ± 0.6	32.8 ± 0.7	41.9 ± 0.5	55.0 ± 0.5	70.5 ± 0.6	47.0	68.2 ± 2.0	86.4 ± 0.8
RadJEPA	**19.2 ± 0.2**	72.6 ± 0.3	**42.2 ± 0.4**	**58.6 ± 0.2**	**63.9 ± 0.3**	74.4 ± 0.4	**55.2**	**72.7 ± 1.5**	**89.2 ± 0.4**

Table 6: Image classification performance with linear probing on VinDr-CXR and RSNA-Pneumonia. AUPRC denotes area under the precision-recall curve and AUC denotes area under the ROC curve. Agg. denotes mean AUPRC over the six VinDr-CXR classes: LO (Lung Opacity), CM (Cardiomegaly), PL-T (Pleural Thickening), AE (Aortic Enlargement), PF (Pulmonary Fibrosis), and PE (Pleural Effusion). The RAD-DINO and CheXWorld rows are reproduced from publicly released checkpoints under our evaluation pipeline. VinDr-CXR uses 15,000/3,000 train/test splits; RSNA results are reported on the test set (5,337 images). Numbers in bold are best per column. Results are averaged over five random seeds; the full statistical protocol is in Section 12.

gains on subtle findings (PL-T, AE, PF), improving PF by **+4.5** over RAD-DINO and **+8.9** over I-JEPA. Compared to I-JEPA, consistent improvements are observed across all VinDr-CXR classes. Despite using a ViT-B/14 backbone (as shown in Table 2) with only **86M** parameters (vs. 1.1B for DINO-v2 and 0.6B for I-JEPA), RadJEPA achieves superior downstream performance.

10.2 Semantic Segmentation

10.2.1 Datasets and Implementation Details

Lung and Lung Zone Segmentation masks are obtained using Chest ImaGenome bounding boxes for six regions (left/right upper, middle, lower) (Wu et al., 2021). This yields 1,138 images from 1,138 subjects. For **Rib Segmentation**, we use the VinDR-RibCXR dataset (Nguyen et al., 2021), which provides expert annotations for 20 ribs (L1–L10, R1–R10) across 245 subjects, following the official train/test splits. We evaluate RadJEPA for semantic segmentation using frozen backbone encoders with task-specific decoder heads. Experiments are conducted on a single node with 4 NVIDIA A6000 GPUs (batch size 20 per GPU; total 80). Models are optimized with Adam (5×10^{-4}) and cosine scheduling for 100 epochs. Images are center-cropped and resized; training augmentations include random horizontal flipping (except for left-right lung and lung zone tasks), random affine and elastic transformations, brightness/contrast jittering, and random gamma adjustments. Image intensities are normalized using statistics computed from all MIMIC-CXR images (Johnson et al., 2019). Data are split 70/15/15 by subject, and evaluation is performed on the test set using the checkpoint with lowest validation loss (rib segmentation follows the official 196/49 train/test split). Results are reported as mean of Dice score over 5-fold cross-validation.

10.2.2 Results analysis

Table 7 reports semantic segmentation performance. Across all segmentation tasks, *RadJEPA consistently outperforms competing image-only and vision-language pretrained backbones* under identical decoder configurations. With a UPerNet decoder, RadJEPA achieves the highest Dice scores on lung (**98.3**), lung zone (**93.7**), and rib (**89.6**) segmentation, surpassing RAD-DINO and I-JEPA across all three tasks. Notably, RadJEPA shows the largest gains on structurally complex targets such as lung zones and ribs, improving rib segmentation by **+4.8 Dice** over RAD-DINO and **+6.4 Dice** over I-JEPA. Importantly, all predictive and contrastive baselines share the same decoder.

11 Model Checkpoints

Table 8 lists the public Hugging Face checkpoint used for every backbone, decoder, and image encoder evaluated in this paper.

12 Statistical Evaluation Protocol

Motivation: A common concern in deep learning evaluation is the choice of statistical significance analysis under limited experimental runs. Classical paired significance tests such as paired

Encoder	Decoder	# Features	# Params	Lungs	Lung zones	Ribs
NN-UNet (Isensee et al., 2018)	Unet	–	17.9 M	98.0 (1.1)	92.6 (10.2)	86.2 (2.8)
EfficientNet-B6 (Tan and Le, 2019)	Unet	–	45.9 M	98.3 (1.1)	92.7 (10.1)	88.9 (2.6)
BioViL-T (Bannur et al., 2023)	Linear	2048	2049	83.2 (3.2)	69.4 (9.0)	59.1 (4.7)
BiomedCLIP (Zhang et al., 2023)	Linear	768	769	90.4 (2.6)	76.2 (10.2)	67.4 (4.5)
CheXzero (Tiu et al., 2022)	Linear	768	769	84.0 (3.4)	68.3 (9.1)	62.0 (3.3)
RAD-DINO (Pérez-García et al., 2025)	Linear	768	769	95.9 (1.5)	85.7 (9.8)	73.4 (3.6)
DINO-v2	Linear	768	769	90.5 (2.7)	77.9 (9.8)	70.2 (4.2)
DINO-v3	Linear	768	769	89.1 (1.4)	77.4 (9.3)	69.7 (3.1)
I-JEPA	Linear	768	769	91.1 (2.4)	81.6 (9.4)	71.8 (4.0)
RadJEPA	Linear	768	769	96.2 (1.4)	86.5 (9.2)	73.9 (3.4)
DINO-v2	ViTDet	4×768	24.8 M	96.1 (1.4)	88.3 (9.3)	79.4 (3.1)
DINO-v2	UPerNet	4×768	39.3 M	96.4 (1.3)	88.7 (9.1)	80.1 (2.9)
DINO-v3	ViTDet	4×768	24.8 M	96.0 (1.2)	88.4 (9.0)	79.6 (2.6)
DINO-v3	UPerNet	4×768	39.3 M	95.9 (0.9)	89.2 (8.8)	80.3 (2.2)
RAD-DINO (Pérez-García et al., 2025)	ViTDet	4×768	24.8 M	97.8 (1.2)	90.8 (9.9)	83.7 (2.8)
RAD-DINO (Pérez-García et al., 2025)	UPerNet	4×768	39.3 M	98.1 (1.1)	91.5 (9.8)	84.8 (2.7)
CheXWorld (Yue et al., 2025)	ViTDet	4×768	24.8 M	97.6 (1.3)	90.3 (10.0)	83.1 (2.9)
CheXWorld (Yue et al., 2025)	UPerNet	4×768	39.3 M	97.9 (1.2)	91.0 (9.9)	84.2 (2.8)
I-JEPA	ViTDet	4×768	24.8 M	96.6 (1.3)	90.2 (9.7)	82.4 (2.9)
I-JEPA	UPerNet	4×768	39.3 M	96.9 (1.2)	91.0 (9.5)	83.2 (2.8)
RadJEPA	ViTDet	4×768	24.8 M	98.0 (1.1)	92.9 (9.0)	88.0 (2.3)
RadJEPA	UPerNet	4×768	39.3 M	**98.3 (1.0)**	**93.7 (8.8)**	**89.6 (2.1)**

Table 7: Semantic segmentation results obtained with linear, ViTDet (Li et al., 2022), and UPerNet (Xiao et al., 2018) decoders on frozen backbone encoders. Dice scores are reported as mean (standard deviation in parentheses). Lungs denotes left/right lung segmentation, Lung zones the six anatomical lung regions, and Ribs the 20 individual ribs. End-to-end U-Net models indicate an upper-bound reference. RAD-DINO and CheXWorld rows are reproduced from publicly released checkpoints under our evaluation pipeline.

Model	Source
Vision-Language backbones (Section 5.2)	
MedLLaVA	HF
Phi-4	HF
Qwen-2.5	HF
BLIP-2	HF
Language decoders	
Vicuna-7B (v1.5)	HF
Mistral-7B	HF
Qwen2.5-7B	HF
OPT-2.7B	HF
Phi-4-mini	HF
Image encoders evaluated in Tables 3 and 4	
DINO-v2	HF
I-JEPA	HF
RAD-DINO	HF
CheXWorld	(Yue et al., 2025)
Evaluation models	
CheXbert	(Smit et al., 2020)
RadGraph	(Jain et al., 2021)
GREEN	HF

Table 8: Hugging Face checkpoints for every model referenced in the main paper. "HF" is a hyperlink to the repository.

t-tests assume approximate Gaussianity of the sampled differences, while non-parametric alternatives such as the Wilcoxon signed-rank test require a sufficient number of samples to reliably establish significance. In our setting, each experiment requires large-scale training on chest X-ray images and multiple VLM backbones, making a substantially larger number of runs computationally expensive and practically infeasible. Therefore, instead of relying on potentially unstable significance tests with only five runs, we adopt the same statistical evaluation strategy used in RAD-DINO for fair and consistent comparison.

Report Generation. For radiology report generation (cf. Tables 3 and 4), following RAD-DINO, we report the median and 95% confidence intervals obtained from 500 bootstrap samples. Across most evaluation settings, RadJEPA achieves higher median performance along with narrower confidence intervals compared to RAD-DINO, suggesting improved robustness and reduced evaluation uncertainty.

Classification and Segmentation. For disease classification and semantic segmentation tasks (cf. Tables 6 and 7), we report the mean and standard deviation across five runs with different random seeds. Across most settings, RadJEPA achieves higher mean performance while also exhibiting

lower standard deviation compared to RAD-DINO, indicating more stable and consistent representations.

13 CheXWorld Overview

CheXWorld (Yue et al., 2025) is a recently proposed image-only chest X-ray representation learning framework based on an architecture similar to I-JEPA. The model uses a ViT-B/14 backbone with approximately 86M parameters and is pretrained on around 448K frontal chest X-ray images collected from multiple datasets at a resolution of 224×224. Unlike RadJEPA, CheXWorld does not utilize lateral images during pretraining and is trained on approximately 31% fewer images. The pretrained weights are publicly available, which enabled us to reproduce and compare the model across all downstream tasks considered in our work.

We observed trends similar to those reported in the original CVPR 2025 paper: CheXWorld generally performs competitively with RAD-DINO, but in most settings slightly underperforms RAD-DINO and consistently remains below RadJEPA. One possible reason is the comparatively smaller pretraining dataset and lower training compute budget, which may impose limitations on optimization choices such as batch size and large-scale training stability. Furthermore, the original CheXWorld work did not evaluate report generation performance. In fact, several recent chest X-ray representation learning works, including X-Win (Yang et al., 2025b; Jalan et al., 2026; Banerjee et al., 2026), primarily focus on classification and segmentation benchmarks without extensive report generation analysis.

Although the primary objective of our work is radiology report generation, we additionally perform extensive classification and segmentation experiments to provide a more robust and comprehensive comparison across existing representation learning frameworks. Following prior radiology benchmarking practices (including RAD-DINO from Microsoft), we primarily report AUPRC instead of AUROC due to the severe class imbalance and non-mutually exclusive nature of chest X-ray disease labels. Nevertheless, for completeness, we also report reproduced AUROC comparisons for CheXWorld, RAD-DINO, and RadJEPA in Table 9.

Model	**VinDr-CXR**	**RSNA**
CheXWorld	94.23 ± 0.18	73.21 ± 0.34
RAD-DINO	94.92 ± 0.12	75.64 ± 0.63
RadJEPA	$\mathbf{95.36 \pm 0.15}$	$\mathbf{79.12 \pm 0.21}$

Table 9: AUROC comparison reproduced over 5 different random seeds for completeness.

Comparison	Score (QWK)	Acceptable (κ)
H1 vs H2	0.91	0.83
H1 vs LLM Master vote	0.88	0.86
H2 vs LLM Master vote	0.81	0.77
H1 vs MedGemma-4B	0.83	0.75
H1 vs Qwen-3-4B	0.85	0.79
H1 vs Gemma-4B	0.84	0.75
H2 vs MedGemma-4B	0.75	0.65
H2 vs Qwen-3-4B	0.80	0.70
H2 vs Gemma-4B	0.79	0.66

Table 10: Pairwise agreement on the 0–5 clinical score (quadratic-weighted Cohen's κ, QWK) and on the binary Acceptable flag (Cohen's κ). All human-LLM comparisons fall in the substantial-to-almost-perfect range, with the LLM Master vote tracking the human-human reference most closely.

14 Human Evaluation

14.1 Inter-Annotator Agreement

Setup: Two clinical annotators (H1, H2) and three LLM judges (MedGemma-4B, Qwen-3-4B, Gemma-4B) independently graded a stratified sample of $N = 500$ RadJEPA-generated reports drawn equally from the IU-Xray and MIMIC-CXR test sets (250 reports each). Each report was rated along two axes: an overall clinical score $s \in \{0, \ldots, 5\}$ (5 = clinically equivalent to the ground-truth report; 0 = unsafe or dangerously contradicted) and a binary Acceptable flag set to 1 iff $s \geq 4$. The three LLM judgements were combined into a single Master vote by per-metric majority.

Agreement statistics: We report Cohen's quadratic-weighted κ for the ordinal score, Cohen's unweighted κ for the binary flag, Fleiss' κ (Fleiss, 1971) and Krippendorff's ordinal α for multi-rater comparisons, and ICC(2,1) (Shrout and Fleiss, 1979) for absolute-agreement reliability of the ordinal score. Following Landis and Koch (1977), values of 0.81–1.00 are interpreted as almost above average, 0.61–0.80 as substantial, and 0.41–0.60 as moderate.

Summary: The two human annotators show almost-perfect agreement on both the ordinal score

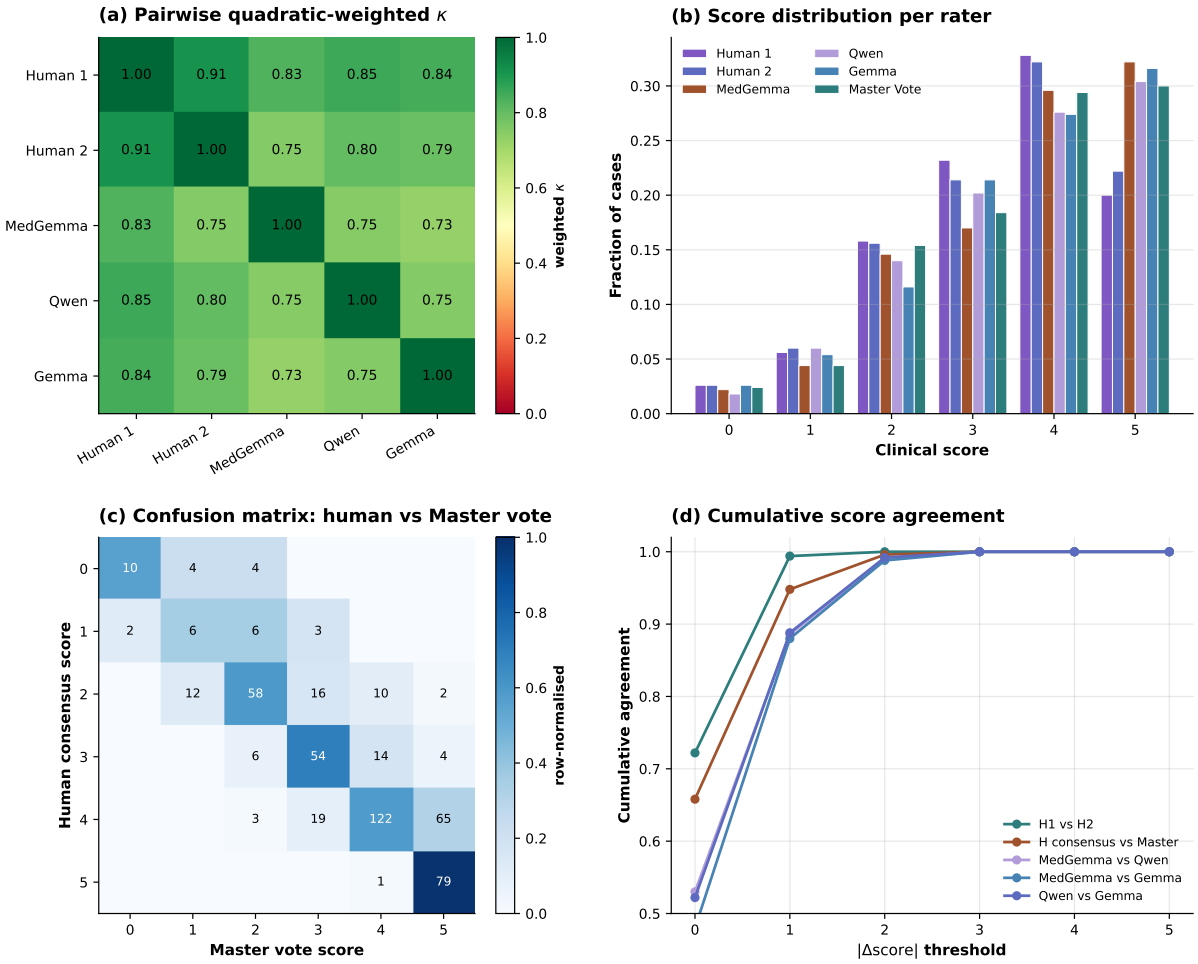


Figure 4: Inter-annotator agreement suite on the $N = 500$ paired evaluation sample. (a) Pairwise quadratic-weighted κ across the two human annotators and the three LLM judges. The two humans agree at κ=0.91; every human-LLM pair lies in the substantial-to-almost-perfect band. (b) Per-rater score distributions: humans and LLMs put most of their probability mass on the same $\{3, 4, 5\}$ region of the scale. (c) Confusion matrix between the rounded human-consensus score and the LLM Master vote (row-normalised), showing concentration on and immediately adjacent to the diagonal. (d) Cumulative agreement: for every plotted rater pair, more than 90% of cases agree within ± 1 score point.

Multi-rater statistic	Value
Krippendorff's α (ordinal, 5 raters, score)	0.77
ICC(2,1) (5 raters, score)	0.80
Fleiss' κ (5 raters, Acceptable)	0.71
Within ± 1 score agreement (H vs Master vote)	94.2%

Table 11: Multi-rater reliability across the two human annotators and the three LLM judges. All coefficients fall in the substantial range by the Landis–Koch scale (Landis and Koch, 1977).

(QWK = **0.91**) and the acceptability flag (κ = 0.83), establishing a high-reliability reference. The LLM majority vote tracks this reference at QWK = **0.84** and κ = 0.81 on average across the two humans, and the multi-rater coefficients (Krippendorff's α = 0.77, ICC = 0.80, Fleiss' κ = 0.71) confirm substantial agreement across the full panel. These results support the use of the majority-vote LLM protocol as a scalable and clinically meaningful complement to human evaluation. The same conclusions are visible in Figure 4, which shows the pairwise weighted-κ matrix, per-rater score distributions, the human-consensus to vote confusion matrix, and cumulative agreement curves.

Qualitative Error Analysis: Aggregate report-generation metrics provide limited insight into the specific clinical statements responsible for individual generation errors. We therefore complement the quantitative evaluation with a case-level qualitative analysis of representative reports from the 500-report human-reviewed sample. Table 12 organizes these observations into four clinically relevant cate-

gories: false-positive findings, hallucinated clinical content, missed subtle findings, and low-frequency disease categories. We show only the clinically relevant portions of the reference and generated reports to make each discrepancy directly interpretable. These examples are intended to characterize representative generation behaviours rather than estimate the prevalence of individual error types.

15 GREEN Score Evaluation

In addition to the lexical and clinical-NLP metrics reported in the main paper, we also evaluate the generated reports with the GREEN score (Ostmeier et al., 2024), a language-model-based clinical-correctness metric that scores a generated report against the reference using a fine-tuned LLM judge. The metric is bounded in $[0, 1]$, with higher values indicating better clinical agreement.

Table 13 shows that RadJEPA achieves the highest GREEN score on both datasets, with absolute margins of **+0.018** on MIMIC-CXR and **+0.017** on IU-Xray over RAD-DINO, the next-best baseline. The relative ranking of the remaining encoders is consistent with the ranking observed under the lexical and clinical-NLP metrics in Table 3 and under the inter-annotator analysis in Appendix 14.1, which provides a third, independent line of evidence that the gains reported in the main paper reflect genuine improvements in clinical report quality rather than artefacts of any single metric family.

16 Per-Metric Model Ranking

Figure 5 ranks every encoder evaluated in Table 3 on the three MIMIC report-quality metrics. Models are sorted within each panel from best to worst on the corresponding metric; the marker area encodes the backbone parameter count on a log scale, so the visualisation also exposes the efficiency profile of each encoder. RadJEPA sits at the top of every panel despite a 86M backbone and a 224×224 input resolution, while the next-best models, RAD-DINO$_{\text{control}}$ and RAD-DINO, require approximately $5.4\times$ more input pixels per image. DINO-v2 occupies the last row of every panel even though it carries 1.1B parameters, confirming that the gain reported in the main table is not explained by either capacity alone.

17 Visual Analysis

17.1 Report Generation Consistency

The report generation results show a consistent pattern across both datasets. RadJEPA is strongest on every plotted metric, and the gains are not explained by token count alone, since DINO-v2, I-JEPA, RAD-DINO, RadJEPA$_{\text{control}}$, and RadJEPA all use 1369 visual tokens. Narrative metrics and clinical F1 improve together, suggesting that the encoder improves not only lexical overlap but also disease-level content. Confidence intervals are narrow for ROUGE-L, BLEU-4, and RG$_{\text{ER}}$, while Macro-F1 is wider, especially on IU-Xray. (cf. Figure 6, Table 14)

17.2 Portability Across VLM Backbones

Replacing the vision encoder across four VLM backbones provides a stronger test than a single decoder setting. RadJEPA leads in nearly all cells, including the weaker Phi-4 setting. Gains are most consistent for ROUGE, BLEU, and RG$_{\text{ER}}$, while Macro-F1 is more decoder-sensitive, with RAD-DINO narrowly ahead for Phi-4. Thus, the encoder consistently improves report language and entity recovery, while disease-label extraction still depends on decoder use of the visual features. (Figure 7, Table 16)

17.3 Classification Transfer Profile

The classification profile is label-dependent rather than uniformly superior across diseases. RadJEPA is strongest for lung opacity, pleural thickening, aortic enlargement, and pulmonary fibrosis, while MRM remains strongest for cardiomegaly and RAD-DINO for pleural effusion. This highlights the distinction between broad transfer strength and label-specific performance. RadJEPA also shows low run-to-run variation on VinDr-CXR, with standard deviations of 0.2–0.4 across the six displayed classes. (cf. Figure 8, Table 15)

Metric	Mean RadJEPA − RAD-DINO
ROUGE-L	+1.35
BLEU-4	+0.71
RG$_{\text{ER}}$	+0.79
Macro-F1-14	+0.89
All metrics	+0.93

Table 16: Mean performance margins of RadJEPA over RAD-DINO across report-generation metrics, evaluated under the same VLM and dataset settings. Positive values indicate improvements with RadJEPA, with an average gain of +0.93 across all metrics.

Category	Ground Truth	Generated Report	Human Interpretation
False positive	No focal air-space opacity. No pleural effusion or pneumothorax.	Mild bibasilar atelectatic change with a **small left pleural effusion**.	A clinically plausible pleural abnormality is reported despite its absence in the reference.
	Cardiomediastinal silhouette is within normal limits. No pulmonary vascular congestion or focal consolidation.	Mild cardiomegaly with **interstitial pulmonary edema**.	The generation introduces both cardiac enlargement and edema in an otherwise negative reference report.
	Lungs are clear without focal consolidation. No acute cardiopulmonary abnormality.	There is a **small right lower-lobe opacity**, likely representing early air-space disease.	A focal pulmonary finding is asserted despite no corresponding abnormality in the reference.
Hallucinated finding	No focal pulmonary opacity, pleural effusion, or pneumothorax.	No focal consolidation. A **right internal jugular central venous catheter terminates in the lower SVC**.	The generation introduces a support device that is not present in the reference information.
	Stable cardiomediastinal contours with mild bibasilar atelectatic change.	Stable postoperative appearance following **median sternotomy and CABG**, with mild bibasilar atelectasis.	The pulmonary finding is compatible with the reference, but unsupported surgical history is added.
	Mild left basilar opacity, likely atelectatic. No pleural effusion.	Left basilar atelectatic opacity is present. **Postoperative changes following left upper lobectomy are again noted**.	A valid image finding is accompanied by unsupported procedure-related clinical context.
Missed subtle finding	Heart size is moderately enlarged. Mild left costophrenic angle blunting is present. Impression includes a **small left pleural effusion versus pleural thickening**, with streaky and patchy bibasilar opacities.	Moderate cardiomegaly and left costophrenic angle blunting are described, together with bibasilar streaky opacities; however, the report concludes, "There is **no pleural effusion or pneumothorax**."	Several accompanying abnormalities are retained, but the small left pleural abnormality is explicitly negated.
	The heart size and pulmonary vascularity are within normal limits. A **right pleural effusion is present and appears increased**; the impression again identifies a small increased right pleural effusion.	The heart and mediastinum are described as unremarkable, followed by: "There is no focal consolidation, **effusion**, or pneumothorax."	The right pleural effusion is omitted and directly contradicted by an explicit negative statement.
	Stable elevation of the right hemidiaphragm with questionable increased right basilar air-space opacity. The impression additionally reports a **small right pleural effusion**.	The generated report recovers the elevated right hemidiaphragm and right basilar air-space opacity, but states: "No pneumothorax or **pleural effusion**."	Neighbouring radiographic findings are preserved while the subtle pleural abnormality is missed.
Less common finding	Small-volume **pneumomediastinum** with mild associated subcutaneous emphysema. No pneumothorax.	Subcutaneous emphysema is present along the chest wall. No pneumothorax or pleural effusion.	A related abnormality is recognized, while the less commonly represented pneumomediastinum finding is omitted.
	Acute mildly displaced **right posterior sixth-rib fracture**. No pneumothorax.	No focal pulmonary abnormality, pleural effusion, or pneumothorax. **No acute osseous abnormality**.	The focal osseous abnormality is missed and additionally contradicted by the generated report.
	Extensive **subcutaneous emphysema** over the right chest wall and lower neck without pneumothorax.	Right chest-wall and lower-neck **subcutaneous emphysema** is present. No pneumothorax.	The example illustrates successful recovery of a less commonly represented finding rather than an error.

Table 12: Examples from false-positive findings, hallucinated clinical content, missed subtle abnormalities, and behaviour on low-frequency disease categories. Reference and generated reports are shortened to the clinically relevant content, and **bold text** marks the key finding responsible for the agreement or discrepancy.

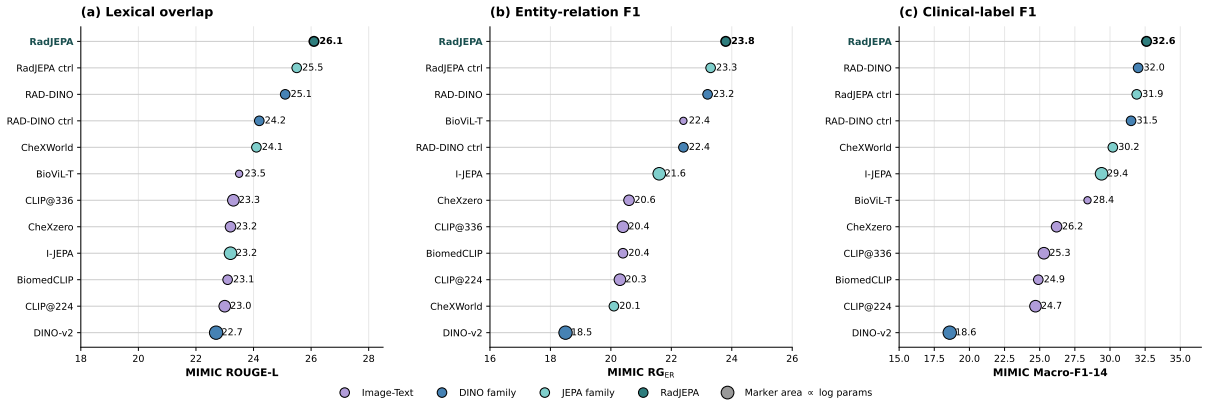


Figure 5: Per-metric ranking of frozen encoders on the MIMIC-CXR report-generation benchmark (cf. Table 3). Within each panel, encoders are sorted from best to worst on the corresponding metric. Marker area encodes the backbone parameter count on a log scale; colour encodes the model family. RadJEPA sits at the top of every panel.

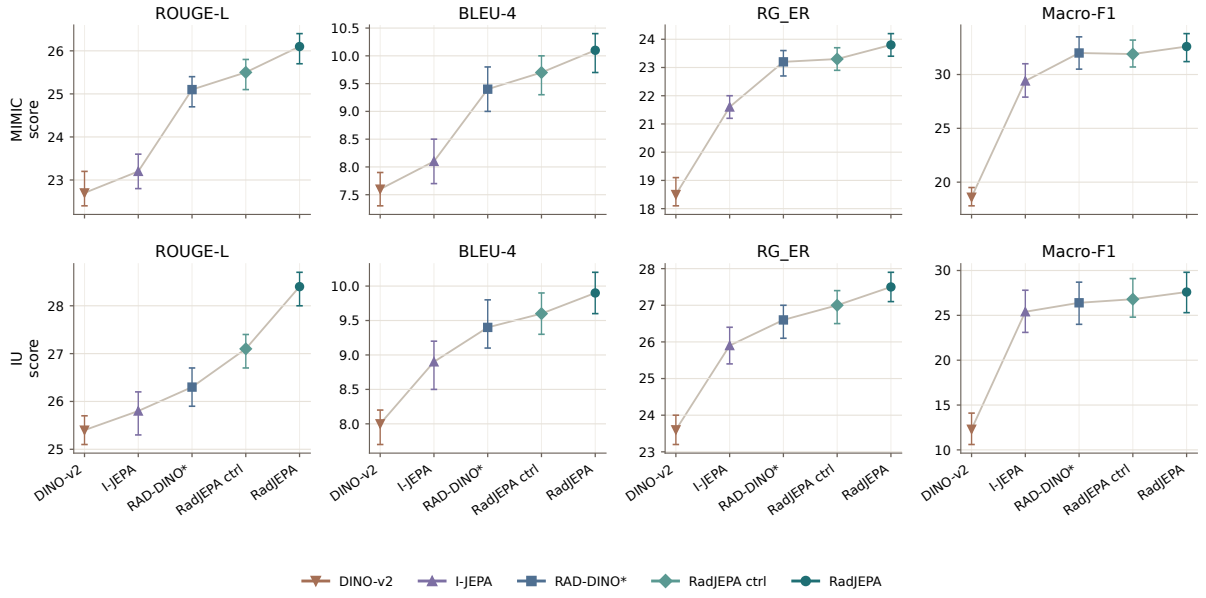


Figure 6: Report generation medians with 95% confidence intervals. Each row is one dataset and each column is one metric.

Model	MIMIC-CXR	IU-Xray
CheXWorld	0.471	0.498
RAD-DINO	0.479	0.508
I-JEPA	0.455	0.482
RadJEPA$_{\text{control}}$	0.457	0.486
RadJEPA	**0.497**	**0.525**

Table 13: GREEN score (Ostmeier et al., 2024) comparison across image-only encoders on MIMIC-CXR and IU-Xray. All numbers are computed under our evaluation pipeline using the publicly released GREEN model. Higher is better; bold marks the best value per column.

Model	Input	Tokens	MIMIC	IU	Mean CI width
DINO-v2	518×518	1369	16.85	17.32	1.19
I-JEPA	224×224	1369	20.57	21.50	1.60
RAD-DINO	518×518	1369	22.43	22.18	1.56
RadJEPA$_{\text{control}}$	224×224	1369	22.60	22.62	1.40
RadJEPA	224×224	1369	**23.15**	**23.35**	1.42

Table 14: Report generation summary. Mean CI width averages the reported interval widths over both datasets and all four metrics.

17.4 Segmentation Decoder Scaling

The segmentation results show how the representation behaves as the decoder becomes stronger. With a linear decoder, RadJEPA is only slightly ahead of RAD-DINO on mean Dice. With ViT-Det and UPerNet, the separation becomes much clearer, especially for lung zones and ribs. Lungs are already close to saturation for most strong models, so the more informative targets are the

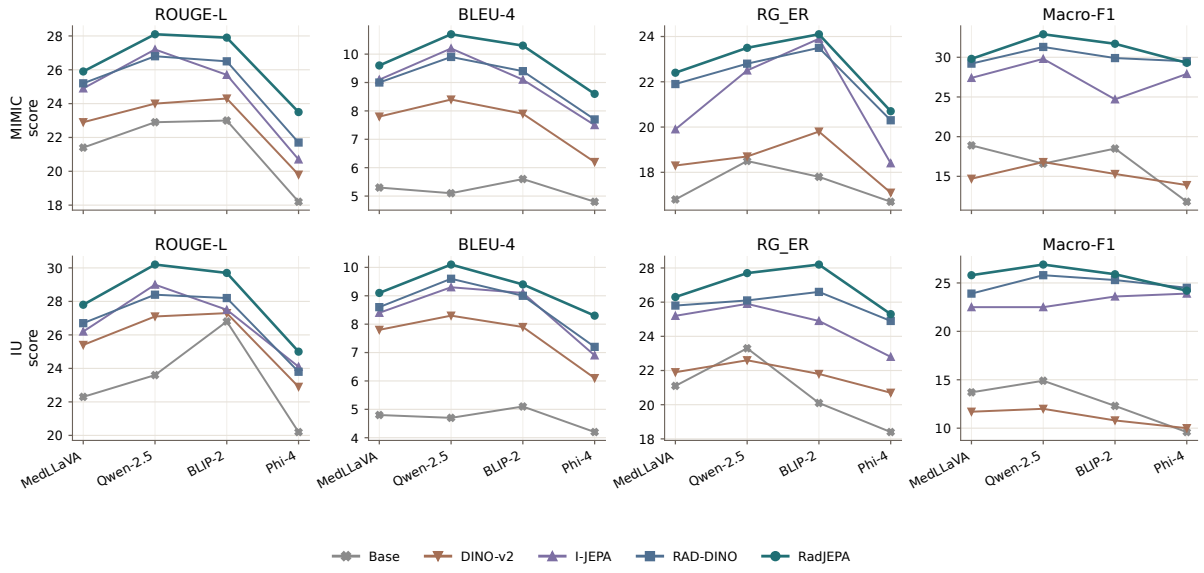


Figure 7: Encoder replacement results across VLM backbones. Lines track the same encoder across MedLLaVA, Qwen-2.5, BLIP-2, and Phi-4.

Target	Strongest non-RadJEPA	RadJEPA	Comparator	Margin
Lung opacity	RAD-DINO	19.2	14.9	+4.3
Cardiomegaly	MRM	72.6	79.7	-7.1
Pleural thickening	RAD-DINO	42.2	36.6	+5.6
Aortic enlargement	RAD-DINO	58.6	48.5	+10.1
Pulmonary fibrosis	RAD-DINO	63.9	59.4	+4.5
Pleural effusion	RAD-DINO	74.4	77.8	-3.4
RSNA AUPRC	MRM	72.7	71.4	+1.3
RSNA AUC	MRM	89.2	89.0	+0.2

Table 15: Target-wise comparison against the strongest non-RadJEPA model for each endpoint.

harder anatomical structures. RadJEPA-UPerNet also has lower reported standard deviations than RAD-DINO-UPerNet on lungs, lung zones, and ribs, which makes the gain look stable rather than noisy. (cf. Figure 9, Table 17, 18)

Model	Lin.	ViTDet	UPer.	SD	Δ
DINO-v2	79.5	87.9	88.4	4.4	+8.9
I-JEPA	81.5	89.7	90.4	4.5	+8.9
RAD-DINO	85.0	90.7	91.5	4.6	+6.5
RadJEPA	**85.5**	**93.0**	**93.9**	**4.0**	+8.3

Table 17: Frozen-encoder segmentation summary using mean Dice across lungs, lung zones, and ribs denotes the linear segmentation setting, ViTDet denotes the ViTDet decoder setting, and UPer. denotes the UPerNet decoder setting. SD reports the standard deviation associated with the UPerNet Dice scores, and Δ reports the absolute gain from the linear setting to the UPerNet setting. All values are reported on the 0–100 Dice scale, with higher mean Dice indicating better segmentation performance and lower SD indicating more consistent performance across the evaluated anatomical structures.

Anatomy	RadJEPA-UPerNet	RAD-DINO-UPerNet	EfficientNet-B6 U-Net	RadJEPA − RAD-DINO	RadJEPA SD
Lungs	98.3	98.0	98.3	+0.3	1.0
Lung zones	93.7	91.2	92.7	+2.5	8.8
Ribs	89.6	85.3	88.9	+4.3	2.1

Table 18: Anatomy-level UPerNet comparison. The larger separation appears on lung zones and ribs, where the task is less saturated than whole-lung segmentation.

Training choice	**RadJEPA (ours)**	**I-JEPA (Assran et al., 2023)**	**CheXWorld (Yue et al., 2025)**	**RAD-DINO (Pérez-García et al., 2025)**
Pretraining data	839,364 open-source CXRs from BRAX, CheXpert, MIMIC-CXR, ChestX-ray14, and PadChest.	ImageNet-1K ($\sim$ 1.28M natural images).	$\sim$ 0.5M frontal CXRs from MIMIC-CXR, CheXpert, and NIH ChestX-ray14.	882,775 CXRs in the public release; the paper checkpoint used a different corpus/configuration.
Views used	635,086 frontal and 204,278 lateral images.	Natural images; one crop per sample.	Frontal images only.	Frontal and lateral CXRs in the public release.
Initialization	Trained from scratch.	Trained from scratch.	Trained from scratch.	Continued from pretrained DINOv2 weights.
Image preprocessing	DICOM	RGB images; tensor conversion and ImageNet mean/std normalization.	X-ray loading and normalization implemented in the released data pipeline.	DICOM
Image augmentation	Random resized crop, scale 0.3–1.0; no flip, color distortion, or Gaussian blur.	Same released transform.	Random-resized crops plus brightness, contrast, gamma, and Gaussian-blur transformations for domain-transition modeling.	Random-resized multi-crop, horizontal flip, color jitter, grayscale, and Gaussian blur.
Optimizer	AdamW, $\beta = (0.9, 0.999)$; no decay on bias/1-D parameters.	AdamW, $\beta = (0.9, 0.999)$; no decay on bias/1-D parameters.	AdamW, $\beta = (0.9, 0.999)$.	AdamW, $\beta = (0.9, 0.999)$, with layer-wise decay and patch-embedding LR multiplier.
Learning-rate schedule	$2 \times 10^{-4} \rightarrow 10^{-3}$ over 40 warmup epochs, then cosine to 10^{-6}.	Same released ViT-H/14 config.	Linear warmup for 40 epochs to 2×10^{-4}, then cosine to 10^{-6}.	Base LR 10^{-3} in the CXR override, 5 warmup epochs, then cosine to 10^{-6}.
Weight decay	Cosine $0.04 \rightarrow 0.4$.	Cosine $0.04 \rightarrow 0.4$.	Fixed at 0.05 in the released command.	Cosine $0.04 \rightarrow 0.4$.
EMA / teacher schedule	Linear $0.996 \rightarrow 1.0$.	Linear $0.996 \rightarrow 1.0$.	Cosine $0.996 \rightarrow 1.0$.	Teacher momentum $0.992 \rightarrow 1.0$.
Epochs / updates	300 epochs.	300 epochs.	300 epochs.	100-epoch/iteration-based released configuration; public checkpoint selected at 35K iterations from a 100K-iteration run.
Numerical precision	bfloat16 autocast.	bfloat16 autocast.	bfloat16 autocast with automatic mixed precision.	fp16 mixed precision with FSDP.
Gradient clipping	None.	None in released code.	Global norm clipped at 1.0.	Global norm clipped at 3.0.
Random seed	0	0 in released training code.	0 in released command.	0 in released configuration.

Table 19: Comparison of pretraining configurations for RadJEPA and related self-supervised representation-learning methods. We compare training data, input views, initialization, augmentation, optimization, scheduling, and numerical settings.

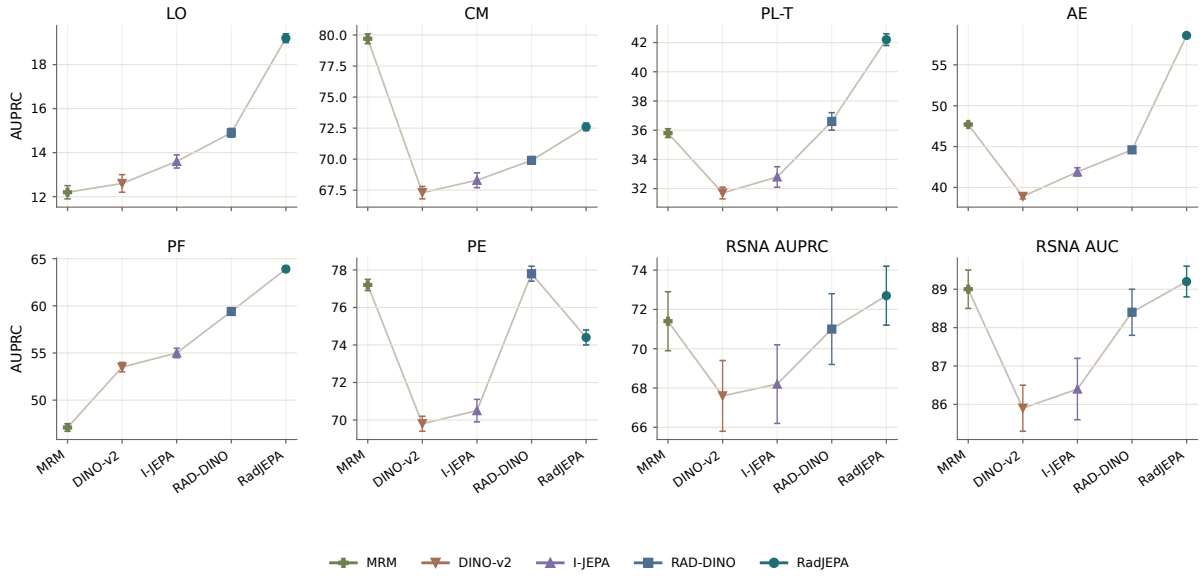


Figure 8: Classification performance with reported standard deviations. The six VinDr-CXR panels use AUPRC, followed by RSNA AUPRC and RSNA AUC.

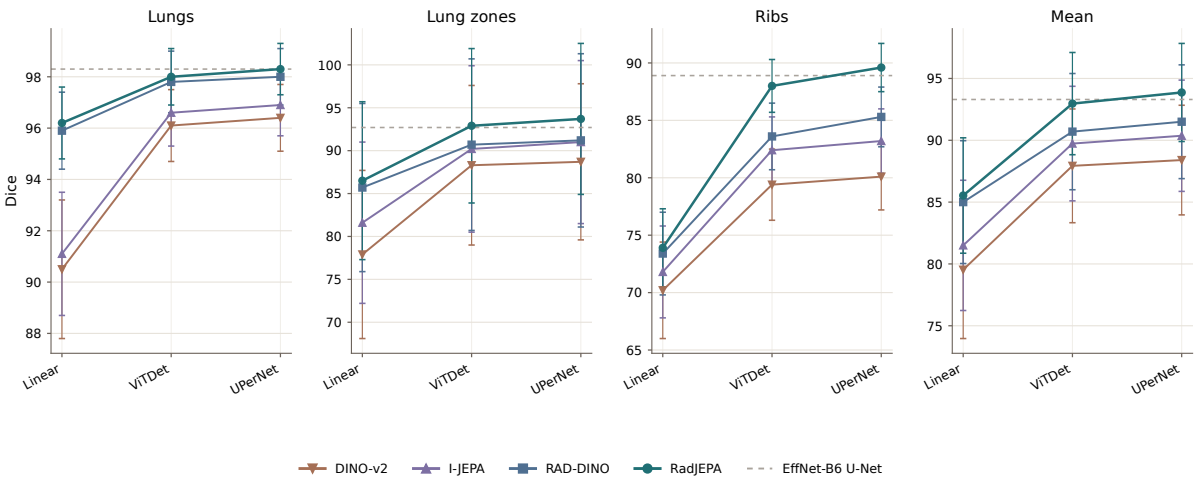


Figure 9: Dice scores across decoder families with reported standard deviations. The dashed reference line marks the EfficientNet-B6 U-Net value for each panel.