

What Do Medical Vision–Language Models Learn in Radiology? Transfer, Alignment, and Source-Proxy Leakage Under Distribution Shift

Ayoub Louaye Bouaziz¹ Lokmane Chebouba² Yassine Himeur³

¹LaTIM, Inserm, University of Western Brittany, Brest, France

²University of Constantine, Algeria

³University of Dubai, UAE

ayoublouaye.bouaziz@student.umc.edu.dz lokmane.chebouba@umc.edu.dz yhimeur@ud.ac.ae

Abstract

Medical vision–language models (VLMs) can appear reliable in-domain while failing when acquisition domain, paired supervision, or evaluation protocol changes. We study this failure mode as a representation-level blind spot relevant to epistemic intelligence, without claiming a formal estimator of epistemic uncertainty. Using NIH ChestXray14 and CheXpert, we first isolate source-only cross-dataset visual transfer from unsupervised domain-adaptation diagnostics. Using PadChest and OpenI, we then evaluate multimodal alignment under strict pair-index retrieval and quantify metadata-derived source-proxy information retained in frozen embeddings. Self-supervised visual initialization improves NIH-to-CheXpert transfer over supervised ImageNet initialization in matched ResNet-18 comparisons, whereas adversarial adaptation is useful only in a narrow regime and becomes unstable as adversarial pressure increases. Multimodal exact-pair retrieval remains low under external OpenI stress testing, and source-proxy information remains recoverable from learned representations. Qualitative nearest-neighbor and Grad-CAM analyses show clinically plausible cross-dataset structure and thoracic attention patterns in many cases, while device-heavy and false-positive cases remain ambiguous. Auxiliary architecture checks are task-dependent and do not support a universal backbone ranking. Overall, the study shows that apparent competence under a single protocol can conceal transfer, alignment, and shortcut-related failure modes, motivating stress-tested evaluation of medical VLMs under distribution shift.

1. Introduction

Deep learning models for chest X-ray analysis can degrade across hospitals because acquisition pipelines, patient populations, and reporting conventions change across institutions. Models may therefore exploit scanner, view, or documentation regularities that do not transfer with pathology [18, 27]. This problem becomes more difficult for vision–language models (VLMs), where image and text streams can each carry dataset-specific signals.

Medical VLMs align radiographs with clinical text and can support zero-shot prediction or retrieval [13, 19, 26, 28, 29]. Yet strong in-domain performance does not establish that the learned representation remains useful when the image domain, report style, or supervision regime changes. Recent reviews likewise identify robustness, grounding, and evaluation under dataset variation as open problems for medical vision–language systems [20].

This paper asks a narrower question: what failure modes become visible when chest X-ray representations are stress-tested across datasets? This framing is directly relevant to epistemic intelligence under distribution shift. A system may appear competent on familiar data while its representation fails to support the same decision or alignment externally. We do not estimate epistemic uncertainty and do not claim a formal notion of what a model “knows.” Instead, we operationalize one EIML-relevant blind spot through controlled tests of transfer, alignment, and recoverable source-proxy information.

We use NIH ChestXray14 [25] and CheXpert [14] for visual transfer, and PadChest [5] and OpenI [6] for paired multimodal evaluation. The primary tier uses matched ResNet-18 encoders so that initialization effects are not confounded with backbone changes. A separate auxiliary tier probes architecture sensitivity without using it to support causal claims about architecture superiority.

Our prior work studied feature-level site leakage in a single-modal cross-hospital setting [4]. The present study changes the scientific scope in four ways: it evaluates image–text alignment, separates paired-supervision effects from source-only visual transfer, tests external paired retrieval, and measures metadata-derived source-proxy recoverability in multimodal representations. This distinction is important because the present claims concern multimodal evaluation behavior rather than a new site-invariant encoder.

Contributions.

- We present a controlled stress-test of medical VLM representations under distribution shift, separating source-only visual transfer, unsupervised domain-adaptation diagnostics, multimodal pair-index retrieval, and source-proxy leakage analysis.
- We show that self-supervised visual initialization improves matched NIH-to-CheXpert transfer over supervised ImageNet initialization, while stronger adversarial adaptation becomes unstable.
- We define a dual-scale multimodal *evaluation setting*, not a new benchmark, using PadChest for paired training/in-domain analysis and OpenI for external stress testing.
- We make the leakage claim precise: OpenI has no explicit hospital-site labels, so we report *metadata-derived source-proxy leakage* rather than site leakage.
- We complement scalar metrics with qualitative cross-dataset retrieval and Grad-CAM analyses, and we treat architecture results as task-dependent sensitivity checks rather than a backbone ranking.

2. Related Work

2.1. Chest X-ray Generalization and Self-Supervision

Cross-institution generalization remains difficult in chest radiography because models can exploit acquisition and cohort-specific shortcuts [3, 18, 27]. Self-supervised learning can improve medical image transfer when labeled data are limited or shifted [2, 22]. Our visual-transfer tier uses this literature as a controlled representation baseline rather than as an architectural contribution.

2.2. Medical Vision–Language Representation Learning

ConVIRT, GLoRIA, CheXzero, and CXR-CLIP illustrate the progression from paired medical contrastive pretraining to chest X-ray-specific language–image systems [13, 26, 28, 29]. These methods demonstrate useful image–text structure, but evaluation can depend strongly on dataset construction, prompt formulation, and candidate-pool definition. Our focus is therefore diagnostic: we ask which conclusions survive conservative cross-dataset stress tests.

2.3. Leakage, Grounding, and Evaluation

Shortcut learning and hidden stratification motivate patient-disjoint splits and explicit nuisance probes [3, 18]. In multimodal data, the problem can be amplified because image and report modalities may share source-specific regularities. We therefore pair utility metrics with source-proxy probes and avoid equating metadata-derived proxies with true hospital-site labels.

3. Experimental Setup

3.1. Datasets and Roles

We use four public chest X-ray datasets with fixed roles. NIH ChestXray14 is the labeled source for visual transfer and CheXpert is the external target. PadChest is the paired training corpus and in-domain multimodal evaluation set. OpenI is the external paired stress-test set. All images are converted to grayscale, resized to 224×224 , and normalized under a common pipeline. Patient-disjoint splits are used when patient identifiers are available. Grouping and label-mapping details are provided in the supplement.

3.2. Primary Matched Visual Encoders

The primary controlled tier uses ResNet-18 [11] and varies initialization while keeping architecture fixed:

- **ImageNet**: supervised natural-image initialization [7].
- **BYOL**: self-supervised chest X-ray initialization [10].
- **OpenI CLIP-init**: a CLIP-style image encoder pre-trained on OpenI, used only where this does not contaminate the evaluation set.

The OpenI-pretrained CLIP initialization is **not** a held-out OpenI baseline and is excluded from external OpenI quantitative claims.

3.3. Source-Only Transfer and Unsupervised Adaptation

For source-only transfer, the classification head and any trainable visual layers use NIH labels only. CheXpert labels are never used for parameter updates, hyperparameter selection, or early stopping. We report a frozen-backbone linear probe and partial fine-tuning of the final residual block plus classifier.

DANN [9] and CORAL [21] are reported separately as *unsupervised domain-adaptation* diagnostics because they use unlabeled target-domain features to construct domain or covariance losses. They are therefore not presented as pure domain-generalization methods. We use scheduled adversarial strength and monitor the complete training trajectory to expose instability.

3.4. Multimodal Construction and Retrieval

We pair an image encoder with BioClinicalBERT [1], project both modalities to a normalized shared space, and

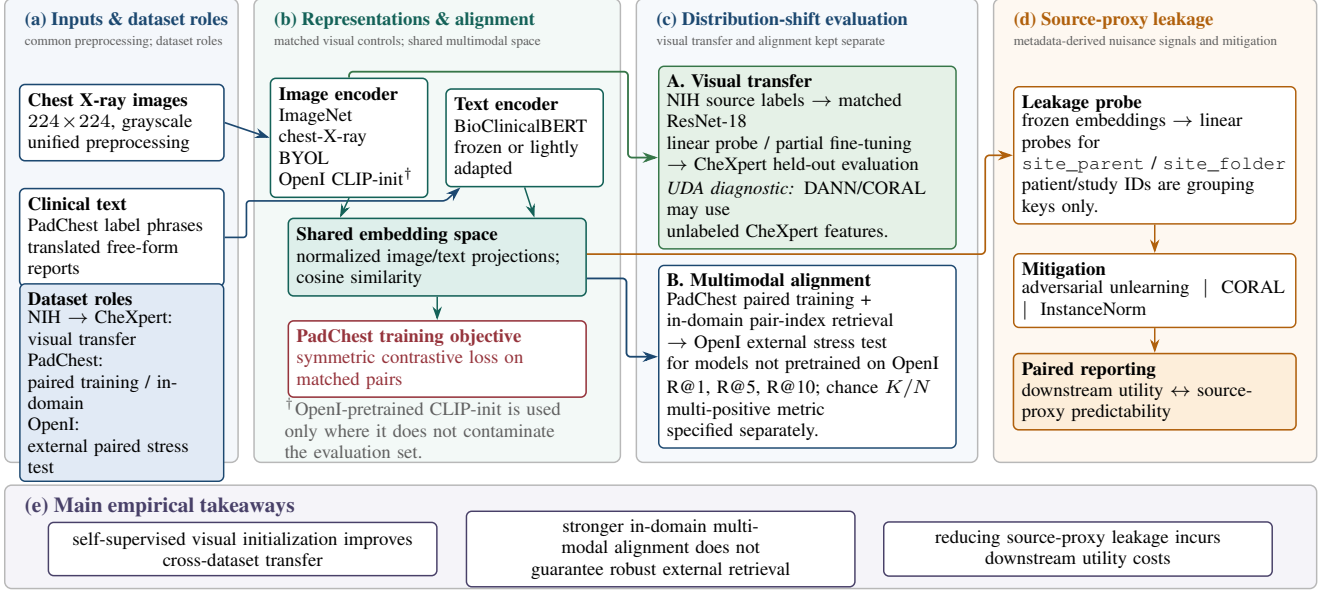


Figure 1. **Overview of the corrected leakage-aware evaluation framework.** The pipeline separates three questions that must not be conflated: source-only visual transfer, multimodal alignment under an external dataset shift, and recoverability of metadata-derived source proxies. OpenI is considered held out only for models whose pretraining did not use OpenI. The OpenI-pretrained CLIP initialization is therefore excluded from held-out OpenI quantitative claims.

optimize symmetric contrastive loss on paired PadChest data. We examine two text constructions: standardized label-derived English phrases and automatically translated free-form reports.

Retrieval is reported as a strict *pair-index retrieval diagnostic* with Recall@ K , $K \in \{1, 5, 10\}$. This wording is deliberate. PadChest templates can repeat across samples, and OpenI can associate multiple images with report-level text. Consequently, an index-exact miss is not necessarily a semantic mismatch. We therefore do not interpret these values as a clinical semantic-retrieval benchmark. For OpenI, the processed evaluation index contains $N = 6800$ candidate entries. We do not equate this count with the number of unique OpenI reports. Chance is K/N . The PadChest diagnostic analogously uses $N = 15000$ candidate entries.

The label-derived PadChest templates are used to test sensitivity to controlled linguistic formulation. Because duplicate templates can create multiple valid semantic positives, they are not used to support claims about exact one-to-one semantic retrieval. A duplicate-aware or multi-positive retrieval protocol would be required for that stronger claim.

3.5. Source-Proxy Leakage

OpenI does not provide explicit hospital-site labels. We therefore use two metadata-derived source proxies available in the processed data: `site_parent` and `site_folder`. These names are retained to match the implementation, but they are not treated as verified site iden-

tities. Linear probes predict the proxy class from frozen embeddings.

The split rule is group-aware: proxy classes may occur in both train and test because they are the prediction targets, while patient or study groups are kept disjoint to prevent memorization of repeated samples. We do not use patient identity as a probe target. Patient/study identifiers are grouping keys only.

3.6. Leakage Mitigation and Architecture Sensitivity

We evaluate adversarial proxy unlearning, CORAL alignment, and an InstanceNorm ablation [24]. Each method is reported jointly with downstream utility and proxy predictability.

A separate architecture-sensitivity tier uses ResNet-50, DenseNet-121, EfficientNet-B0, ViT-S, Swin-T, and a CLIP-initialized ResNet-50 [8, 11, 12, 16, 23]. These runs are auxiliary. Their task definitions differ from the primary matched transfer experiment, so we do not combine heterogeneous metrics into a single architecture-ranking table.

3.7. Implementation Notes

Experiments are implemented in PyTorch. Adam or AdamW [15, 17] is used depending on the experiment, with fixed settings within each matched comparison and validation-based early stopping where applicable. CheXpert uncertainty handling and the cross-dataset pathology

Table 1. NIH ChestXray14 $\rightarrow$ CheXpert transfer AUC with matched ResNet-18 encoders. The archived manuscript contains point estimates but not the seed-level records required to reconstruct mean $\pm$ std; no significance claim is made.

Initialization	Linear probe	Partial fine-tune
ImageNet	0.82	0.85
BYOL	0.87	0.88

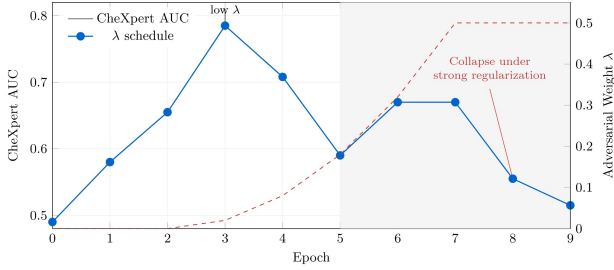


Figure 2. **DANN sensitivity on CheXpert.** Target AUC across training epochs under a scheduled gradient-reversal weight. The trajectory is used as a stability diagnostic, not as evidence for a universal benefit of domain adaptation.

mapping are specified in the supplement. We report variability only for experiments for which repeated-run statistics are available, rather than implying a uniform seed count across all runs. The supplement includes an explicit reproducibility audit that separates archived settings from run-level metadata that were not preserved.

4. Results

4.1. Matched NIH-to-CheXpert Transfer

Table 1 reports the central quantitative evidence for the matched ResNet-18 comparison. BYOL initialization improves AUC over ImageNet initialization under both reported regimes. The comparison is architecture-matched and uses NIH labels only.

The result supports a limited claim: under this matched setup, chest X-ray SSL provides a better starting representation than supervised ImageNet initialization. It does not establish that initialization dominates architecture in general.

4.2. Unsupervised Domain-Adaptation Behavior

Figure 2 shows that DANN behavior is sensitive to adversarial strength. Target AUC rises early and then degrades as adversarial pressure becomes stronger. CORAL is less erratic in the observed runs, but neither method is treated as a substitute for source representation quality.

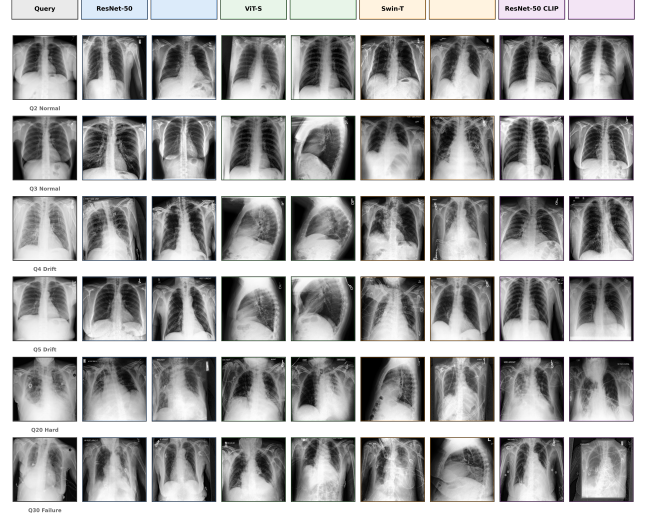


Figure 3. **Cross-dataset nearest neighbors in frozen image-representation space.** For each NIH query, top CheXpert neighbors are shown for an SSL ResNet-50, ViT-S, Swin-T, and a CLIP-style ResNet-50 image encoder. Clean cases often preserve plausible anatomy/view similarity, while harder cases expose ambiguity and view sensitivity.

4.3. Multimodal Pair-Index Retrieval

Table 2 reports the strict pair-index diagnostic. The processed OpenI evaluation index contains 6800 candidate entries. This count is not interpreted as the number of unique reports. The OpenI-pretrained CLIP initialization is omitted from the held-out OpenI rows because it has seen OpenI during pretraining.

The external OpenI numbers are low in absolute terms and remain close to chance at larger K . The defensible conclusion is therefore specific: exact pair identity learned from PadChest does not transfer strongly to the processed OpenI pool under this strict diagnostic. Because duplicate or report-level positives are not modeled, we do not generalize this result to all forms of semantic medical image-text retrieval.

4.4. Qualitative Representation Structure

The two figures provide evidence about representation behavior that scalar AUC alone cannot show. They support a modest claim that clinically plausible visual structure is often retained. They do not establish lesion localization accuracy or causal grounding.

4.5. Source-Proxy Leakage and Mitigation

The OpenI probes show that metadata-derived source proxies remain recoverable from frozen representations. Because the proxies come from dataset organization metadata rather than verified hospital acquisition labels, the correct

Table 2. Strict pair-index image-to-text retrieval. OpenI and PadChest pool sizes refer to candidate entries in the processed evaluation index. These numbers should not be interpreted as semantic retrieval accuracy when multiple indices can share equivalent report text. OpenI CLIP-init is intentionally omitted from the held-out OpenI comparison because that initialization was pretrained on OpenI.

Dataset	Initialization	R@1	R@5	R@10	External held-out?
OpenI, $N = 6800$ entries	Chance	0.000147	0.000735	0.001471	–
OpenI, $N = 6800$ entries	ImageNet + Text	0.0006	0.0010	0.0020	Yes
OpenI, $N = 6800$ entries	BYOL + Text	0.0002	0.0010	0.0022	Yes
PadChest, $N = 15000$ entries	Chance	0.000067	0.000333	0.000667	–
PadChest, $N = 15000$ entries	ImageNet + Text	0.0015	0.0019	0.0024	In-domain
PadChest, $N = 15000$ entries	BYOL + Text	0.0007	0.0015	0.0026	In-domain
PadChest, $N = 15000$ entries	OpenI CLIP-init + Text	0.0009	0.0014	0.0022	In-domain

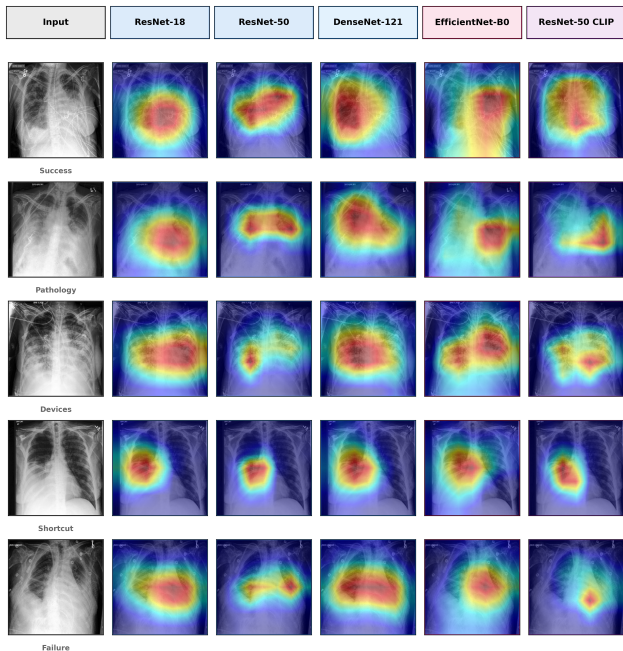


Figure 4. **Grad-CAM for consolidation classification on representative CheXpert cases.** The maps frequently concentrate attention within thoracic regions, while device-heavy and false-positive cases are more ambiguous. These visualizations are qualitative attention diagnostics, not a localization benchmark.

interpretation is source-proxy recoverability, not proof of hospital-site identification.

Figure 5 shows that InstanceNorm reduces proxy accuracy most aggressively and incurs the largest utility cost. Adversarial unlearning provides an intermediate reduction with a smaller, but still visible, utility loss. This result supports a trade-off claim only for the evaluated proxy and task.

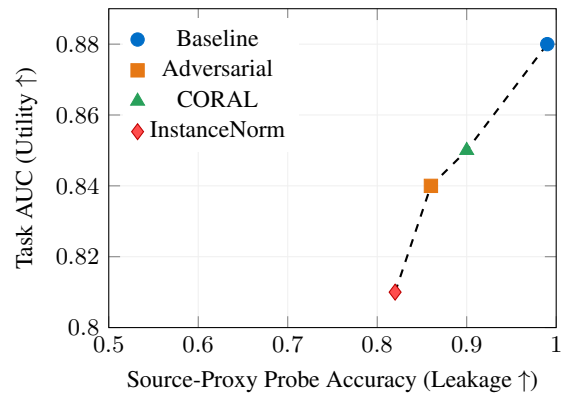


Figure 5. **Utility–source-proxy leakage trade-off.** InstanceNorm gives the largest reduction in proxy accuracy and also the largest utility loss. Adversarial unlearning is intermediate, while CORAL produces a milder change.

Table 3. Auxiliary consolidation-classifier AUC. These results are a sensitivity check and are not directly comparable with the primary NIH-to-CheXpert transfer table.

Backbone	AUC	Note
ResNet-18	0.9257	Stable
ResNet-50	0.9163	Stable
DenseNet-121	0.9200	Stable
EfficientNet-B0	0.9203	Stable
ResNet-50 CLIP	0.9155	Stable
ViT-S	0.8774	Lower on this task
Swin-T	0.5373	Unstable exploratory

4.6. Architecture Sensitivity Is Task-Dependent

Table 3 reports a single auxiliary consolidation task, separate from the different CheXpert-wide protocol. Swin-T reaches 0.5373 in this run and is marked unstable, consistent with the supplementary table.

These results do not justify the previous statement that architecture differences are universally modest. The defensible interpretation is task-dependent. CNN-family encoders cluster closely on this consolidation check, ViT-S is lower, and Swin-T is unstable. A separate CheXpert-only architecture table in the supplement uses a different label space and should be read only as an auxiliary evaluation.

5. Limitations

First, the multimodal study uses one large paired training corpus and one external paired stress-test dataset. Second, the current retrieval archive supports exact pair-index Recall@ K but does not preserve a validated equivalence map for all duplicate or report-level positives. We therefore define a duplicate-aware multi-positive metric in the supplement but do not fabricate an uncomputed score; the reported exact-index metric remains a conservative stress-test diagnostic. Third, the central matched transfer values are archived as point estimates, while the seed-level records required to reconstruct mean $\pm$ std are unavailable in the supplied experiment archive. We consequently make no significance claim for the 0.82–0.88 differences. Fourth, the source bundle does not preserve a complete run-level hyperparameter ledger for every experimental family, so the supplement explicitly marks which settings are known and which exact values are unavailable. Fifth, the exact GPT-4 API/model snapshot used for PadChest translation was not preserved; temperature zero reduces decoding variability but does not guarantee identical regeneration. Beyond these four reproducibility limitations, OpenI provides no verified hospital-site labels, the architecture sweep is auxiliary and heterogeneous across tasks, the study is restricted to chest radiography, and we do not measure calibrated epistemic uncertainty, abstention, or formal unknown-unknown detection.

6. Conclusion

We studied chest X-ray VLMs through controlled stress tests of representation transfer, multimodal pair-index alignment, and metadata-derived source-proxy recoverability. In matched NIH-to-CheXpert comparisons, BYOL initialization improves over ImageNet initialization. Unsupervised adversarial adaptation is useful only in a narrow regime and becomes unstable as adversarial pressure increases. Under external OpenI evaluation, strict exact-pair retrieval remains weak, while source-proxy signals remain recoverable from frozen embeddings. Qualitative retrieval and Grad-CAM results show plausible cross-dataset structure and thoracic attention patterns but also persistent ambiguity in difficult cases. Auxiliary architecture checks are task-dependent and do not support a universal ranking. These findings fit an epistemic-intelligence perspective as an evaluation result: apparent in-domain competence can coexist with representation-level blind spots under distribution shift. The paper therefore argues for explicit stress tests, contamination-aware baselines, and precise claims about

what evaluation protocols actually measure.

References

- [1] Emily Alsentzer, John Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *NAACL Clinical NLP Workshop*, 2019.
- [2] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, et al. Big self-supervised models advance medical image classification. In *ICCV*, 2021.
- [3] Marcus Badgeley et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2019.
- [4] Ayoub Louaye Bouaziz and Lokmane Cheboub. Feature-level site leakage reduction for cross-hospital chest x-ray transfer via self-supervised learning, 2026.
- [5] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and María de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Scientific Data*, 2020.
- [6] Dina Demner-Fushman, Marc Kohli, Marc Rosenman, Sonya Shooshan, Laritza Rodriguez, Sameer Antani, George Thoma, and Clement McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 2016.
- [7] Jia Deng et al. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [8] Alexey Dosovitskiy et al. An image is worth 16x16 words. In *ICLR*, 2021.
- [9] Yaroslav Ganin et al. Domain-adversarial training of neural networks. In *JMLR*, 2016.
- [10] Jean-Bastien Grill et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [12] Gao Huang et al. Densely connected convolutional networks. In *CVPR*, 2017.
- [13] Xinyue Huang et al. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *ICCV*, 2021.
- [14] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviu Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Babak Haghighi, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019.
- [15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- [16] Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.
- [17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019.
- [18] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Nature Medicine*, 2020.
- [19] Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- [20] Ji Seung Ryu, Hyunyoung Kang, Yuseong Chu, and Sejung Yang. Vision-language foundation models for medical imaging: a review of current practices and innovations. *Biomedical Engineering Letters*, 15:809–830, 2025.
- [21] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshops*, 2016.
- [22] A. Taleb et al. 3d self-supervised methods for medical imaging. *Medical Image Analysis*, 2020.
- [23] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.
- [24] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance

- normalization: The missing ingredient for fast stylization. In *arXiv*, 2016.
- [25] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification. In *CVPR*, 2017.
 - [26] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun Kyoung Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2023.
 - [27] John Zech, Marcus Badgeley, Manway Liu, Andrew Costa, Joshua Titano, and Eric Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs. *PLoS Medicine*, 2018.
 - [28] Yuhao Zhang et al. Convirt: Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.
 - [29] Yuhao Zhang et al. Chexzero: Zero-shot chest x-ray classification via contrastive language-image pretraining. *Nature Biomedical Engineering*, 2022.

What Do Medical Vision–Language Models Learn in Radiology? Transfer, Alignment, and Source-Proxy Leakage Under Distribution Shift

Supplementary Material

A. External Zero-Shot Contextual Baselines

This appendix reports the external zero-shot baselines omitted from the main paper for clarity. These models are included only as contextual reference points under the same unified preprocessing, label mapping, and AUROC-based evaluation protocol. Because they are not architecture-matched controls, they are not used as primary evidence for the paper’s controlled conclusions.

Table 4. Zero-shot contextual baselines on CheXpert under the unified evaluation protocol. We report mean AUC over the 5-label setting for calibration only. These results are not architecture-matched to the primary controlled experiments. All models are evaluated without retraining under the same preprocessing and label-mapping pipeline.

Model	Mean AUC (macro, 5-label)
CheXzero	0.56 ± 0.021
CXR-CLIP	0.58 ± 0.0175

The zero-shot results are reported only to contextualize the numerical scale of the primary experiments. Pretraining-data provenance differs across external checkpoints. In particular, public CXR-CLIP checkpoints exist with training compositions that include CheXpert, so the value shown here must not be interpreted as held-out CheXpert generalization unless the exact checkpoint provenance is verified. These baselines are not used for architecture or transfer claims.

B. Extended Architecture Tables and Exploratory Results

This appendix reports auxiliary architecture-family comparisons. These runs use task definitions that differ from the primary NIH-to-CheXpert matched transfer experiment and therefore must not be combined into a single architecture-ranking claim. The wide per-pathology table below is a CheXpert-only auxiliary evaluation over the native CheXpert observation set, not the seven-label cross-dataset transfer task defined later in the appendix.

The consolidation check is a separate auxiliary task. It should not be numerically combined with the CheXpert-only table above or with the primary NIH-to-CheXpert transfer experiment. Swin-T reached 0.5373 in this auxil-

iary run and was unstable, so it is reported for completeness rather than used as primary evidence.

C. Leakage-Aware Evaluation Protocol

C.1. Protocol Overview

Leakage-Aware Evaluation Protocol (Summary)

This appendix provides the full specification of the leakage-aware evaluation protocol used in this study. The protocol standardizes dataset roles, preprocessing, split construction, evaluation metrics, and diagnostic probes in order to isolate representation transfer, multimodal alignment quality, and potential shortcut signals.

The protocol consists of three complementary experimental settings:

- (1) Source-only cross-dataset transfer.** Visual representations are trained with NIH ChestXray14 labels and evaluated on CheXpert. CheXpert labels are not used for model updates, selection, or early stopping.
- (2) Multimodal pair-index retrieval.** Image–text encoders are trained on paired PadChest data. PadChest is used for in-domain analysis and OpenI (IU X-ray) is used as an external pair-index stress test. OpenI-pretrained CLIP initialization is excluded from held-out OpenI claims.
- (3) Source-proxy diagnostics.** Frozen embeddings are probed for metadata-derived OpenI source proxies. Patient/study identifiers are grouping keys, not probe targets.

Across all experiments the protocol enforces:

- **Patient-disjoint splits** whenever patient identifiers are available.
- **Unified preprocessing** across datasets.
- **Controlled architectures and training regimes.**
- **Explicit chance baselines** for retrieval tasks:
 $R@K = \frac{K}{N}$.
- **Paired reporting of utility and leakage metrics.**

Table 7 summarizes dataset roles.

C.2. Unified Preprocessing

To minimize dataset-specific artifacts and ensure fair cross-dataset comparison, all images are processed using a unified preprocessing pipeline before training and evaluation. The same preprocessing steps are applied across NIH

Table 5. CheXpert-only auxiliary per-pathology AUROC over the broader CheXpert observation set. This table is not the NIH-to-CheXpert seven-label transfer experiment and is not directly comparable with the primary matched transfer results. Best result per pathology is highlighted in bold.

Model	NF	ECM	Cardio	Opac	Les	Edema	Cons	Pneum	Atel	PTX	Eff	PIO	Frac	Supp	Mean
<i>CNN backbones</i>															
ResNet50	0.513	0.500	0.479	0.499	0.542	0.500	0.511	0.520	0.494	0.496	0.564	0.531	0.495	0.527	0.512
ResNet18	0.544	0.493	0.508	0.480	0.512	0.412	0.518	0.541	0.484	0.514	0.537	0.558	0.560	0.510	0.514
DenseNet121	0.567	0.499	0.479	0.512	0.511	0.424	0.514	0.508	0.460	0.516	0.503	0.529	0.526	0.606	0.511
ResNet50d	0.475	0.501	0.499	0.514	0.579	0.422	0.502	0.507	0.539	0.519	0.464	0.551	0.508	0.503	0.506
EfficientNet-V2-S	0.469	0.495	0.495	0.475	0.476	0.482	0.466	0.528	0.474	0.532	0.488	0.533	0.549	0.578	0.503
EfficientNet-B3	0.455	0.499	0.508	0.517	0.527	0.482	0.465	0.501	0.485	0.521	0.541	0.450	0.476	0.605	0.502
EfficientNet-B0	0.409	0.493	0.503	0.482	0.508	0.499	0.526	0.497	0.495	0.544	0.523	0.469	0.505	0.500	0.497
ResNet101	0.496	0.508	0.513	0.538	0.522	0.505	0.492	0.528	0.460	0.469	0.494	0.453	0.496	0.520	0.500
<i>Modern architectures</i>															
ViT-Base	0.478	0.492	0.505	0.505	0.534	0.517	0.517	0.530	0.513	0.506	0.475	0.481	0.519	0.535	0.508
ConvNeXt-Base	0.445	0.491	0.510	0.474	0.468	0.482	0.433	0.488	0.491	0.512	0.473	0.436	0.520	0.547	0.484
Swin-Base	0.496	0.505	0.495	0.524	0.432	0.470	0.479	0.472	0.485	0.459	0.500	0.595	0.466	0.473	0.489
<i>Self-supervised representations</i>															
DINO-ViT	0.676	0.519	0.512	0.452	0.480	0.444	0.479	0.509	0.520	0.480	0.471	0.443	0.515	0.474	0.498
MAE-ViT	0.419	0.494	0.501	0.468	0.533	0.435	0.518	0.518	0.486	0.437	0.512	0.440	0.502	0.496	0.483

Table 6. Auxiliary consolidation classifier check for the expanded backbone tier. Entries marked unstable are reported for completeness only and are not used as primary qualitative evidence in the main paper.

Backbone	Consolidation AUC	Note
ResNet-18	0.9257	Stable
ResNet-50	0.9163	Stable
DenseNet-121	0.9200	Stable
EfficientNet-B0	0.9203	Stable
ResNet-50 CLIP	0.9155	Stable
ViT-S	0.8774	Weaker than CNN family
Swin-T	0.5373	Unstable exploratory result

Table 7. Dataset roles in the leakage-aware evaluation protocol.

Dataset	Modality	Role in Protocol
NIH ChestXray14	Image + labels	Source dataset for transfer
CheXpert	Image + labels	Target dataset for transfer
PadChest	Image + reports	Multimodal training + in-domain retrieval
OpenI (IU X-ray)	Image + reports	External retrieval evaluation

ChestXray14, CheXpert, PadChest, and OpenI datasets.

Image normalization pipeline. All radiographs are converted to grayscale and resized to a fixed spatial resolution of 224×224 pixels using bilinear interpolation. Pixel intensities are normalized to the range $[0, 1]$ and subsequently standardized using dataset-independent statistics. No dataset-specific histogram equalization or contrast normalization is applied.

Color channel handling. Because chest radiographs are inherently grayscale images, single-channel images are replicated across three channels when required by architectures initialized with ImageNet weights.

Spatial processing. Images are resized directly to 224×224 without additional cropping or padding in order to preserve anatomical coverage. No center cropping or aspect-ratio preserving padding is applied.

Augmentation policy. During training, lightweight data augmentation is applied to improve robustness. This includes random horizontal flipping and small affine transformations (translation and scaling). No dataset-specific augmentations are used. Augmentations are disabled during validation and testing.

Preprocessing Pipeline

For each image x :

1. Load image from dataset.
2. Convert to grayscale if necessary.
3. Resize to 224×224 using bilinear interpolation.
4. Normalize pixel intensities to $[0, 1]$.
5. Replicate channel to three channels if required by the encoder.
6. Apply lightweight augmentation during training only.

This unified preprocessing pipeline ensures that performance differences across datasets and models arise from representation quality and training regime rather than differences in image formatting or resolution.

C.3. Split Construction and Grouping Rules

To prevent information leakage and ensure realistic evaluation conditions, all experiments use grouped data splits that enforce disjointness at the patient level whenever patient identifiers are available. This ensures that images from the same patient do not appear in both training and evaluation sets.

Patient-disjoint splits. For datasets that provide patient identifiers (NIH ChestXray14, CheXpert, and PadChest), splits are constructed such that all studies from a given patient are assigned to a single partition. This prevents models from exploiting patient-specific visual characteristics that could artificially inflate performance.

Cross-dataset transfer setting. In the single-modal transfer experiments, models are trained on NIH ChestXray14 and evaluated on CheXpert without any shared patients or studies between datasets. Because the datasets originate from different institutions, cross-dataset transfer provides a realistic domain shift scenario.

Multimodal training and evaluation splits. For multimodal experiments, PadChest is used for both training and in-domain retrieval evaluation. The dataset is partitioned into training, validation, and test sets using patient-disjoint splits. OpenI (IU X-ray) is used as an external dataset for cross-dataset retrieval evaluation.

External retrieval dataset (OpenI). The OpenI dataset contains image–report pairs collected from a different institution. When patient identifiers are available, patient-disjoint evaluation is enforced. Otherwise, study-level disjointness is maintained to prevent identical image–report pairs from appearing in multiple evaluation subsets.

Grouped splits for source-proxy probes. For source-proxy diagnostics, the proxy classes being predicted may occur in both probe-training and probe-test partitions. Disjointness is instead enforced over patient or study groups, so repeated samples from the same patient/study cannot appear in both partitions. This preserves a valid supervised probe while limiting sample-level memorization.

Table 8 summarizes the grouping keys and split policies used for each dataset.

These split construction rules ensure that performance differences reflect representation quality and cross-domain generalization rather than overlap between training and evaluation data.

Table 8. Dataset grouping keys and split policies used in the leakage-aware evaluation protocol.

Dataset	Group Key	Split Type	Usage
NIH ChestXray14	Patient ID	Patient-disjoint	Training (transfer source)
CheXpert	Patient ID	Patient-disjoint	Evaluation (transfer target)
PadChest	Patient ID	Patient-disjoint	Multimodal training + in-domain retrieval
OpenI (IU X-ray)	Patient / Study ID	Study-disjoint	External retrieval evaluation

C.4. Label Mapping and Uncertainty Handling

To enable cross-dataset transfer evaluation between NIH ChestXray14 and CheXpert, we construct a consistent label space across the two datasets. Because the two datasets differ in annotation taxonomy and label extraction pipelines, we define an explicit mapping between overlapping pathology categories.

Shared pathology set. Only pathologies that appear in both datasets are used in the transfer experiments. Labels that do not have a clear correspondence between datasets are excluded from the evaluation.

CheXpert uncertainty labels. CheXpert annotations include an explicit uncertainty label (U) produced by the automatic report labeler. Following common practice in chest X-ray classification benchmarks, uncertainty labels are handled using a fixed policy applied consistently across all experiments.

In our experiments, uncertainty labels are treated as **negative labels (U-Zeros)**. That is, uncertain observations are mapped to the negative class during training and evaluation. This strategy avoids introducing additional label noise while maintaining consistency with several prior studies.

Label filtering. Images that contain missing labels for the selected pathology set are excluded from the corresponding evaluation task. All remaining labels are treated as binary classification targets.

Table 9 shows the mapping between NIH ChestXray14 labels and CheXpert observations used in the transfer experiments.

This mapping ensures that transfer evaluation reflects differences in learned representations rather than inconsistencies in dataset annotation schemas.

C.5. Training Regimes

The primary transfer table reports two regimes whose semantics are explicit in the manuscript.

Linear probing. The backbone is frozen and only a linear classification head is optimized using NIH source labels.

Table 9. Mapping between NIH ChestXray14 and CheXpert labels used for cross-dataset transfer evaluation. Only pathologies with clear correspondence across datasets are retained.

NIH ChestXray14 Label	CheXpert Observation
Atelectasis	Atelectasis
Cardiomegaly	Cardiomegaly
Consolidation	Consolidation
Edema	Edema
Pleural Effusion	Pleural Effusion
Pneumonia	Pneumonia
Pneumothorax	Pneumothorax

Partial fine-tuning. The final residual block (layer4 for ResNet-18) and the classifier head are optimized using NIH source labels, while earlier layers remain frozen.

Target-label isolation. CheXpert labels are never used for parameter updates, validation selection, or early stopping in the source-only transfer experiment. When DANN or CORAL is enabled, unlabeled CheXpert features may enter the adaptation loss, which is why those runs are reported separately as unsupervised domain adaptation.

Optimization. Adam/AdamW-based optimization and validation-based early stopping are used depending on the experiment. Hyperparameters are held fixed within each matched comparison.

C.6. Multimodal Alignment and Retrieval Protocol

Multimodal alignment is evaluated using a strict pair-index image-text retrieval diagnostic. Each query has one designated positive index in the processed evaluation table. This is intentionally not described as a semantic one-to-one benchmark because multiple indices can share equivalent text or report-level content.

Image and text embeddings are produced using the visual encoder and the BioClinicalBERT text encoder, respectively. Both embeddings are projected into a shared latent space and normalized prior to similarity computation.

Retrieval performance is measured using Recall@ K ($R@K$), which evaluates whether the designated positive index appears among the top- K candidates. Explicit chance baselines are reported as K/N .

Algorithm 1: Multimodal Retrieval Evaluation Protocol

- 1: **Input:** trained image encoder f_θ , text encoder g_ϕ , evaluation dataset $\mathcal{D}$ with image-report pairs
- 2: Encode images: $z_i = f_\theta(x_i)$
- 3: Encode reports: $z_t = g_\phi(t_i)$
- 4: Project both embeddings into a shared space and normalize:

$$\tilde{z}_i = \frac{P_i(z_i)}{\|P_i(z_i)\|}, \quad \tilde{z}_t = \frac{P_t(z_t)}{\|P_t(z_t)\|}$$

- 5: Construct retrieval candidate pool of size N
- 6: **for** each query image x_q **do**
- 7: Compute cosine similarity with all report embeddings:

$$s(q, j) = \tilde{z}_i^{(q)} \cdot \tilde{z}_t^{(j)}$$

- 8: Rank candidate text indices by similarity score
- 9: Record Recall@ K if the designated positive index appears within the top- K
- 10: **end for**
- 11: Compute overall Recall@ K across all queries
- 12: Compute chance baseline:

$$R@K_{\text{chance}} = \frac{K}{N}$$

- 13: **Output:** R@1, R@5, R@10 and fold-over-chance improvement

External dataset evaluation. Cross-dataset pair-index retrieval is evaluated using OpenI. The processed evaluation index contains $N = 6800$ candidate entries. We do not interpret this number as the count of unique OpenI reports. The corresponding chance baselines are:

$$R@1 = 0.000147, \quad R@5 = 0.000735, \quad R@10 = 0.00147$$

Interpretation caveat. OpenI may contain multiple images associated with report-level text, and PadChest label-derived templates can repeat. Therefore, an index-exact miss can still be semantically plausible. The reported metric is used as a stringent transfer diagnostic, not as a clinical semantic-retrieval benchmark.

C.7. Source-Proxy Leakage Diagnostics

To assess whether learned representations encode dataset-specific shortcut signals, we train linear probes on frozen

image embeddings to predict metadata-derived source proxies. OpenI does not provide verified hospital-site labels, so these probes measure source-proxy recoverability rather than hospital-site identification.

Embedding extraction. For each trained model, image embeddings are extracted from the frozen visual encoder prior to the task-specific classifier. These embeddings serve as input features for leakage probes.

Metadata-derived source proxy variables. Because verified institutional site labels are unavailable in OpenI, we use two fields from the processed dataset organization as source proxies:

- **site_parent:** a coarse grouping derived from directory-level dataset structure.
- **site_folder:** a finer-grained grouping corresponding to image subdirectories.

These proxies can capture dataset-organization or acquisition-related regularities, but they are not assumed to correspond one-to-one with hospitals or scanners.

Linear probe training. For each proxy variable, a linear classifier is trained on top of the frozen image embeddings. The probe model consists of a single fully connected layer optimized using cross-entropy loss. No updates are applied to the underlying visual encoder.

Group-disjoint splits. The source-proxy class is the supervised target and may therefore be represented in both train and test. Patient/study groups are kept disjoint between partitions to prevent repeated samples from the same group from driving probe accuracy.

Evaluation metrics. Probe performance is measured using classification accuracy. High probe accuracy indicates that the representation retains information predictive of the chosen source proxy. It does not by itself establish hospital-site leakage.

Utility-leakage analysis. To study the trade-off between clinical utility and shortcut suppression, leakage probe accuracy is reported alongside task performance metrics (e.g., AUROC for pathology prediction). This allows models to be compared along a utility-leakage frontier, where improvements in task performance can be evaluated in relation to the degree of recoverable source-proxy information.

C.8. Reporting Checklist

To ensure consistent and transparent evaluation across experiments, we adopt a standardized reporting checklist for all results presented in this work. Each experiment follows

the same reporting protocol to allow fair comparison between representation initializations, training regimes, and adaptation methods.

For each experiment, the following quantities are reported:

- **Task utility metric.** Primary task performance is reported using the appropriate evaluation metric for the task. For pathology classification tasks, we report area under the receiver operating characteristic curve (AUROC). For multimodal retrieval tasks, we report Recall@ K metrics (R@1, R@5, R@10).
- **Chance baselines.** For retrieval experiments with large candidate pools, chance performance is explicitly reported using the analytical baseline

$$R@K_{\text{chance}} = \frac{K}{N},$$

where N denotes the candidate pool size.

- **Leakage probe performance.** For models evaluated with leakage diagnostics, probe accuracy for metadata-derived source proxies is reported alongside the main task metric to quantify shortcut signal strength.
- **Multiple training seeds.** Repeated-run dispersion is reported only where repeated runs are available. We do not assume a uniform seed count across all experimental families.
- **Dataset split specification.** Patient/study grouping follows the split rules specified above whenever identifiers are available.

This reporting protocol ensures that model performance is interpreted jointly with potential shortcut signals and evaluation baselines, enabling more reliable comparison across experimental settings.

D. Reproducibility and Evidence Audit

The final audit separates what can be reconstructed from the archived experiment bundle from information that was not preserved. Missing run-level metadata are reported explicitly rather than reconstructed from assumptions.

Four Evidence Gaps and Their Treatment

1. **Duplicate-aware retrieval.** Exact pair-index Recall@ K is available, but the archived candidate table does not contain a validated semantic-equivalence map for every duplicate/report-level positive. We specify a multi-positive metric below and do not invent an uncomputed score.
2. **Repeated-run dispersion for the central transfer table.** The archived manuscript retains the point estimates but not the seed-level values required to reconstruct mean $\pm$ standard deviation. The main table therefore reports point estimates only and makes no statistical-significance claim.
3. **Complete hyperparameter ledger.** Core protocol choices are recoverable, but exact run-level learning rates, batch sizes, epoch caps, and early-stopping patience are not available for every experimental family. Table 10 records this status explicitly.
4. **Translation model snapshot.** The translation archive records the GPT-4 model family and temperature 0, but not the exact API/model snapshot identifier. Regenerating the translations may therefore produce small differences even with the same prompt.

D.1. Duplicate-Aware Multi-Positive Retrieval Definition

For a query image q , let $\mathcal{P}(q)$ denote the set of candidate text indices that are valid positives after report-ID matching or validated text-equivalence grouping. A duplicate-aware Recall@ K can be defined as

$$R@K_{\text{multi}} = \frac{1}{Q} \sum_{q=1}^Q \mathbb{1}[\text{TopK}(q) \cap \mathcal{P}(q) \neq \emptyset].$$

This metric prevents an image from being marked incorrect when the retrieved candidate belongs to the same validated report-level or equivalent-text positive set. The current archived evaluation index preserves the designated pair index but not a validated $\mathcal{P}(q)$ for every candidate. We therefore retain exact pair-index Recall@ K as the reported conservative diagnostic and leave $R@K_{\text{multi}}$ unreported rather than manufacturing equivalence labels post hoc.

D.2. Run-Level Reproducibility Ledger

Recommended release artifact. For exact reproduction, the release should include a machine-readable run manifest containing optimizer, learning rate, batch size, epoch cap, early-stopping rule, random seed, checkpoint provenance, candidate-index construction, and the exact translation model snapshot. The current paper does not claim that these missing fields can be reconstructed from prose alone.

E. PadChest Text Construction and Translation

This appendix describes how textual inputs are constructed for multimodal experiments using the PadChest dataset. Because PadChest radiology reports are written in Spanish, we consider two alternative textual representations: (1) structured label-derived English phrases and (2) translated free-form radiology reports. This design allows us to analyze the effect of textual formulation on multimodal alignment and retrieval performance.

E.1. Label-Derived English Phrase Templates

In the first representation, text inputs are generated directly from structured PadChest annotation labels. This approach avoids translation noise and ensures consistent phrasing across samples.

For each image, the associated clinical findings are converted into a standardized English sentence using a fixed template:

Label-Derived Phrase Template

Template:

Chest X-ray showing: [LABEL_1], [LABEL_2], ...

Examples:

- “Chest X-ray showing: cardiomegaly and pleural effusion.”
- “Chest X-ray showing: pulmonary edema.”
- “Chest X-ray showing: no acute cardiopulmonary abnormality.”

When multiple findings are present, labels are concatenated into a single sentence separated by conjunctions. Negative findings are represented using standardized negation phrases when available in the PadChest annotation metadata.

This label-derived formulation provides a controlled textual representation with minimal linguistic variability and allows the multimodal alignment task to focus on the correspondence between visual findings and structured clinical descriptions.

E.2. Spanish-to-English Radiology Report Translation

In the second representation, the original PadChest radiology reports are translated from Spanish to English. Translation is performed using GPT-4 with temperature set to 0 to reduce decoding variability while preserving the source report content.

A controlled prompting strategy is used to preserve clinical meaning while preventing the model from introducing hallucinated findings. The prompt explicitly instructs the

Table 10. Reproducibility ledger for the archived experiments. “Archived” means the value is recoverable from the supplied manuscript/-source bundle. “Not preserved” means no exact value is available in the supplied archive and none is inferred.

Component	Setting supported by archive	Status	Consequence for interpretation / re-production
Image preprocessing	Grayscale, 224×224, common normalization; three-channel replication when required	Archived	Reconstructable from protocol description
CheXpert uncertainty	U-Zeros policy for the mapped transfer labels	Archived	Reconstructable from protocol description
Primary transfer trainability	Linear probe; partial fine-tuning of ResNet layer4 + classifier; NIH labels only	Archived	Reconstructable at the regime level
Optimizer family	Adam/AdamW depending on experiment; fixed within matched comparisons	Partly archived	Exact branch-specific optimizer is not recoverable for every run
Learning rate / batch size / epoch cap	Exact run-level values are absent for part of the experiment suite	Not preserved	Exact numerical rerun cannot be guaranteed from the manuscript bundle alone
Early stopping	Validation-based where applicable	Partly archived	Exact patience/criterion threshold is not preserved for every run
DANN schedule	Scheduled gradient-reversal strength; full trajectory retained in the reported diagnostic figure	Partly archived	Qualitative instability is documented; exact schedule reconstruction is incomplete
Multimodal text encoder	BioClinicalBERT with normalized shared-space projection and symmetric contrastive loss	Archived	Core multimodal construction is reconstructable
Central transfer repetitions	Point estimates 0.82/0.85 and 0.87/0.88 retained; seed-level values absent	Not preserved	No mean±std or significance test is reported
Translation prompt	Full Spanish-to-English prompt and temperature 0	Archived	Prompt can be reused
GPT translation snapshot	Exact GPT-4 API/model snapshot identifier	Not preserved	Regeneration may differ from the archived translated text

model to perform faithful translation without adding or removing clinical information.

Radiology Report Translation Prompt

System Prompt

You are a clinical translation assistant. Translate Spanish radiology reports into English faithfully. Do not add, remove, infer, or diagnose findings that are not explicitly stated in the original report.

User Prompt

Translate the following Spanish chest X-ray report into English.

Rules:

1. Preserve negation and uncertainty expressions (e.g., “no se observa”, “posible”, “sugiere”).
2. Maintain the structure of the original report when possible.
3. Do not introduce measurements, devices, or diagnoses that are not explicitly present in the source text.
4. Preserve medical abbreviations when uncertain.
5. If a term cannot be confidently translated, retain the Spanish term and indicate it in a short note.

TEXT:

<SPANISH RADIOLOGY REPORT>

Temperature 0 reduces decoding variability but does not guarantee bitwise or API-level reproducibility. The archived source bundle identifies the model family as GPT-4 but does not preserve the exact API/model snapshot identifier. The translated text used in the experiments should therefore be treated as the primary artifact; regeneration from the prompt may differ slightly.

This prompting strategy preserves the semantic content of radiology reports while minimizing translation artifacts that could affect multimodal alignment.

E.3. Translation Quality Control

To ensure translation fidelity, additional quality checks are applied to the translated reports. These checks include verifying that negation expressions are preserved, confirming that no additional clinical findings are introduced, and identifying ambiguous or untranslated terms.

In addition, retrieval experiments are performed using both textual formulations (label-derived phrases and translated reports). Comparing retrieval performance across these two textual representations allows us to evaluate the sensitivity of multimodal alignment to linguistic variability and translation noise.

The label-derived formulation can produce duplicate or semantically equivalent strings across samples. It is therefore treated as a controlled text-construction ablation rather than evidence for one-to-one semantic retrieval. The translated free-form condition retains more linguistic variability but also introduces translation dependence.

F. Diagnostic Experiments

This appendix provides additional diagnostic analyses that support the interpretation of the experimental results reported in the main paper. These diagnostics are not used as primary claims but serve to validate the reliability of the evaluation pipeline under the strict protocol used in this study.

Because our evaluation uses strict cross-dataset protocols and large candidate pools for retrieval, absolute performance values can appear numerically small. The diagnostics presented here check representation similarity and candidate-pool scaling. They do not prove semantic grounding and are not used as primary evidence for the paper’s main claims.

F.1. Representation Similarity Analysis

To analyze how different initialization strategies influence learned representations, we compute Centered Kernel Alignment (CKA) similarity between feature embeddings produced by models trained under different pretraining regimes.

Given two feature matrices X and Y extracted from the

Table 11. Representative CKA similarity values between representations obtained from different initialization strategies.

Model Pair	Dataset	CKA
ImageNet vs BYOL	NIH	0.81
ImageNet vs CLIP-init	NIH	0.74
BYOL vs CLIP-init	NIH	0.79

Table 12. Sensitivity of retrieval performance to candidate pool size. Absolute Recall@ K values decrease as the candidate pool grows, consistent with theoretical expectations.

Pool Size (N)	R@1	R@5	R@10
500	0.0060	0.021	0.041
1000	0.0030	0.013	0.028
6800	0.0013	0.0062	0.013

same set of images, linear CKA is defined as:

$$\text{CKA}(X, Y) = \frac{\|Y^\top X\|_F^2}{\|X^\top X\|_F \cdot \|Y^\top Y\|_F}.$$

Higher CKA values indicate stronger similarity between representations.

Table 11 reports representative CKA similarities between visual encoders initialized with different pretraining strategies. These values are computed on embeddings extracted from the NIH validation split.

These similarities suggest that self-supervised and multimodal pretraining produce related but not identical feature structures, reflecting differences in the supervision signals used during representation learning.

F.2. Retrieval Pool Size Sensitivity

Retrieval performance depends on the size of the candidate pool used during evaluation. To assess the sensitivity of Recall@ K metrics to pool size, we evaluate retrieval performance on OpenI under different candidate pool sizes.

Table 12 reports retrieval performance for several pool sizes.

As expected, $\text{Recall}@K$ values decrease as the number of candidate entries increases. This behavior confirms that the retrieval evaluation behaves consistently with theoretical scaling and that the reported results are not artifacts of a specific candidate-pool size configuration.