MediRely: Reliability-Aware Retrieval for Robust Multimodal Clinical Decision Support

Jia-Xian Jian[1], Jenq-Neng Hwang[2], and Pau-Choo Chung[1]

[1] Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan
n28101173@gs.ncku.edu.tw, pcchung@ee.ncku.edu.tw
[2] Department of Electrical & Computer Engineering, University of Washington, Seattle, USA
hwang@uw.edu

Abstract. Multimodal clinical decision support (CDS) systems often rely on medical images and complementary clinical context, yet such context may be unavailable at deployment. Existing missing-modality methods typically require retraining or generative imputation, increasing deployment cost and potentially introducing unreliable content. We propose *MediRely*, a lightweight, reliability-aware retrieval framework for missing-context multimodal CDS. Given a test image without its context, MediRely retrieves visually similar training cases from an image-embedding memory bank and aggregates their paired context embeddings into a similarity-weighted pseudo-modality for the existing multimodal classifier. A neighbour-agreement score estimates retrieval reliability and supports trust gating and selective prediction. On IU X-Ray, using a strict patient-disjoint study-level split, MediRely improves the image-only baseline from 0.722 macro AUC and 0.144 expected calibration error (ECE) to 0.777 macro AUC and 0.056 ECE without retraining the deployed model. With optional cross-modal distillation and reliability gating, it reaches 0.807 macro AUC and 0.018 ECE, comparable to retraining-based baselines while enabling reliability-aware deferral. Zero-shot evaluation on NIH ChestX-ray14 further improves over the image-only baseline, showing preliminary cross-site robustness. MediRely therefore provides an efficient, calibrated, and auditable robustness layer for multimodal CDS when clinical context is unavailable. Code and models will be publicly available at `https://github.com/JianJiaXian/MediRely`.

Keywords: Multimodal learning · Clinical decision support · Missing modality · Retrieval · Reliability · Chest X-ray.

1 Introduction

Clinical decision support (CDS) increasingly relies on *multimodal* models that combine medical images with complementary clinical text, such as radiology reports and clinical notes. While image–text models have shown clear benefits on

paired datasets [13,12,2,16], they typically assume that all modalities are available at inference time. In practice, textual context may be delayed, inaccessible, or unavailable across institutions, causing multimodal models to degrade when the text modality is missing [5]. Reliable inference under missing clinical context is therefore an important deployment challenge.

Existing missing-modality approaches commonly rely on model retraining, architectural modification, or reconstruction of the missing input [11,14,7,4,15,9,17]. These strategies can increase deployment cost, while generative alternatives may introduce unsupported clinical content. We instead explore a retrieval-based alternative: recovering missing context from visually similar training cases using a non-parametric memory [8,6].

We propose **MediRely**, a lightweight reliability-aware retrieval framework for multimodal CDS with missing clinical context. Given a test image without its report, MediRely retrieves visually similar training cases and aggregates their report embeddings into a similarity-weighted *pseudo-report*, which is processed by the existing multimodal classifier. To determine when this retrieved context is trustworthy, MediRely introduces a *neighbour-agreement* signal that, together with retrieval similarity, supports trust gating, fallback, and selective prediction. Because the recovered context comes from real training cases rather than generated text, predictions remain traceable to inspectable clinical evidence.

Contributions.

1. **Training-free retrieval for missing context.** MediRely recovers pseudo-report embeddings from visually similar training cases without retraining the deployed multimodal model or generating synthetic reports.
2. **Reliability-aware pseudo-modality inference.** Neighbour agreement and retrieval confidence are used to control trust mixing, fallback, and selective prediction.
3. **Auditable missing-report evaluation.** We evaluate MediRely on IU X-Ray in terms of predictive performance, calibration, reliability, deferral behavior, and efficiency.

2 Method

MediRely is a lightweight, modular robustness layer for existing multimodal CDS models. It requires no retraining of the deployed model or generative imputation and provides an interpretable retrieval process. The image encoder is interchangeable, supporting a lightweight ResNet or a frozen TorchXRayVision DenseNet-121, with only a small projection layer trained for feature alignment.

2.1 Problem setup and memory bank

Let each training case be denoted by (x_i, r_i, y_i), where x_i is the medical image, r_i is its paired report text, and $y_i \in \{0,1\}^C$ is the multi-label target over C

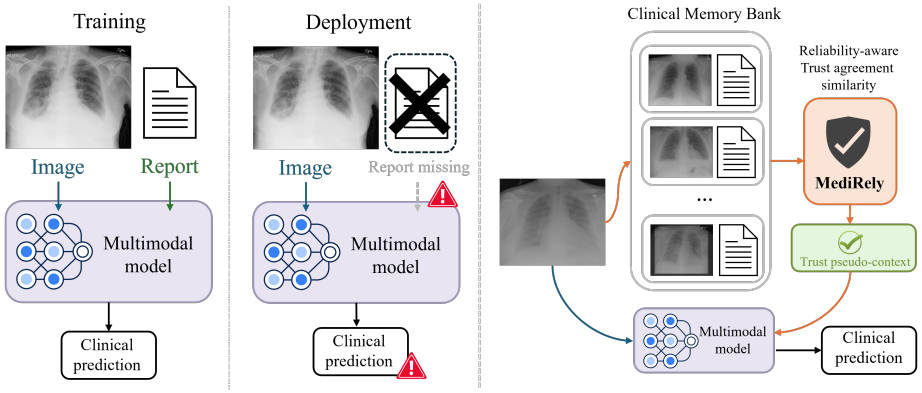

Fig. 1. Concept of MediRely. When the clinical report is missing, visually similar training cases provide reliability-aware pseudo-context for multimodal prediction.

clinical findings. A full multimodal classifier consists of an image encoder f_θ, a text encoder g_ϕ, a fusion module h_ψ, and a classification head c_ω, and maps a case to prediction probabilities as

$$\mathbf{u}_i = f_\theta(x_i) \in \mathbb{R}^{d_v}, \quad \mathbf{v}_i = g_\phi(r_i) \in \mathbb{R}^{d_t}, \quad \hat{\mathbf{y}}_i = c_\omega(h_\psi(\mathbf{u}_i, \mathbf{v}_i)) \in [0,1]^C, \qquad (1)$$

where d_v and d_t denote the dimensions of the image and text representations. At inference, we consider a test image $x^\star$ whose corresponding report $r^\star$ is unavailable. Our goal is to estimate the multi-label prediction $\hat{\mathbf{y}}^\star \in [0,1]^C$ from the available image while recovering as much of the full-modality performance as possible, without retraining the deployed model or generating the missing report $r^\star$.

We first train the full image–text classifier with the multi-label binary cross-entropy (BCE) loss, $\mathcal{L} = \frac{1}{C}\sum_j \mathrm{BCE}(\hat{y}_{ij}, y_{ij})$, using AdamW [10] and a cosine learning-rate schedule. Optionally, modality dropout is applied during training: with probability p_{drop} the report input is replaced by a learned "missing" token, encouraging the fusion module to remain robust when textual context is unavailable. After training, all model parameters are frozen and the N training cases are encoded into a memory bank $\mathcal{M} = \{(\mathbf{u}_i, \mathbf{v}_i, y_i, r_i, \texttt{study}_i, \texttt{path}_i)\}_{i=1}^N$, where $\mathbf{u}_i$ and $\mathbf{v}_i$ are the image and report embeddings, y_i is the multi-label target, r_i is the raw report, $\texttt{study}_i$ identifies the source study, and $\texttt{path}_i$ identifies the source image. The bank is constructed once with a single forward pass over the training set and is the only additional structure required at inference; for IU X-Ray, it occupies approximately 11 MB. The bank is indexed per image: because the study-level split keeps all views of a study together, a retrieved neighbour set may contain multiple views of the same training study, which share one report. We do not deduplicate by study, and study-level retrieval is a straightforward refinement. Throughout, the image-only prediction $\hat{y}^{\mathrm{img}}$ is obtained from the frozen classifier by replacing the report embedding with a fixed "missing" representation. When modality dropout is enabled this representation is a learnable

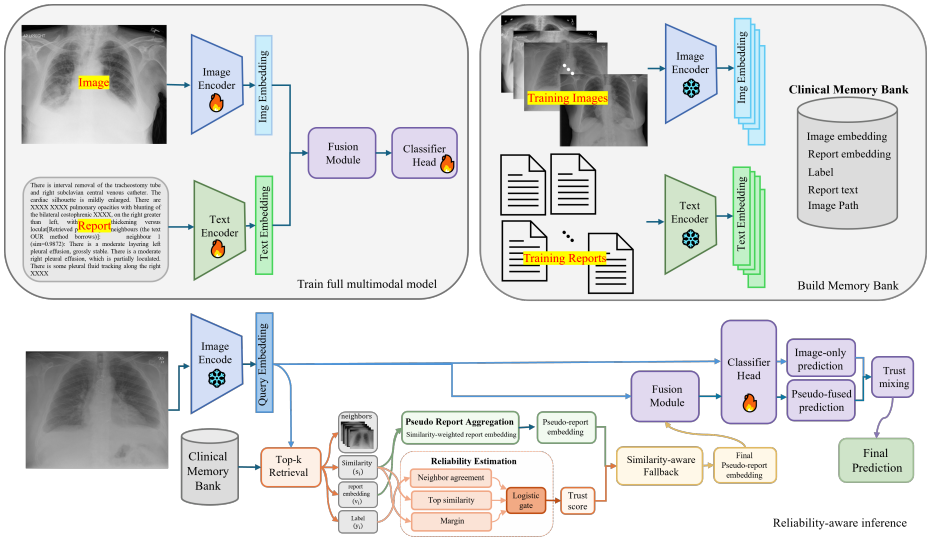

Fig. 2. MediRely architecture: memory-bank construction, pseudo-report retrieval, reliability estimation, and trust-aware prediction.

token; otherwise—as in our drop-in base model, unless stated otherwise—it is a fixed zero text embedding. The naive drop-text baseline uses this same image-only pathway.

2.2 Pseudo-modality retrieval

Given a test image $x^\star$, we compute its image embedding $\mathbf{u}^\star = f_\theta(x^\star)$ and rank the memory-bank images by cosine similarity to it:

$$s_i = \frac{\langle \mathbf{u}^\star, \mathbf{u}_i \rangle}{\|\mathbf{u}^\star\| \, \|\mathbf{u}_i\|}. \tag{2}$$

Let $\mathcal{N}_k$ denote the indices of the top-k most similar training cases, and let $s^\star = \max_i s_i$ denote the top-1 similarity. We then construct a *pseudo-report* embedding by aggregating the report embeddings of the retrieved neighbours, each weighted by a temperature-scaled softmax over its similarity:

$$\alpha_i = \frac{\exp(s_i/\tau_s)}{\sum_{j\in\mathcal{N}_k} \exp(s_j/\tau_s)}, \qquad \tilde{\mathbf{v}}^\star = \sum_{i\in\mathcal{N}_k} \alpha_i \, \mathbf{v}_i, \tag{3}$$

where τ_s is the softmax temperature controlling the concentration of the retrieval weights: a smaller τ_s assigns greater weight to the most similar neighbours, whereas a larger value produces more uniform weights.

Equation (3) estimates the missing report representation from the available image by transferring contextual information from visually similar training cases.

Several imputation baselines used in Sec. 3 can be expressed as alternative choices of the pseudo-report, including the global mean report embedding $\bar{\mathbf{v}}$ (mean-report), the embedding of a randomly selected report (random-report), and $\tilde{\mathbf{v}}^\star = \mathbf{0}$ (zero-text). We additionally evaluate uniform neighbour weighting and Euclidean-distance retrieval.

2.3 Reliability-aware gating and fallback

A pseudo-report is useful only when its retrieved neighbours provide reliable context. We therefore estimate the reliability of the recovered context for each test case and use it to control pseudo-modality fusion, selective prediction, and a hard fallback.

Neighbour-agreement signal. Let $\{y_i\}_{i\in\mathcal{N}_k}$ denote the multi-label targets of the retrieved neighbours, with per-finding empirical frequency $\bar{\mathbf{p}} = \frac{1}{k}\sum_{i\in\mathcal{N}_k} y_i \in [0,1]^C$. We define the neighbour-agreement score as

$$a^\star = \frac{1}{C}\sum_{c=1}^{C} \max\big(\bar{p}_c,\ 1-\bar{p}_c\big) \in [0.5, 1]. \tag{4}$$

The score $a^\star$ is high when the retrieved neighbours consistently agree on the presence or absence of each finding, and decreases as their labels become more inconsistent. This provides information that the top similarity $s^\star$ alone cannot capture, since highly similar neighbours may still disagree clinically. Because thoracic findings are sparse, $a^\star$ partly reflects agreement on *negative* labels; class-balanced or finding-specific agreement is a natural refinement. We use three reliability features: the top-1 similarity $s^\star$, the similarity margin $s_1 - s_k$ (where s_1 and s_k are the highest and k-th highest retrieved similarities), and the neighbour-agreement score $a^\star$.

Soft trust gate. We combine these reliability features using a lightweight logistic gate,

$$\gamma^\star = \sigma\big(\mathbf{w}^\top[\, s^\star,\ s_1 - s_k,\ a^\star\,] + b\big) \in [0,1], \tag{5}$$

where $\gamma^\star$ estimates whether incorporating the retrieved pseudo-context is likely to improve the prediction. The parameters $(\mathbf{w}, b)$ are fitted on the validation set using the binary target $\mathbb{1}[\,\ell(\hat{y}^{\text{pseudo}}, y) < \ell(\hat{y}^{\text{img}}, y)\,]$, which indicates whether the pseudo-report reduces the prediction loss relative to the image-only pathway; the pretrained multimodal model remains frozen during this calibration step. The resulting trust score then controls a convex combination of the image-only and pseudo-fused predictions:

$$\hat{y}^{\text{soft}} = (1-\gamma^\star)\hat{y}^{\text{img}} + \gamma^\star \hat{y}^{\text{pseudo}}. \tag{6}$$

A high $\gamma^\star$ places greater trust in the retrieved pseudo-context, whereas a low value shifts the prediction toward the image-only pathway.

Selective prediction. For deployment, reliability can also drive selective prediction. We use the neighbour-agreement score $a^\star$ as the primary deferral criterion, separately from the learned gate $\gamma^\star$ used for soft prediction mixing: test cases are ranked by $a^\star$ and the least reliable fraction is deferred for additional clinical evidence rather than forcing a prediction, giving a controllable coverage–reliability trade-off. In our experiments $a^\star$ is an effective and computationally inexpensive deferral criterion (Sec. 3).

Similarity-aware fallback. When retrieval confidence is low, the soft mechanism reduces to a simple hard fallback that gates the pseudo-report on the top-1 similarity $s^\star$ and a threshold τ:

$$\mathbf{v}^\star = \begin{cases} \tilde{\mathbf{v}}^\star, & s^\star \geq \tau \quad \text{(retrieve)}, \\ \mathbf{v}_{\text{safe}}, & s^\star < \tau \quad \text{(fallback)}, \end{cases} \tag{7}$$

where $\mathbf{v}_{\text{safe}}$ is a conservative alternative, such as the learned missing token, a zero vector, or the global mean report embedding, and the final prediction reuses the frozen classifier, $\hat{\mathbf{y}}^\star = c_\omega(h_\psi(\mathbf{u}^\star, \mathbf{v}^\star))$. This limits the influence of poorly matched retrieved cases by reverting to a conservative representation when retrieval confidence is insufficient (Fig. 2).

2.4 Optional: cross-modal distillation to close the gap

The training-free bridge can be pushed closer to the full model with a single additional training run. We treat the trained image–text model as a frozen teacher with access to the real report and train a student that consumes the image and the retrieved pseudo-report, supervised by the labels, the teacher's softened predictions, and its fused representation:

$$\mathcal{L}_{\text{distill}} = \text{BCE}(\hat{y}^s, y) + \lambda\, T^2\, \text{BCE}\big(\sigma(q^s/T),\ \sigma(q^t/T)\big) + \mu\, \|\mathbf{z}^s - \mathbf{z}^t\|_2^2, \tag{8}$$

where q^t and q^s are the teacher and student logits, $\hat{y}^t = \sigma(q^t)$ and $\hat{y}^s = \sigma(q^s)$ the corresponding predictions, and $\mathbf{z}^t$ and $\mathbf{z}^s$ their fused representations. The temperature T controls the softness of the teacher targets, and λ and μ balance prediction distillation and feature alignment, respectively.

During student training, retrieval is performed in a leave-one-*study*-out manner: all memory-bank images belonging to the query study are excluded, so a view cannot retrieve another view paired with the same report. This removes study-level leakage and matches the retrieval setting used at inference, where the query study is never in the bank. The student thus learns to compensate for imperfect pseudo-context using guidance from the full-modality teacher. This stage requires one additional training run but is optional; the core retrieval bridge remains applicable without retraining the deployed model, and at inference the student requires only image encoding, retrieval, and multimodal fusion.

3 Experiments and Results

3.1 Experimental setup

Datasets. We use IU X-Ray with eight findings derived from human-curated MeSH annotations: cardiomegaly, effusion, atelectasis, pneumonia, edema, pneumothorax, consolidation, and no finding. A fixed 70%/10%/20% patient-disjoint split contains 5,217/736/1,517 images from 2,695/385/771 studies for training/validation/test. All views sharing a report remain in the same split, preventing report leakage through the memory bank. For zero-shot cross-site evaluation, we use 2,459 images from the official patient-level test partition of NIH ChestX-ray14 [16], restricted to the first two public image archives, and evaluate the seven findings shared with IU. No NIH data are used for training, model selection, calibration, or hyperparameter tuning.

Implementation. Images are resized to 224×224. We use a frozen TorchXRayVision DenseNet-121 [1] with 256-dimensional image and text representations; ResNet-18 [3] is used for encoder ablation. Models are trained for 20 epochs using AdamW (learning rate 3×10^{-4}, weight decay 10^{-4}, batch size 32), multi-label BCE, and validation macro-AUC selection. Retrieval uses cosine similarity, $k = 10$, $\tau_s = 0.1$, and fallback threshold $\tau = 0.5$. Optional distillation uses $T = 2$, $\lambda = 1$, and $\mu = 10$ with leave-one-study-out retrieval. We report macro AUC and ECE, with ECE computed using 15 equal-width bins and confidence intervals obtained from 2,000 paired study-level bootstrap samples. Baselines include drop-text, mean- and random-report imputation, modality dropout [11], and privileged knowledge distillation [15,14]; the real-report model serves only as an oracle.

3.2 Comparison with missing-report methods

Table 1 compares missing-report methods according to whether retraining is required. In the *drop-in* setting, MediRely improves the image-only baseline from 0.722 macro AUC and 0.144 ECE to 0.777 macro AUC and 0.056 ECE without retraining the deployed model. It also outperforms mean-report imputation (0.738 AUC, 0.071 ECE), showing that visually grounded retrieval provides more useful context than query-independent substitution.

When retraining is allowed, modality dropout [11], privileged knowledge distillation [15,14], and MediRely with optional distillation achieve similar performance around 0.80 macro AUC. MediRely reaches 0.807 AUC, compared with 0.810 for privileged KD, with no significant difference ($\Delta = -0.003$, 95% paired study-level bootstrap CI $[-0.007, +0.002]$, 2000 resamples). It also achieves lower ECE (0.018 vs. 0.028) while supporting reliability-aware selective deferral, allowing uncertain cases to be routed for additional clinical evidence.

The oracle model using the real report reaches 0.970 macro AUC and is included only as an upper reference. Because the report may explicitly contain the

Table 1. Missing-report performance on IU X-Ray. Results are grouped by whether retraining is required; the real-report model is shown as an oracle.

Method	Macro AUC	ECE↓	Retrain	Defer
Oracle (reads real report)	0.970	0.006	–	–
Drop-in: no retraining of the deployed model				
Drop text (naive)	0.722	0.144	–	–
Mean-report imputation	0.738	0.071	–	–
MediRely (ours, training-free)	**0.777**	0.056	–	✓
Requires retraining the model				
Modality dropout [11]	0.801	0.007	✓	–
Privileged KD [15,14]	**0.810**	0.028	✓	–
MediRely + distill (ours, optional)	0.807	0.018	✓	✓

target findings, it is not a fair missing-modality target; the more relevant objective is to recover the image-only performance gap while maintaining calibration and reliability.

Finally, directly transferring retrieved neighbours' labels as a kNN prior underperforms report-embedding transfer (0.679 vs. 0.777 macro AUC), suggesting that semantic report embeddings provide more informative pseudo-context than hard neighbour labels.

3.3 Module ablation and reliability

Table 2 ablates three components of MediRely: pseudo-modality retrieval (R), cross-modal distillation (D), and reliability gating (G). In the training-free setting, retrieval improves macro AUC from 0.722 to 0.777, while adding the reliability gate reduces ECE from 0.056 to 0.024 with only a small decrease in AUC. When retraining is allowed, distillation alone achieves 0.808 macro AUC and 0.008 ECE, while the full model combining retrieval, distillation, and gating reaches 0.807 macro AUC and 0.018 ECE. These results show that the components serve different roles across deployment settings: retrieval recovers missing context without retraining, distillation further improves predictive performance, and gating substantially improves calibration in the training-free setting—though it does not further reduce ECE once distillation is applied—while enabling selective prediction.

The learned gate further suggests that neighbour agreement is a useful reliability signal beyond raw similarity: the predicted trust score closely follows the observed probability that pseudo-context improves the prediction (e.g. 0.72 predicted vs. 0.76 observed). This signal also supports selective prediction: abstaining on cases with the lowest neighbour agreement reduces selective BCE risk from 0.255 at full coverage to 0.221 at 50% coverage and 0.200 at 10% coverage, outperforming selection based on similarity or model confidence alone.

Table 2. Ablation of retrieval (R), distillation (D), and reliability gating (G) on IU X-Ray.

R	D	G	Macro AUC	ECE↓	
–	–	–	0.722	0.144	
✓	–	–	0.777	0.056	
✓	–	✓	0.772	0.024	
–	✓	–	**0.808**	**0.008**	
✓	✓	–	0.798	0.029	
✓	✓	✓	**0.807**	0.018	**(full, ours)**

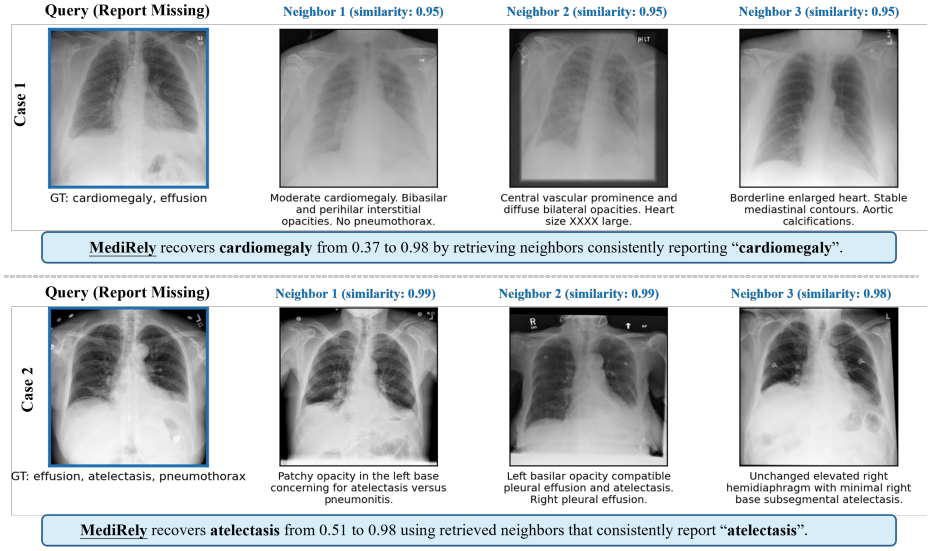

Fig. 3. Neighbour-driven recovery on IU X-Ray (two selected successful cases). Retrieved reports provide pseudo-context for findings missed by the image-only pathway.

3.4 Qualitative recovery and generalization

Figure 3 shows two selected IU X-Ray test cases in which MediRely recovers cardiomegaly and atelectasis missed by the image-only pathway. The retrieved neighbours exhibit similar CXR appearances and consistently mention the missed findings, shifting predictions toward the full image–text oracle and illustrating the value of neighbour agreement.

MediRely remains robust across missing-report ratios, consistently outperforming the image-only baseline, including 0.816 vs. 0.776 macro AUC at 75% missing. Performance varies by less than ± 0.01 across $k \in \{1, 3, 5, 10\}$, weighting strategies, and distance metrics. The framework is also encoder-agnostic and efficient: replacing ResNet-18 with a frozen CXR DenseNet-121 improves macro AUC from 0.733 to 0.777, while the memory bank requires only $\sim$11 MB and re-

Table 3. Zero-shot cross-site evaluation on NIH ChestX-ray14 using the IU-trained model and memory bank.

Method (NIH, zero-shot)	Macro AUC	ECE↓
Image-only (report missing)	0.639	0.099
Random-report (control)	0.522	0.120
MediRely (retrieval)	**0.648**	**0.084**

trieval takes ∼0.4 ms/query. Total latency remains close to image-only inference (0.44 vs. 0.25 ms/sample).

For zero-shot cross-site evaluation on 2,459 NIH ChestX-ray14 images, MediRely improves macro AUC from 0.639 to 0.648 and reduces ECE from 0.099 to 0.084 over the seven findings shared with IU (Table 3). Random-report imputation instead reduces macro AUC to 0.522, while retrieval performance increases with k (0.603, 0.626, 0.637, and 0.648 for $k = 1, 3, 5, 10$). These results provide preliminary evidence that retrieval-based pseudo-context remains useful under cross-site distribution shift.

4 Conclusion

We presented *MediRely*, a reliability-aware pseudo-modality bridge for multimodal clinical decision support when clinical reports are unavailable. Instead of generating missing text, MediRely retrieves visually similar training cases, aggregates their report embeddings as pseudo-context, and reuses the existing multimodal classifier without retraining the deployed model. On IU X-Ray, it improves macro AUC from 0.722 to 0.777 and reduces ECE from 0.144 to 0.056 in the drop-in setting under a strict patient-disjoint split. MediRely further estimates the reliability of retrieved context using neighbour agreement and retrieval confidence, enabling trust-aware fusion, selective prediction, and fallback when retrieval is unreliable. With optional distillation and reliability gating, it reaches 0.807 macro AUC and 0.018 ECE, remaining competitive with retraining-based baselines while supporting reliability-aware deferral. Because the pseudo-context is derived from real training cases rather than generated text, predictions remain traceable to inspectable neighbours. Our findings also emphasize the need for careful evaluation in report-based multimodal CDS: because reports may explicitly contain target findings, the full image–text model should be regarded as an oracle reference rather than a fair missing-modality target. Overall, MediRely provides a practical, calibrated, and auditable robustness layer for multimodal CDS when clinical context is unavailable.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cohen, J.P., Viviano, J.D., Bertin, P., et al.: Torchxrayvision: A library of chest x-ray datasets and models. In: Medical Imaging with Deep Learning (MIDL) (2022)
2. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., et al.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association (JAMIA) **23**(2), 304–310 (2016)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
4. Hu, K., et al.: Learning missing modal electronic health records with unified multimodal data embedding and modality-aware attention. In: Machine Learning for Healthcare Conference (MLHC) (2023)
5. Jian, J.X., Lin, S.C., Sun, W.C.: Robust pediatric brain tumor segmentation: Evaluating modality importance and resilience in missing-modality scenarios. In: IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–4 (2026). https://doi.org/10.1109/ISBI61048.2026.11515316
6. Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., Lewis, M.: Generalization through memorization: Nearest neighbor language models. In: International Conference on Learning Representations (ICLR) (2020)
7. Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y.: Multimodal prompting with missing modalities for visual recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14943–14952 (2023)
8. Lewis, P., Perez, E., Piktus, A., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems (NeurIPS) **33**, 9459–9474 (2020)
9. Li, M., Yang, D., Zhao, X., Wang, S., Wang, Y., Yang, K., Sun, M., Kou, D., Qian, Z., Zhang, L.: Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12458–12468 (2024)
10. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)
11. Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: Smil: Multimodal learning with severely missing modality. In: AAAI Conference on Artificial Intelligence. pp. 2302–2310 (2021)
12. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: AAAI Conference on Artificial Intelligence (2018)
13. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
14. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: Multi-modal learning with missing modality via shared-specific feature modelling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15878–15887 (2023)
15. Wang, S., Yan, Z., Zhang, D., et al.: Prototype knowledge distillation for medical segmentation with missing modality. arXiv preprint arXiv:2303.09830 (2023)
16. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2097–2106 (2017)

17. Zhu, S., Chen, Y., Chen, W., Wang, Y., Liu, C., Jiang, S., Qin, F., Wang, C.: Bridging the gap in missing modalities: Leveraging knowledge distillation and style matching for brain tumor segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2025)