PaCX-MAE: Physiology-Augmented Chest X-Ray Masked Autoencoder

Yancheng Liu [1] **Kenichi Maeda** [1] **Manan Pancholy** [1]

Abstract

Clinical diagnosis often requires combining imaging with physiological measurements, yet deployed models typically operate on unimodal data. We present **PaCX-MAE**, a cross-modal distillation framework that injects physiological priors into chest X-ray (CXR) encoders while remaining strictly unimodal at inference. PaCX-MAE augments in-domain masked autoencoding with a dual contrastive-predictive objective, aligning CXR representations with paired ECG and laboratory embeddings. Extensive evaluation across nine benchmarks demonstrates consistent improvements over domain-specific MAE, particularly on physiology-dependent tasks (e.g., **+2.7 AUROC** on MedMod; **+6.5 F1** on VinDr-CXR). The method proves highly label-efficient in the **1%** regime and preserves anatomical fidelity, achieving parity with MAE on segmentation tasks. Zero-shot and attention analyses confirm that PaCX-MAE successfully learns to attend to physiological indicators, such as the cardiac silhouette, absent in standard visual pretraining.

1. Introduction

Self-supervised learning on medical imaging typically considers modalities in isolation or assumes full modality availability at inference. While multimodal models can leverage physiological signals (e.g., ECG, PPG, labs) to represent dynamic states, such as cardiac electrical activity or fluid status, clinical realities often force a disconnect between training and deployment. In acute settings, a Chest X-ray is frequently the only immediate acquisition. Consequently, models trained on rich multimodal data fail to deploy because the auxiliary modalities are missing, while unimodal models fail to capture the systemic physiological context invisible to the naked eye.

[1]Department of Computer Science, Brown University, Providence, RI, USA. Correspondence to: Yancheng Liu <yancheng_liu@brown.edu>.

Assignment created by Randall Balestriero for CSCI2952X: Research Topics in Self-Supervised Learning, offered in Fall 2025.

This raises a critical question: *Can a model learn to hallucinate physiological context from anatomical data alone?* We hypothesize that by explicitly aligning visual features with physiological encodings during training, a vision encoder can learn to recognize the subtle anatomical correlates of systemic physiology.

To bridge this gap, we introduce **PaCX-MAE**, a framework that distills physiological priors into a strictly unimodal vision encoder. Unlike standard multimodal fusion, PaCX-MAE uses paired data only during pretraining to align CXR representations with physiological embeddings via a dual contrastive-predictive objective. This allows the model to internalize systemic context without requiring auxiliary modalities at inference. We validate PaCX-MAE across nine benchmarks, showing that it: (1) outperforms standard unimodal MAE, particularly on physiology-heavy tasks (e.g., MedMod); (2) maintains pixel-level fidelity for segmentation; and (3) significantly improves label efficiency in low-data regimes.

2. Related Work

SSL for Medical Imaging Self-supervised learning in medical imaging has largely relied on contrastive methods, which learn invariance by maximizing agreement between augmented views (Azizi et al., 2021; Cho et al., 2023; Gorade et al., 2025). However, recent work indicates that Masked Autoencoders (MAE) often yield superior representations for downstream diagnosis by forcing the reconstruction of fine-grained anatomical details rather than global invariants (Xiao et al., 2023; Zhou et al., 2023; Gupta et al., 2024). **PaCX-MAE** takes advantage of this reconstructive strength, initializing with an MAE backbone to ensure robust anatomical features before injecting physiological priors.

Multimodal Distillation & Missing Modalities While standard multimodal learning fuses data streams (e.g., imaging, text, continuous waveforms) during both training and inference (Zhang et al., 2022; Radford et al., 2021), clinical deployment is often constrained by missing modalities. To address this, cross-modal knowledge distillation transfers "privileged information" from a multimodal teacher to a unimodal student (Lopez-Paz et al., 2016; Gupta et al., 2015). In medical domains, this paradigm has been applied to distill radiology reports (Tiu et al., 2022; Boecking et al., 2022) or

missing MRI sequences (Dou et al., 2020; Wang et al., 2023) into image-only encoders. **PaCX-MAE** extends this framework to physiology, distilling dense signals (ECG, labs) into CXR representations to enable physiological reasoning at inference time without requiring auxiliary sensors.

3. Methodology

We introduce **PaCX-MAE**, a framework for distilling physiological priors into a unimodal vision backbone. As illustrated in Figure 1, our approach decouples *representation learning* from *cross-modal alignment* via a two-stage curriculum: (1) independent unimodal pretraining to establish robust feature spaces, followed by (2) cross-modal distillation that aligns the visual encoder with frozen physiological targets via a dual contrastive-predictive objective.

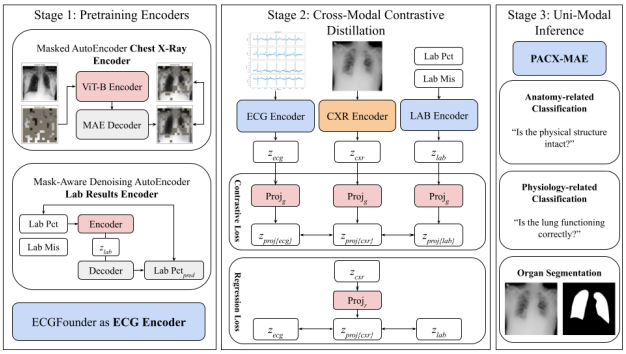

Figure 1. **Overview of the PaCX architecture.** The pipeline comprises unimodal pretraining (Stage 1) and cross-modal distillation (Stage 2). Colors indicate optimization status: red (trainable), orange (LoRA-adapted), and blue (frozen). During distillation, the CXR encoder learns to predict physiological embeddings via lightweight heads, which are discarded at inference.

3.1. Stage 1: Unimodal Pretraining

We utilize the Symile-MIMIC dataset (Saporta et al., 2025), containing $N \approx 10$k paired triplets of CXR (x_v), ECG (x_e), and Laboratory (x_l) data. To mitigate the risk of overfitting to limited paired data, we first initialize modality-specific encoders on large-scale external datasets.

Vision Encoder (CXR). We initialize the vision backbone (f_v) using a ViT-B architecture trained via Masked Autoencoding (MAE) on CheXpert (Irvin et al., 2019). We employ an aggressive masking ratio of 0.90, as it forces the model to infer global anatomical semantics (e.g., cardiac silhouette, mediastinal width) from limited visual cues rather than exploiting local texture shortcuts (Gupta et al., 2024). We prioritize this reconstructive objective over contrastive alternatives (e.g., MoCo, DINO). Unlike contrastive methods, which enforce invariance to augmentations that can obscure subtle intensity cues (e.g., fluid opacity) (Huang et al., 2023), MAE preserves fine-grained details critical for physiological inference and avoids the semantic collapse often associated with false-negative sampling in medical imaging (see Fig. 2). Implementation details, including the specific optimizer and scheduler configurations, are provided in Appendix A.1.1.

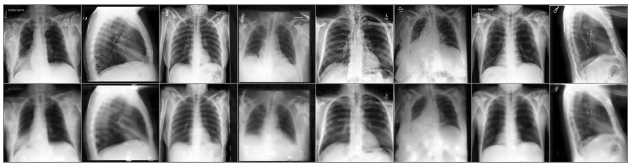

Figure 2. **MAE Pretraining Reconstructions.** Top: Original CXRs; Bottom: Reconstructions under 90% masking. Despite extreme sparsity, the model accurately recovers key physiological indicators such as the cardiac boundary and diaphragm curvature.

Physiological Targets (ECG & Labs). We employ high-fidelity, frozen encoders to serve as distillation targets. For Laboratory data, we pretrain a mask-aware Denoising Autoencoder that models both measured values and structured missingness typical of tabular clinical records. For ECG, we utilize **ECGFounder** (Li et al., 2025), a transformer pretrained on 10 million recordings to capture high-frequency morphological patterns. Crucially, these encoders (f_e, f_l) remain **frozen** during Stage 2. This prevents degenerate co-adaptation and ensures that physiological structure is transferred into the visual manifold rather than jointly relearned. Architectural and training details are provided in Appendix A.1.1.

3.2. Stage 2: Cross-Modal Distillation

The objective of Stage 2 is to inject physiological priors into the visual backbone f_v without degrading its anatomical fidelity. We employ a dual-branch architecture that aligns the CXR embeddings with the frozen physiological targets (f_e, f_l) via a hybrid contrastive-regression loss.

Parameter-Efficient Adaptation (LoRA). To prevent catastrophic forgetting of the dense anatomical priors learned during MAE pretraining, we freeze the majority of the visual backbone, keeping only the normalization layers trainable to stabilize feature distributions. We inject Low-Rank Adaptation (LoRA) matrices into both the attention ('qkv') and feed-forward ('fc') modules. This ensures that the core visual manifold remains stable, while the lightweight LoRA parameters ($< 1\%$ of total weights) learn the specific projections required to bridge the modality gap.

Dual Distillation Objective. We define a hybrid objective $\mathcal{L}_{total} = \lambda_C \mathcal{L}_{contrastive} + \lambda_R \mathcal{L}_{regression}$ to balance global semantic alignment with latent feature reconstruction.

1. **Global Contrastive Alignment ($\mathcal{L}_C$):** We project CXR, ECG, and Lab embeddings into a shared latent space and align them via a symmetric InfoNCE loss with a learnable temperature parameter. To prevent the model from exploiting local batch biases, we employ **global negative sampling**, gathering embeddings across all dis-

tributed GPUs to maximize the diversity of negative pairs. We further apply label smoothing ($\epsilon = 0.02$) to prevent overfitting to noisy medical labels.

2. **Latent Regression ($\mathcal{L}_R$):** Contrastive learning alone can settle for "shortcut" features sufficient for discrimination but insufficient for detailed reasoning. To counter this, we introduce modality-specific regression heads that predict the *exact* unprojected frozen embedding vectors of the physiological encoders. We minimize the **Cosine Distance** ($1 - \cos(\hat{y}, y)$) between the predicted and target embeddings, forcing the visual encoder to internalize the dense semantic structure of the physiological signals.

Both the projection (contrastive) and regression heads are discarded at inference, leaving only the enhanced visual encoder. Implementation details are provided in Appendix A.1.2.

4. Experiments

Datasets and Baselines. We evaluate PaCX-MAE across a comprehensive suite of 9 public benchmarks covering binary, multiclass, and multilabel classification, as well as semantic segmentation tasks. Detailed descriptions and statistics for each dataset are provided in Appendix A.2.1. We compare our method against two primary baselines: a standard **ImageNet-pretrained** ViT-B/16 and a domain-specific **Masked Autoencoder (MAE)** pretrained on unimodal CXRs. Since PaCX utilizes the same backbone architecture as the MAE baseline, any performance variance directly isolates the efficacy of our cross-modal physiological distillation.

Evaluation Protocols. Our evaluation strategy assesses representation quality through **linear probing**, where a linear classifier is trained on top of the frozen encoder backbone. We rigorously test **data efficiency** by training on restricted subsets (e.g., 1%, 10%) of the available training data. To deconstruct the impact of specific physiological signals, we conduct **modality ablations** and loss component analyses. Furthermore, we evaluate the alignment of learned representations via **zero-shot retrieval** metrics (Recall@K, Cosine Similarity) and provide qualitative interpretations using **Attention Rollout**. Complete implementation details, including hyperparameters, optimization schedules, and ablation protocols, are provided in Appendix A.2.2.

5. Results

5.1. Clinical Transfer & Data Efficiency

Domain-Specific Pretraining. Table 1 confirms that domain-specific pretraining consistently outperforms ImageNet initialization. Crucially, PaCX retains the dense anatomical competence of the strong MAE baseline, achieving parity on pixel-precise segmentation benchmarks like **CXL-Seg** (**0.996** IoU) and **COVID-QU-Ex** (**0.942** IoU), demonstrating that physiological distillation does not induce catastrophic forgetting of structural features.

Physiological Awareness. PaCX demonstrates superior transfer to tasks reflecting latent physiological states rather than purely visible structures. On physiology-dense benchmarks—**CheXchoNet** (Fluid Overload), **VinDr-CXR**, and **MedMod**—it significantly outperforms the unimodal MAE, achieving gains of **+2.4%** AUROC and **+6.5%** F1 on VinDr-CXR, **+2.7%** AUROC on MedMod, and **+5.1%** F1 on CheXchoNet. This supports our hypothesis that distilled "phantom" physiological signals enhance diagnosis where visual cues are subtle.

Dataset	Metric	ImageNet	MAE	PACX-MAE
TB[1]	AUROC	0.887	0.899	**0.910**
	F1	0.818	0.814	**0.846**
CheXchoNet[1]	AUROC	0.728	0.788	**0.803**
	F1	0.147	0.215	**0.266**
ChestX6[2]	AUROC	0.983	0.988	**0.989**
	F1	0.876	0.905	**0.906**
VinDr-CXR[3]	AUROC	0.751	0.847	**0.871**
	F1	0.097	0.191	**0.256**
NIH-14[3]	AUROC	0.721	0.772	**0.783**
	F1	0.048	0.113	**0.115**
MedMod[3]	AUROC	0.612	0.695	**0.722**
	F1	0.091	0.231	**0.253**
COVID-QU-Ex[4]	IoU	0.894	**0.942**	**0.942**
	Dice	0.943	**0.970**	**0.970**
QaTa-COV19[4]	IoU	0.622	**0.726**	0.715
	Dice	0.766	**0.841**	0.833
CXL-Seg[4]	IoU	0.984	**0.996**	**0.996**
	Dice	0.992	**0.998**	**0.998**

Table 1. **Comparison of classification and segmentation performance with different pretraining methods.** Superscripts denote task type: [1] binary, [2] multiclass, [3] multilabel, [4] segmentation.

Low-Data Efficiency. PaCX significantly lowers sample complexity. As illustrated in Figure 3, the performance gap is widest in extremely low-data regimes. In the **1% data regime**, PaCX consistently surpasses the MAE baseline, showing AUROC improvements of **+8.2%** on CheXchoNet, and ∼**+5%** on MedMod and VinDr. Even at **10% data**, it maintains a robust lead of **+1.7%–3.8%** across benchmarks. This consistent advantage indicates that learned physiological priors act as effective regularizers, enabling generalization even when visual examples are scarce.

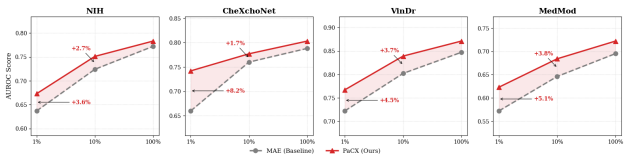

Figure 3. **Label Efficiency.** PaCX (red) outperforms MAE (grey) consistently at 1% and 10% training data, demonstrating robust few-shot transfer.

5.2. Physiological Alignment & Interpretability

Zero-Shot Alignment. Table 2 confirms that PaCX effectively distills physiological signals, achieving superior structural alignment between frozen CXR embeddings and ground-truth targets. PaCX surpasses the MAE baseline in Cosine Similarity across both modalities (**0.229** ECG, **0.252** Labs), indicating it successfully pulls anatomically distinct X-rays closer to their correct physiological profiles in the latent space without explicit supervision.

	ECG Targets			Lab Targets		
Metric	**ImNet**	**MAE**	**PaCX**	**ImNet**	**MAE**	**PaCX**
Cos Sim	0.143	0.204	**0.229**	0.187	0.239	**0.252**
R@5	1.51%	5.17%	**5.60%**	1.51%	3.66%	**3.66%**

Table 2. **Zero-Shot Alignment.** PaCX improves latent structure across both modalities, demonstrating superior or equivalent retrieval capability (R@5) and vector alignment (Cosine Similarity).

Attention Rollout. We visualize the impact of this alignment using Attention Rollout on validation "rescue cases"—instances where PaCX correctly classified pathology that the baseline missed. Figure 4 illustrates a representative case of cardiomegaly. While the MAE baseline scatters attention over bony structures like clavicles, PaCX tightens focus on the cardiac silhouette and mediastinum, as highlighted by the red regions in the Difference Map. This confirms that physiological signals effectively guide the visual encoder toward clinically relevant soft-tissue anatomy. Additional cases in Appendix A.3.1 confirm this is a systematic trend, with PaCX consistently exhibiting more focused attention on relevant organ systems.

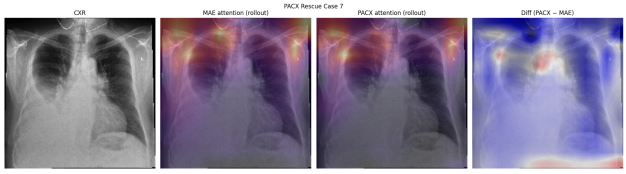

Figure 4. **Attention Shift.** PaCX (middle-right) shifts focus from bony structures to the cardiac silhouette (red in Difference Map).

5.3. Component Analysis

To disentangle the contributions of specific signals and objectives, we analyze component-wise performance on three physiology-dense benchmarks (Table 3).

Modality Synergy. The exclusion of either physiological source degrades overall performance, confirming that ECG and Laboratory data provide complementary rather than redundant signals. For instance, removing Lab priors causes a notable drop in F1 on MedMod ($0.253 \rightarrow 0.243$) and VinDr ($0.256 \rightarrow 0.233$), while the full PaCX configuration generally yields the most robust balance across metrics.

Objective Function. Comparing loss components reveals that regression alone ($\mathcal{L}_{reg}$) is insufficient for robust representation learning, significantly underperforming the contrastive objective ($\mathcal{L}_{cont}$). However, the hybrid objective ($\mathcal{L}_{cont} + \mathcal{L}_{reg}$) achieves the best overall stability, particularly in F1 scores (e.g., VinDr F1 $+1.5\%$ over $\mathcal{L}_{cont}$ alone), validating the design choice to combine global structural alignment with dense continuous regression.

		Modality Ablation			Loss Ablation		
Dataset	**Metric**	**ECG**	**Lab**	**PaCX**	**Cont**	**Reg**	**PaCX**
CheXchoNet	AUC	0.801	0.795	**0.803**	0.799	0.789	**0.803**
	F1	**0.296**	0.275	0.266	**0.273**	0.227	0.266
MedMod	AUC	0.717	0.721	**0.722**	**0.722**	0.673	**0.722**
	F1	0.243	0.245	**0.253**	**0.258**	0.131	0.253
VinDr	AUC	0.871	**0.875**	0.871	0.866	0.843	0.871
	F1	0.233	0.248	**0.256**	0.241	0.130	**0.256**

Table 3. **Unified Ablation Study.** Comparing single-modality sources (Left) and isolated loss components (Right). The full PaCX configuration yields the most consistent performance across tasks.

6. Conclusion & Discussion

In this work, we introduced **PaCX-MAE**, a framework for distilling latent physiological signals captured via ECG and laboratory values into a frozen visual encoder. By implementing a multimodal contrastive and regression-based objective, we demonstrated that "phantom" physiological priors can be effectively embedded into standard chest X-ray representations without requiring paired data at inference time. Our extensive evaluation across 9 benchmarks reveals that PaCX not only matches state-of-the-art domain-specific baselines on structural tasks like segmentation but significantly outperforms them on physiology-dense diagnostic tasks. Notably, these gains are most prominent in low-data regimes, suggesting that physiological grounding acts as a powerful regularizer, guiding the model to attend to clinically relevant soft-tissue structures rather than spurious correlations.

Limitations include our reliance on single-center data (MIMIC-IV), which restricts phenotypic diversity; multi-center validation is essential to ensure these priors generalize across populations. Additionally, our global alignment strategy overlooks dense, region-specific correlations, such as mapping specific ECG waveforms to localized cardiac sub-regions. Future work will explore granular token-level distillation and longitudinal modeling to capture disease progression and anatomical nuance more holistically.

7. Code Availability

The code for PaCX-MAE is available on GitHub at https: https://github.com/Lyce24/PACX-MAE/. This repository includes the complete implementation and scripts needed to reproduce our results.

References

Adel, M. Chestx6: Multi-class x-ray dataset. Kaggle, 2025.

Ahishali, M., Degerli, A., Yamac, M., Kiranyaz, S., Chowdhury, M. E. H., Hameed, K., Hamid, T., Mazhar, R., and Gabbouj, M. Advance warning methodologies for covid-19 using chest x-ray images. *IEEE Access*, 9:41052–41065, 2021. doi: 10.1109/ACCESS.2021.3064927.

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., and Norouzi, M. Big self-supervised models advance medical image classification, 2021. URL https://arxiv.org/abs/2101.05224.

Bhave, S., Rodriguez, V., Poterucha, T., Mutasa, S., Aberle, D., Capaccione, K. M., Chen, Y., Dsouza, B., Dumeer, S., Goldstein, J., Hodes, A., Leb, J., Lungren, M., Miller, M., Monoky, D., Navot, B., Wattamwar, K., Wattamwar, A., Clerkin, K., Ouyang, D., Ashley, E., Topkara, V. K., Maurer, M., Einstein, A. J., Uriel, N., Homma, S., Schwartz, A., Jaramillo, D., Perotte, A. J., and Elias, P. Deep learning to detect left ventricular structural abnormalities in chest x-rays. *European Heart Journal*, 45(22): 2002–2012, 2024.

Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., Poon, H., and Oktay, O. *Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing*, pp. 1–21. Springer Nature Switzerland, 2022. ISBN 9783031200595. doi: 10.1007/978-3-031-20059-5_1. URL http://dx.doi.org/10.1007/978-3-031-20059-5_1.

Candemir, S., Jaeger, S., Palaniappan, K., Musco, J. P., Singh, R. K., Xue, Z., Karargyris, A., Antani, S., Thoma, G., and McDonald, C. J. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33(2): 577–590, 2014. doi: 10.1109/TMI.2013.2290491.

Cho, K., Kim, K. D., Nam, Y., Jeong, J., Kim, J., Choi, C., Lee, S., Lee, J. S., Woo, S., Hong, G.-S., Seo, J. B., and Kim, N. Chess: Chest x-ray pre-trained model via self-supervised contrastive learning. *Journal of Digital Imaging*, 36:902–910, 2023.

Chowdhury, M. E. H., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., Islam, K. R., Khan, M. S., Iqbal, A., Emadi, N. A., Reaz, M. B. I., and Islam, M. T. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020. doi: 10.1109/ACCESS.2020.3010287.

Degerli, A., Ahishali, M., Kiranyaz, S., Chowdhury, M. E. H., and Gabbouj, M. Reliable covid-19 detection using chest x-ray images. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 185–189, 2021a. doi: 10.1109/ICIP42928.2021.9506442.

Degerli, A., Ahishali, M., Yamac, M., Kiranyaz, S., Chowdhury, M. E. H., Hameed, K., Hamid, T., Mazhar, R., and Gabbouj, M. Covid-19 infection map generation and detection from chest x-ray images. *ealth Inf Sci Syst.*, 15, 2021b.

Degerli, A., Kiranyaz, S., Chowdhury, M. E. H., and Gabbouj, M. Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2306–2310, 2022. doi: 10.1109/ICIP46576.2022.9897412.

Dou, Q., Liu, Q., Heng, P.-A., and Glocker, B. Unpaired multi-modal segmentation via knowledge distillation. *IEEE TMI*, 2020.

Elias, P. and Bhave, S. Chexchonet: A chest radiograph dataset with gold standard echocardiography labels. PhysioNet, 2024.

Elsharief, S., Shurrab, S., Jorf, B. A., Lopez, L. J. L., Geras, K. J., and Shamout, F. E. Medmod: Multimodal benchmark for medical prediction tasks with electronic health records and chest x-ray scans. In Xu, X. O., Choi, E., Singhal, P., Gerych, W., Tang, S., Agrawal, M., Subbaswamy, A., Sizikova, E., Dunn, J., Daneshjou, R., Sarker, T., McDermott, M., and Chen, I. (eds.), *Proceedings of the sixth Conference on Health, Inference, and Learning*, volume 287 of *Proceedings of Machine Learning Research*, pp. 781–803. PMLR, 25–27 Jun 2025. URL https://proceedings.mlr.press/v287/elsharief25a.html.

Gorade, V., Sing, A., and Mishra, D. Otcxr: Rethinking self-supervised alignment using optimal transport for chest x-ray analysis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7143–7152, 2025. doi: 10.1109/WACV61041.2025.00694.

Gupta, A., Osman, I., Shehata, M. S., and Braun, J. W. Medmae: A self-supervised backbone for medical imaging tasks, 2024. URL https://arxiv.org/abs/2407.14784.

Gupta, S., Hoffman, J., and Malik, J. Cross modal distillation for supervision transfer, 2015. URL https://arxiv.org/abs/1507.00448.

Huang, S.-C., Pareek, A., Jensen, M., Lungren, M. P., Yeung, S., and Chaudhari, A. S. Self-supervised learning

for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine*, 6(1):74, Apr 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00811-0. URL https://doi.org/10.1038/s41746-023-00811-0.

Indeewara, W., Hennayake, M., Rathnayake, K., Ambegoda, T., and Meedeniya, D. Chest x-ray dataset with lung segmentation. PhysioNet, 2023.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. URL https://arxiv.org/abs/1901.07031.

Jaeger, S., Karargyris, A., Candemir, S., Siegelman, J., Folio, L., Antani, S., Thoma, G., and McDonald, C. J. Automatic screening for tuberculosis in chest radiographs: a survey. *Quantitative Imaging in Medicine and Surgery*, 3(2), 2013. ISSN 2223-4306. URL https://qims.amegroups.org/article/view/1813.

Li, J., Aguirre, A. D., Junior, V. M., Jin, J., Liu, C., Zhong, L., Sun, C., Clifford, G., Westover, M. B., and Hong, S. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7):AIoa2401033, 2025. doi: 10.1056/AIoa2401033. URL https://ai.nejm.org/doi/full/10.1056/AIoa2401033.

Lopez-Paz, D., Bottou, L., Schölkopf, B., and Vapnik, V. Unifying distillation and privileged information, 2016. URL https://arxiv.org/abs/1511.03643.

Nguyen, H. Q., Pham, H. H., le tuan linh, Dao, M., and lam khanh. Chest x-ray dataset with lung segmentation. PhysioNet, 2021.

Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T. T., Dinh, D. H., Do, C. D., Doan, L. T., Nguyen, C. N., Nguyen, B. T., Nguyen, Q. V., Hoang, A. D., Phan, H. N., Nguyen, A. T., Ho, P. H., Ngo, D. T., Nguyen, N. T., Nguyen, N. T., Dao, M., and Vu, V. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2022. URL https://arxiv.org/abs/2012.15029.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S. B. A., Islam, M. T., Maadeed, S. A., Zughaier, S. M., Khan, M. S., and Chowdhury, M. E. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in Biology and Medicine*, 132:104319, 2021.

Saporta, A., Puli, A., Goldstein, M., and Ranganath, R. Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities, 2024. URL https://arxiv.org/abs/2411.01053.

Saporta, A., Puli, A. M., Goldstein, M., and Ranganath, R. Symile-mimic: a multimodal clinical dataset of chest x-rays, electrocardiograms, and blood labs from mimic-iv. PhysioNet, 2025.

Tahir, A. M., Chowdhury, M. E. H., Qiblawey, Y., Khandakar, A., Rahman, T., Kiranyaz, S., Khurshid, U., Ibtehaz, N., Mahmud, S., and Ezeddin, M. Covid-qu-ex. Kaggle, 2021. URL https://doi.org/10.34740/kaggle/dsv/3122958.

Tiu, E., Talius, E., Patel, P., Langlotz, C. P., Ng, A. Y., and Rajpurkar, P. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, Dec 2022. ISSN 2157-846X. doi: 10.1038/s41551-022-00936-9. URL https://doi.org/10.1038/s41551-022-00936-9.

Wang, H. et al. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. *arXiv preprint arXiv:2310.01035*, 2023.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471, 2017. URL https://api.semanticscholar.org/CorpusID:263796294.

Xiao, J., Bai, Y., Yuille, A., and Zhou, Z. Delving into masked autoencoders for multi-label thorax disease classification. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3577–3589, 2023. doi: 10.1109/WACV56688.2023.00358.

Yamaç, M., Ahishali, M., Degerli, A., Kiranyaz, S., Chowdhury, M. E. H., and Gabbouj, M. Convolutional sparse support estimator-based covid-19 recognition from x-ray images. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):1810–1820, 2021. doi: 10.1109/TNNLS.2021.3070467.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text, 2022. URL https://arxiv.org/abs/2010.00747.

Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., and Prasanna, P. Self pre-training with masked autoencoders for medical image classification and segmentation, 2023. URL https://arxiv.org/abs/2203.05573.

A. Appendix

A.1. Methodology

A.1.1. Unimodal Encoder Pretraining

CheXpert Dataset (Pretraining Data). We initialize the visual backbone using the CheXpert dataset (Irvin et al., 2019), a large-scale collection of chest radiographs. The training split consists of approximately **224k** frontal and lateral chest X-rays. All images are resized to 224×224 pixels and normalized using standard ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]) prior to patch embedding. No patient labels are utilized during this stage; the objective is purely self-supervised reconstruction.

CXR Encoder (Masked Autoencoder). We employ a Vision Transformer (ViT-B/16) architecture. The input image $x \in \mathbb{R}^{224\times224\times3}$ is divided into non-overlapping patches of size 16×16, resulting in a sequence of $N = 196$ tokens.

- **Masking Strategy:** We employ an aggressive random masking ratio of **0.90** (90% of patches are discarded), significantly higher than the standard 0.75 used in natural image MAE (Gupta et al., 2024), to prevent the model from relying on local interpolation of smooth tissues.
- **Architecture:** The encoder operates only on the visible set of patch embeddings (latent dim $D = 768$, depth=12, heads=12). A lightweight decoder (depth=8, dim=512, heads=16) reconstructs the pixel values of the masked patches.
- **Objective:** We minimize the Mean Squared Error (MSE) between the reconstructed and original patches. We enable **per-patch normalization** ('norm_pix_loss'), where the target pixels for each patch are normalized by their local mean and variance to improve contrast representation:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{i\in\mathcal{M}} \left\| \hat{x}_i - \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} \right\|_2^2 \tag{1}$$

  where $\mathcal{M}$ is the set of masked indices.
- **Optimization:** The model is trained for 400 epochs using the **AdamW** optimizer ($\beta_1 = 0.9, \beta_2 = 0.95$, weight decay 0.05). The learning rate follows a schedule with a base value of 1.5×10^{-4}, featuring a linear warmup for 10 epochs followed by a cosine decay to zero.

Laboratory Encoder (Denoising Autoencoder). We train a modality-specific encoder on the **Symile-MIMIC dataset** to handle the sparse, tabular nature of laboratory measurements.

- **Data Representation:** The input consists of N laboratory tests converted to percentiles $x \in [0, 1]^N$ and a binary missingness mask $m \in \{0, 1\}^N$.
- **Architecture:** The model is a symmetric MLP. The input layer concatenates values and masks ($x \oplus m \in \mathbb{R}^{2N}$). The encoder projects this to a hidden layer of size 512 (with LayerNorm and GELU activation) and then to a bottleneck latent representation $z \in \mathbb{R}^{256}$. The decoder mirrors this structure ($256 \to 512 \to N$), concluding with a Sigmoid activation to ensure outputs remain in the valid percentile range $[0, 1]$.
- **Denoising Objective:** We employ a dynamic corruption strategy where **15%** of the *observed* values in x are randomly zeroed out during training. The model is trained to reconstruct the original values minimizing the MSE only on the valid indices:

$$\mathcal{L}_{\text{DAE}} = \frac{\sum_{j=1}^{N} m_j \cdot (\hat{x}_j - x_j)^2}{\sum_{j=1}^{N} m_j + \epsilon} \tag{2}$$

- **Optimization:** We train for 50 epochs (batch size 256) using **AdamW** (learning rate 3×10^{-4}, weight decay 0.01). We use a step-based scheduler with a linear warmup of **200 steps** (approx. 1 epoch) followed by cosine decay.

A.1.2. Cross-modal Distillation Learning

Parameter-Efficient Fine-Tuning (PEFT). To align the CXR encoder (f_v) with the frozen physiological targets (f_e, f_l), we freeze the entire ViT backbone except for:

1. **Normalization Layers:** All 'LayerNorm' parameters are unfrozen to adapt to the domain shift from CheXpert to MIMIC-CXR.
2. **LoRA Modules:** We inject Low-Rank Adaptation matrices ($r = 8, \alpha = 8$) into the Query, Key, Value ('qkv') and Feed-Forward ('fc') projection layers of every transformer block. This accounts for $< 1\%$ of total trainable parameters.

Dual Objective Details. We train with a weighted sum of contrastive and regression losses: $\mathcal{L}_{total} = 0.5 \cdot \mathcal{L}_{CLIP} + 1.0 \cdot \mathcal{L}_{Reg}$.

(i) Cross-modal CLIP Alignment. We map the CXR (h_{cxr}) and physiological (z_m) embeddings to a shared 256-dimensional space via linear projection heads. We compute the symmetric InfoNCE loss:

$$\ell_{\text{clip}}(u, v) = -\log \frac{\exp(u^\top v/\tau)}{\sum_{k=1}^{B} \exp(u^\top v_k/\tau)} \tag{3}$$

Key implementation details:

- **Temperature (τ):** initialized at 0.07 and learned during training.
- **Global Negatives:** We utilize 'all_gather' to collect embeddings from all available GPUs, effectively scaling the number of negative samples by the number of devices (e.g., $B_{eff} = B \times N_{GPUs}$).
- **Label Smoothing:** We apply $\epsilon = 0.02$ smoothing to the target distribution to mitigate overfitting to noisy clinical pairings.

(ii) Physiology Prediction (Regression). We employ modality-specific regression heads (Linear $768 \to D_{target}$) to predict the raw, unprojected output of the frozen physiological encoders. The loss is the Cosine Distance:

$$\mathcal{L}_{\text{reg}} = \sum_{m \in \{ecg, lab\}} \frac{1}{2} \left(1 - \frac{r_m(h_{cxr}) \cdot \text{sg}[z_m]}{\|r_m(h_{cxr})\| \|\text{sg}[z_m]\|} \right) \tag{4}$$

where $\text{sg}[\cdot]$ indicates the stop-gradient operator, ensuring the physiological encoders remain purely frozen targets.

A.2. Experiments

A.2.1. Datasets

We evaluate our method across a diverse suite of public benchmarks, categorized below by task.

Classification Benchmarks

- **TB**: Provided by the National Library of Medicine (NIH), this dataset contains 662 CXRs categorized as tuberculosis or non-tuberculosis.
- **VinDr-CXR**: A large-scale benchmark of 18,000 images. Each image is annotated with one or more of 28 disease diagnoses (Nguyen et al., 2021).
- **NIH-14**: A standard benchmark comprising 112,120 CXRs from 30,805 unique patients, annotated with 14 common thorax diseases (Wang et al., 2017).
- **ChestX6**: Contains 18,036 images labeled with 6 distinct chest conditions (Adel, 2025).
- **CheXchoNet**: Comprises 71,589 images from 24,689 patients. We utilize the composite label indicating the presence of severe left ventricular hypertrophy or dilated left ventricle (Elias & Bhave, 2024).
- **MedMod**: A multimodal benchmark extracted from MIMIC-IV and MIMIC-CXR (Elsharief et al., 2025). We utilize a subset of 9,098 CXRs from 8,035 patients (excluding those in our Symile-MIMIC training set), which are labeled with 28 binary categories representing cardiopulmonary, circulatory, gastrointestinal, and endocrine diagnoses.

Segmentation Benchmarks

- **CXLSeg**: A massive dataset of 243,324 images from MIMIC-CXR with corresponding lung masks (Indeewara et al., 2023).
- **COVID-QU-Ex (Segmentation)**: A dataset of 33,920 chest X-rays (CXRs) for COVID-19 detection. Images are labeled as either COVID-positive, Non-COVID infections (Viral or Bacterial Pneumonia), or COVID-negative. Utilizing the segmentation masks provided with the COVID-QU-Ex dataset for lung and infection regions (Tahir et al., 2021).
- **QaTa-COV19**: A specialized dataset for COVID-19 pneumonia segmentation, consisting of 9,258 CXRs with infected region masks.

A.2.2. Implementation Details

Downstream Classification For standard downstream classification, we employ a linear probing protocol: a linear classifier is trained on top of the frozen backbone for 40 epochs using the Adam optimizer with a learning rate of $\eta = 3 \times 10^{-3}$ and a cosine decay schedule. Performance is reported via AUROC and F1-score. We used effective batch sizes of 512 distributed across GPUs. To simulate low-data regimes, we trained on stratified subsets of the training data (e.g., 10%), monitoring validation AUROC (macro or binary) to checkpoint the best-performing models.

Semantic Segmentation For segmentation, we attach a lightweight 5-layer decoder to the encoder backbone. Models are trained for 50 epochs ($\eta = 1 \times 10^{-4}$) using a composite objective summing Dice loss and Binary Cross-Entropy: $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{BCE}}$.

Physiological Alignment Evaluation We assess the alignment between image representations and physiological states using the Symile-MIMIC dataset (paired CXR, ECG, and Labs) (Saporta et al., 2025).

- **Regression Probe:** We extract frozen embeddings from the ViT-Base backbone (16×16 patch size, 224^2 resolution) and train a Ridge regression probe ($\alpha = 10.0$, solver="svd") to predict normalized ECG and Lab feature vectors. Performance is measured via R^2 and Cosine Similarity (raw and centered).
- **Cross-Modal Retrieval:** We compute Recall@K ($K \in \{1, 5, 10\}$) to evaluate the model's ability to retrieve the correct physiological profile for a given patient from the test set based solely on their chest X-ray.

Ablation Studies We investigate component contributions through two specific protocols:

- **Modality Ablation:** To measure the impact of specific physiological signals, we employ a masking protocol where specific modalities (ECG or Labs) are set to `None` during training. This removes them from both the cross-modal projection mechanism and regression targets.
- **Loss Ablation:** We isolate the contributions of the contrastive and regression objectives by manipulating their scalar weights (λ_{clip}, λ_{reg}). By setting a specific coefficient to 0, we neutralize that objective's contribution to gradient updates while maintaining consistent training dynamics.

Interpretability (Attention Rollout) To visualize model focus, we implement Attention Rollout. We patch the ViT blocks to cache raw attention scores and recursively multiply attention matrices from the input to the final layer, adding residual connections and averaging across heads. The resulting attention map for the [CLS] token is interpolated from the 14×14 grid to the original 224×224 resolution and min-max normalized for visualization.

A.3. Results

A.3.1. Additional Cases For Attention Rollout

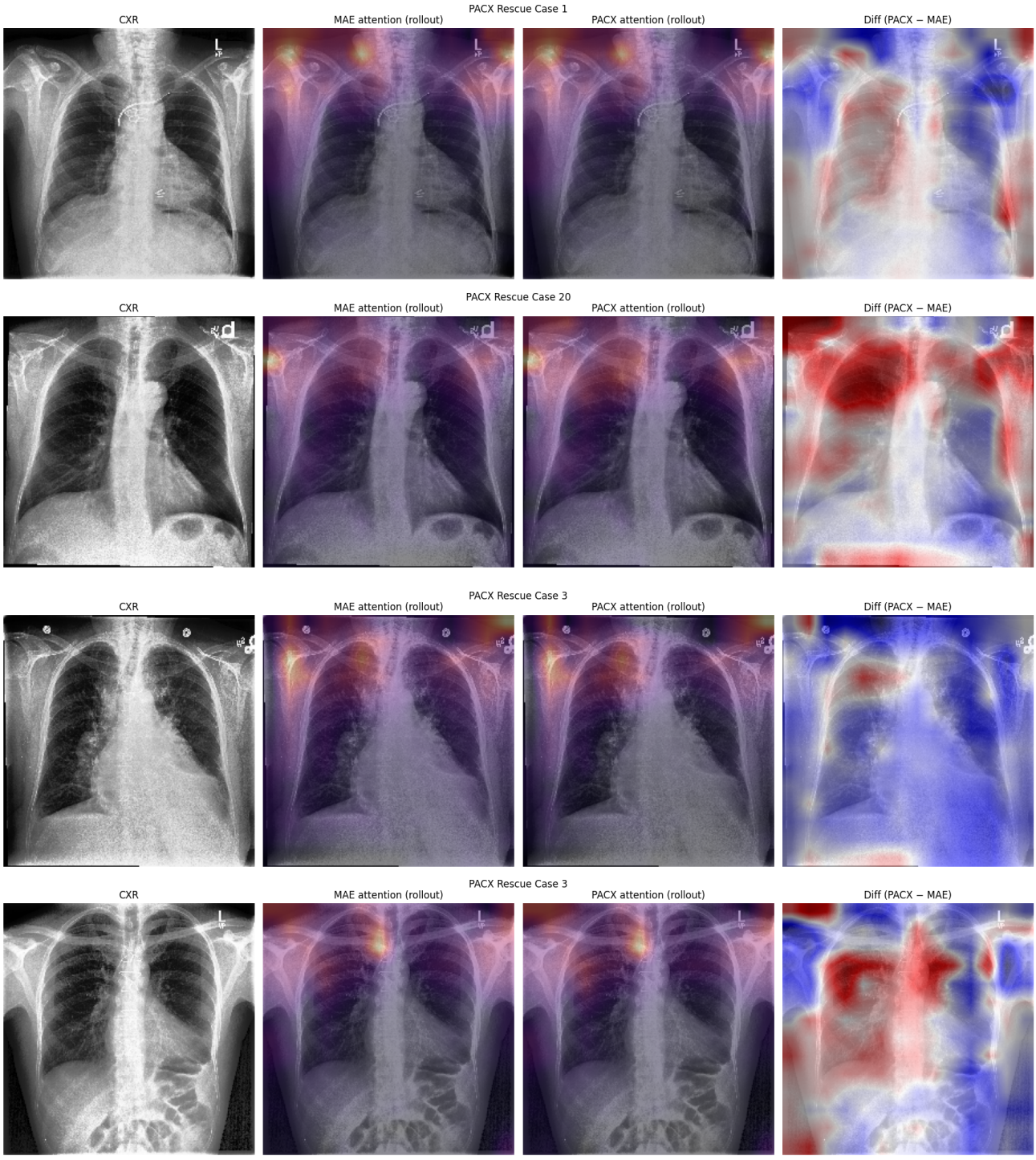

Figure 5. **Additional Attention Rollout Cases.** These visualizations confirm the consistent trend of PaCX attending to soft-tissue anatomy versus the edge-focused attention of the MAE baseline.

B. Acknowledgment

We received assistance from ChatGPT and Gemini throughout this project. This included support with literature studies, data processing, model implementation, and refinement of our manuscript, including LaTeX formatting and overall readability.

C. Contribution

Yancheng Liu spearheaded the project's conceptualization and technical execution. He formulated the core research objective of enabling "multimodal training with unimodal inference" and architected the two-stage PaCX framework. His specific contributions include:

- **Framework Design & Pretraining:** He designed and trained the MAE visual backbone, conducting extensive hyperparameter sweeps on the CheXpert dataset to optimize reconstruction quality. He further engineered the physiological encoding stream by training a Denoising Autoencoder for laboratory values and integrating the ECGFounder architecture.
- **Distillation Strategy:** He led the rigorous architectural search for the cross-modal distillation mechanism. This involved designing and evaluating multiple training strategies (fixed temperature scaling, embedding mixup, teacher-student regimes, and partial ViT unfreezing versus LoRA adaptation) to identify the optimal hyperparameters for physiological alignment.
- **Engineering & Implementation:** He developed the unified Lightning codebase, enabling stable pretraining pipelines for the MAE, cross-modal distillation, and evaluation stages. He also managed the complete data lifecycle, implementing preprocessing pipelines, conducting statistical sanity checks, and building robust DataModules for all 9 benchmark datasets.
- **Evaluation & Reporting:** He did the comprehensive experimental roadmap, logging approximately 500 GPU hours to execute linear probing, low-data regime analysis, zero-shot alignment testing, and Attention Rollout visualizations. Finally, he authored the Methodology and Results sections of this report and prepared the associated presentation materials.

Kenichi Maeda initially explored the use of the MC-MED dataset, which was eventually abandoned. He was also involved in literature studies, planning of the multimodal injection step (e.g., adaptation of the Symile-mimic dataset and the original Symile model, initial setup of teacher-student style distillation learning, etc.), segmentation evaluation, UMAP visualization, as well as modality and loss ablation studies.

Manan Pancholy initially proposed the use of the MC-MED dataset, which was eventually abandoned. Afterward, he proposed the use of the Symile-MIMIC dataset and initially implemented a SimCLR-adjacent framework that employed cross-modal attention for the multimodal injection, the latter of which was eventually abandoned. He was also involved in literature review (real-world clinical applications, medically-inspired data augmentation pipelines, and available datasets), classification evaluation (including preprocessing the CheXchoNet and MedMod datasets), modality ablation studies, and loss ablation studies. He wrote and refined their corresponding portions of the manuscript alongside the discussion section.