Demo: Clinically Diverse Chest X-ray Synthesis via Cross-Modal Conditioning

Hassan Hamidi
York University, Toronto, Canada
hhamidi@yorku.ca

Salamata Konate
York University, Toronto, Canada
sala.konate@gmail.com

Sara Hassani
UTHealth Houston, Houston, TX, USA
sara.hassani@uth.tmc.edu

Andrew Sellergren
Google Research, Mountain View, CA, USA
asellerg@google.com

Ali Sadeghi-Naini
York University, Toronto, Canada
asn@yorku.ca

Laleh Seyyed-Kalantari
York University, Toronto, Canada
Vector Institute, Toronto, ON, Canada
lsk@yorku.ca

Abstract

Latent diffusion models (LDMs) generate high-quality synthetic images from conditioning inputs but often face a trade-off between sample diversity and conditional fidelity, a tension that is acute in chest X-rays where subtle clinical cues must be preserved for diagnosis while maintaining variability for downstream tasks. We introduce a multi-conditional module that integrates multiple modalities into the conditioning signal and remains effective with any subset of available inputs, improving both diversity and fidelity. Notably, a classifier trained solely on our synthetic images matches the performance of a real-data baseline, indicating that the samples are both diverse and faithful to the conditioning. We further show that our method yields samples that better cover the real data distribution than strong baselines, and that combining our synthetic data with real images serves as an effective data augmentation strategy, improving both in-distribution and out-of-distribution generalization. These findings highlight the potential of our conditioning method as a data augmentation approach for enhancing model performance in other generative model applications, particularly in data-limited clinical settings.

1 Introduction

Diffusion models, including Latent Diffusion (LDMs) [28], have shown strong success in high-quality, controllable image synthesis. In medical imaging, persistent challenges in privacy, data scarcity, and dataset bias [2, 16, 30] motivate exploring synthetic data as a potential way to alleviate these issues.

LDMs have been applied to medical image synthesis, particularly chest X-rays (CXR) [8, 14, 19, 24, 31, 34], where conditioning guides generation to match clinical findings. Most CXR work uses a single conditioning modality [8, 14, 19, 34]. This design underuses available data and can force discarding samples when modalities are missing; for example, RoentGen [8] removed about 40% of training images that lacked paired reports because the model required them. Some studies condition only on diagnostic labels [19], although large CXR datasets often contain imprecise, machine-derived

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

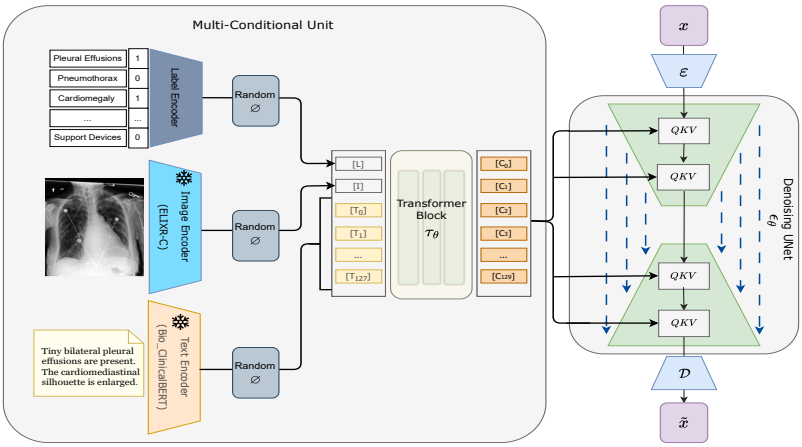

Figure 1: Illustration of the Multi-Conditional Unit: Various conditioning inputs—disease labels, X-ray images, and radiology reports—are processed by specialized encoders. A transformer then integrates these representations into a unified conditioning signal that is injected into the U-Net via cross-attention.

labels [13, 15]. We emphasize that incorporating multiple conditioning modalities (e.g., images, reports, and structured labels) can fuse complementary clinical signals, reduce dependence on any one source, and better capture the full content of an exam, leading to generations that are more clinically diverse and accurate. Another common challenge in synthetic image generation is mode collapse. In this scenario, the generative model produces highly similar, repetitive, or biased samples [9, 26, 32] rather than capturing the full diversity of real-world data. Ensuring that generated CXR images exhibit sufficient diversity to represent the real data distribution has not been a primary focus of existing studies on CXR synthetic image generation [8, 14, 19, 24, 31, 34]. This is a significant limitation, particularly in the medical image domain, given that disease prevalence varies across demographics, institutions, and imaging protocols [1, 10, 35] and insufficient diversity in generated samples lead to biases [6, 17, 25, 29] which can be harmful for the patients [5, 33] and limits the clinical utility of synthetic data in practical deployment. We present a Multi-Conditional Unit that fuses radiographic images, clinical labels, and free-text reports via domain-specific encoders and a transformer to produce a unified conditioning signal for a latent diffusion model (LDM). A stochastic null-assignment mechanism handles missing modalities, enabling training and inference under real-world data incompleteness. As a result of this design, the synthetic distribution more completely covers the real data manifold and improves sample diversity and alignment with real data. Experiments show that a classifier trained exclusively on our synthetic images performs on par with one trained on real images for disease classification. Additionally, the model remains robust across condition settings, and when a specific condition is absent, it leverages the remaining inputs to sustain performance. Finally, augmenting real datasets with our synthetic images boosts the generalization of disease classifiers in both in-distribution and out-of-distribution scenarios.

2 Related work

GANs have been widely applied to synthetic chest X-ray generation [7, 20]. Early DCGANs produced convincing radiographs for dataset augmentation [18]. Later comparisons found WGAN-GP improved realism and training stability over DCGANs [23]. Conditional variants enabled task-specific synthesis; during COVID-19, dedicated models generated positive cases to address data imbalance [21]. These tools proved particularly valuable for addressing the severe data imbalance that emerged during the pandemic. Recent developments in Diffusion Models have emerged as a compelling alternative for chest X-ray image generation [8, 14, 19, 24, 31, 34]. These models offer significant advantages over traditional GANs, particularly in overcoming the training instability, hyperparameter sensitivity, and limited diversity that have long plagued adversarial approaches. The shift toward diffusion-based architectures represents a notable advancement in medical image synthesis, promising more reliable and comprehensive generation of chest radiographs. RoentGen [8] conditions image generation on text prompts derived from radiology reports. Cascaded LDMs utilize a hierarchical architecture to

generate high-resolution CXR images [34]. CXR-IRGen [31] focuses on generating image-report pairs with clinical consistency. Packhäuser et al. [24] focus on privacy-preserving sampling strategies to anonymize synthetic CXRs. DiNO-Diffusion [14] uses self-supervised learning to train LDMs and address challenges related to limited annotated data. Ktena et al. [19] enhance the fairness of medical image classifiers in the presence of distribution shifts with synthetic images. The primary focus of these papers has been on image generation using single-modality data as the conditioning input [8, 14, 19, 34]. Moreover, the diversity of the generated images and the extent to which the synthetic data covers the real data manifold have not been their target.

3 Method

3.1 Latent Diffusion Models

Diffusion models add Gaussian noise and learn to invert it; LDMs [28] do so in a VAE latent $z_0 = \varepsilon(x)$ with a Markov chain $\{z_t\}_{t=0}^{M}$. The forward process [12] is

$$q(z_t \mid z_{t-1}) = \mathcal{N}\Big(z_t; \sqrt{\alpha_t}\, z_{t-1}, (1-\alpha_t)\mathbf{I}\Big). \tag{1}$$

A U-Net denoiser ε_θ predicts the injected noise, optionally conditioned on C, by minimizing

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_0, y, \varepsilon\sim\mathcal{N}(0,1), t}\left[\|\varepsilon - \varepsilon_\theta(z_t, t, C)\|_2^2\right]. \tag{2}$$

3.2 Multi-Conditional Unit

We introduce a multi-conditional add-on unit for LDMs, illustrated in Figure 1. This unit facilitates image generation using an ensemble of heterogeneous conditions: (a) diagnostic labels, (b) images, and (c) radiology reports. Each data point is associated with a condition tuple $y = (\ell, i, r)$, where ℓ is a disease label, i is an X-ray image, and r is a textual radiology report. Three distinct encoders, $L_\theta(\ell)$, $I(i)$, and $T(r)$, encode these respective inputs. To ensure our approach remains functional even when some conditions are missing, we adopt a strategy inspired by classifier-free guidance [11]. Specifically, we randomly replace the outputs of the encoders with a fixed null token $\varnothing$ with probabilities p_L, p_I, and p_T. Formally: $L_\theta(\ell) \leftarrow \varnothing$ w.p. p_L, $I(i) \leftarrow \varnothing$ w.p. p_I, and $T(r) \leftarrow \varnothing$ w.p. p_T. This null token is a fixed-value vector that allows for flexible conditioning, ranging from fully conditioned to entirely unconditional generation using the same backbone. The outputs of the three encoders are then concatenated and processed by a transformer τ_θ to yield a comprehensive conditioning representation C:

$$C = \tau_\theta\big([L_\theta(\ell);\ I(i);\ T(r)]\big) \in \mathbb{R}^{(n_L+n_I+n_T)\times \text{dim}}. \tag{3}$$

Here, n_L, n_I, and n_T denote the number of tokens produced by L_θ, I, and T, each with a dimension of dim. Finally, C is injected into U-Net blocks through cross-attention, and the loss function remains the same as in Eq. 2. Additional implementation details, including the encoders, transformer integration, and training hyperparameters, are provided in Appendix A.

3.3 Evaluation and Metrics

Following prior work in synthetic image generation [4, 27], we adopt a synthetic-to-real evaluation protocol in which classifiers are trained exclusively on synthetic images and tested on a held-out set of real images. Strong performance in this setting serves as a proxy for the diversity and realism of the generated data, highlighting their potential utility for downstream clinical applications. To further assess generation quality, we report two Fréchet-style distances. The first is FID computed with ImageNet-pretrained Inception-V3 features, and the second is FID_{XRV} computed with the TorchXRayVision `densenet121-res224-all` embedding. These metrics capture fidelity from both general-purpose and domain-aware perspectives. For diversity, we additionally report Coverage [22], which estimates the fraction of the real data distribution support covered by generated samples (see Appendix A.3).

.

Table 1: Comparing the performance of models trained under various conditioning strategies. We report the AUC (with 95% confidence intervals from five separate runs) for ten thoracic pathologies—Atelectasis (Atl), Cardiomegaly (Crm), Consolidation (Cns), Edema (Edm), Lung Lesion (Lls), Lung Opacity (Lop), No Finding (Nfd), Pleural Effusion (Ple), Pneumonia (Pnm) , and Pneumothorax (Ptx). Boldfaced entries highlight the best performance among synthetic data methods. The "Avg" column is the mean of the ten AUC values, the higher the better (↑). The final column shows FID scores (Inception v3) where lower values indicate better image quality and diversity (↓).

Training Data	**Atl**	**Crm**	**Cns**	**Edm**	**Lls**	**Lop**	**Nfd**	**Ple**	**Pnm**	**Ptx**	**Avg↑**	**FID↓**
Real Data	.79	.791	.742	.865	.706	.734	.845	.889	.684	.789	$.784 \pm .004$	1.7
C(L,∅,∅)	.757	.760	.723	.851	.664	.718	.828	.871	.664	.763	$.760 \pm .005$	17.2
C(∅,I,∅)	.776	.771	.735	.864	.688	.724	.840	.874	.669	.767	$.771 \pm .002$	15.2
C(∅,∅,T_f)	.746	.762	.709	.834	.685	.701	.820	.865	.646	.771	$.753 \pm .009$	18.1
C(L,∅,T)	.772	**.782**	.735	.866	.676	.729	.835	.873	.679	.767	$.771 \pm .003$	16.3
C(L,I,∅)	**.783**	.774	.733	.868	.678	.729	**.842**	.876	**.687**	.780	$.775 \pm .002$	15.0
C(L,I,T)	.775	.776	**.742**	**.869**	**.695**	**.731**	.840	.876	.677	**.798**	**.777** $\pm.002$	**14.8**
Cheff(T_f)[34]	.759	.775	.723	.853	.642	.711	.835	**.878**	.624	.754	$.755 \pm .003$	19.1
RoentGen(T_f)[8]	.753	.740	.694	.829	.583	.679	.790	.841	.626	.692	$.723 \pm .006$	52.1

Table 2: Diversity Comparison: Finding vs No Finding (↑: higher is better, ↓: lower is better).

Data Source	**Finding**		**No Finding**	
	FID$_{XRV}$ ↓	**Coverage**$_{XRV}$ ↑	**FID**$_{XRV}$ ↓	**Coverage**$_{XRV}$ ↑
Real Data	0.02	0.94	0.04	0.95
$C(L, \varnothing, \varnothing)$	**0.27**	**0.84**	0.32	0.86
$C(\varnothing, I, \varnothing)$	0.29	**0.84**	**0.26**	0.87
$C(\varnothing, \varnothing, T_f)$	0.31	0.81	0.33	0.87
$C(L, I, T)$	0.29	**0.84**	**0.26**	**0.88**
Cheff(T_f) [34]	1.02	0.56	1.44	0.49
RoentGen(T_f) [8]	6.50	0.16	8.10	0.14

4 Experiments and Results

Datasets and Baselines: In this work, we used frontal view images from the publicly available MIMIC-CXR [15] dataset. MIMIC-CXR includes 377,110 image–report pairs from 227,827 studies, with 243,334 frontal images. Following its standard partitioning, we used P10–P18 for training/validation and P19 for testing. To evaluate out-of-distribution (OOD) generalization, we additionally employed the CheXpert [13] test set, which contains 518 frontal samples. We re-evaluated and compared our results with two publicly available diffusion models, RoentGen [8] and Cheff [34], under conditions identical to those of our method. RoentGen is the most similar to ours in architecture and data; it employs a Latent Diffusion backbone and is trained on MIMIC-CXR. Cheff is trained on one of the largest open CXR corpora, and for our experiments we used their text-to-image checkpoint.

Generalization of Synthetic Data to Real Data: We train a DenseNet-121 classifier using fully synthetic data and compare its performance with a baseline trained on real data, as shown in Table 1. The real baseline used split P10 of MIMIC-CXR (23,611 samples). For consistency, we generated synthetic data based on the same split under conditioning schemes ranging from label-only $C(L, \varnothing, \varnothing)$ to full $C(L, I, T)$, enabled by the null assignments described in the methods. Results (Table 1) show that a model trained solely on synthetic data performs comparably to one trained only on real data, validating the overall quality of the generated images. Ensembling multiple conditions with the Multi-Conditional unit further improved performance; e.g., average AUC rose from 0.753 for $C(\varnothing, \varnothing, T)$ to 0.777 for $C(L, I, T)$. We also compared against RoentGen [8] and Cheff [34], generating synthetic images with their checkpoints under RoentGen's criteria, which limit text prompts to 7–77 CLIP tokens due to encoder constraints. The filtered text, denoted T_f, is shown in Table 1. Since split P10 alone did not yield enough samples (23,611) for fair comparison and FID calculation, we supplemented it with part of split P18. Our model under full conditioning $C(L, I, T)$ outperforms other settings and baselines, achieving average AUC 0.777, only slightly below the real baseline of 0.784, and the best FID of 14.8 among synthetic methods, indicating superior visual fidelity.

Diversity of Synthetic Data. We assess clinical diversity using FID_{XRV} and Coverage (Section 3.3). Real and synthetic data are split into "No Finding" and "Finding" groups: in MIMIC-CXR, "No Finding" means absence of all pathologies in the label set; "Finding" indicates at least one. In Table 2, we computed the FID_{XRV} and Coverage metrics for each group. Our method demonstrates superior performance compared to all baseline approaches across both metrics, resulting in more diverse image generation and broader coverage of the data distribution. The improved performance is potentially rooted in our LDM training approach, which incorporates multi-modal conditioning and random null assignments and enables the incorporation of more complex and diverse conditions during training compared to our baselines' single-modality conditioning. Specifically, we achieve an FID_{XRV} score of 0.27 for pathological images when using label-only conditioning, and 0.26 for "No Finding" images when utilizing our full conditioning $C(L, I, T)$. The Coverage metric reaches 0.88 under our full conditioning $C(L, I, T)$ approach. Across both metrics, baselines struggle especially on diverse "No Finding" images, suggesting that generating varied normal chest X-rays (without detectable pathologies) is more challenging for diffusion models.

Table 3: Average AUC for models trained on real, synthetic, and mixed data and tested on in- and out-of-distribution datasets.

Data	**Test Set**	
	MIMIC-CXR in-distribution	**CheXpert** out-of-distribution
Real Data	0.784	0.794
Synthetic Data `C(L,I,T)`	**0.777** (-0.007)	**0.816**(+0.022)
Mixed (Real + Synthetic) `C(L,I,T)`	**0.797** (+0.013)	**0.831** (+0.037)
Synthetic Data `Cheff`	0.755 (-0.029)	0.795 (+0.001)
Mixed (Real + Synthetic) `Cheff`	0.791 (+0.007)	0.830 (+0.036)
Synthetic Data `RoentGen`	0.723 (-0.061)	0.725 (-0.069)
Mixed (Real + Synthetic) `RoentGen`	0.792 (+0.008)	0.812 (+0.018)

Effect of Synthetic Data Augmentation on Generalization: We investigate the role of synthetic data in enhancing in-distribution and OOD generalization for disease classification tasks. Using the data subset and model configuration outlined in Section *Generalization of Synthetic Data to Real Data*, , we evaluate the influence of synthetic augmentation on both in-distribution (MIMIC-CXR) and OOD (CheXpert test set) performance, as summarized in Table 3. Our findings reveal three key insights: First, augmenting real training data with synthetic samples improves classification performance in both settings. Our full conditioning model $C(L, I, T)$ achieves AUC improvements of 0.013 (1.7%) for in-distribution and 0.037 (4.6%) for OOD evaluation. Second, a classifier trained exclusively on synthetic data using full conditioning $C(L, I, T)$ achieves an AUC of 0.816 on the OOD test set, surpassing its real-data counterpart. Third, even baseline models benefit from synthetic–real data mixing. These results align with previous findings [19], which suggest that synthetic images serve as canonical exemplars of clinical conditions, potentially enhancing robustness to prevalence shifts. Despite these promising outcomes, further investigation is required to fully understand the mechanisms driving these improvements and to address safety concerns in clinical applications.

5 Conclusion

In this work, we introduced a novel multi-conditional add-on unit for Latent Diffusion Models that harnesses heterogeneous conditioning signals, including disease labels, X-ray images, and radiology reports, to generate high-fidelity synthetic chest X-ray images. By integrating these diverse modalities, our approach enhances the overall quality and diversity of the generated outputs. Experimental evaluations demonstrate that leveraging multi-conditional information not only improves the visual fidelity of synthetic images, as evidenced by lower FID scores and more comprehensive coverage of the real data manifold, but also yields superior performance in disease classification. Moreover, mixing synthetic data with real data improved model performance in both in-distribution and out-of-distribution settings. These promising results highlight the potential of synthetic data as a powerful tool for data augmentation.

References

[1] M Ahluwalia et al. The subgroup imperative: Chest x-ray classifier generalization gaps in patient, setting, and pathology subgroups. *Radiology: Artificial Intelligence*, 2023.

[2] Sina Akbarian, Laleh Seyyed-Kalantari, Farzad Khalvati, and Elham Dolatabadi. Evaluating knowledge transfer in the neural network for medical images. *IEEE Access*, 11:85812–85821, 2023.

[3] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics, 2019. doi: 10.18653/v1/W19-1909.

[4] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *Transactions on Machine Learning Research*, 2023.

[5] Gebreyowhans Bahre et al. Underdiagnosis bias mitigation with expert foundation model's representation. *IEEE Access*, June 2025.

[6] Bahre, Gebreyowhans and others. Representation is all we need: performance and fairness of google x-ray foundation model representations – a preliminary study. *IEEE International Conference in Healthcare Informatics*, 2025.

[7] Vedant Bhagat and Swapnil Bhaumik. Data augmentation using generative adversarial networks for pneumonia classification in chest xrays. In *2019 Fifth International Conference on Image Information Processing (ICIIP)*, pages 574–579, 2019. doi: 10.1109/ICIIP47207.2019.8985892.

[8] Pierre Chambon et al. Roentgen: vision-language foundation model for chest x-ray generation, 2022.

[9] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, et al. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

[10] Carolina A. M. Heming et al. Benchmarking bias: Expanding clinical ai model card to incorporate bias reporting of social and non-social factors, 2023.

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Adv. Neural Inform. Process. Syst.*, 33:6840–6851, 2020.

[13] Jeremy A Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *AAAI*, 2019.

[14] Guillermo Jimenez-Perez, Pedro Osorio, Josef Cersovsky, Javier Montalt-Tordera, Jens Hooge, Steffen Vogler, and Sadegh Mohammadi. Dino-diffusion. scaling medical diffusion via self-supervised pre-training, 2024.

[15] Alistair E W Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-yin Deng, Roger G Mark, and Steven Horng. MIMIC-CXR: A de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(317), 2019. doi: 10.1038/s41597-019-0322-0.

[16] Faiza Khan Khattak, Vallijah Subasri, Amrit Krishnan, Chloe Pou-Prom, Sedef Akinli-Kocak, Elham Dolatabadi, Deval Pandya, Laleh Seyyed-Kalantari, and Frank Rudzicz. Mlhops: Machine learning health operations. *IEEE Access*, 4:1–47, 2024.

[17] Salamata Konate et al. Interpretability of ai race detection model in medical imaging with saliency methods. *Computational and Structural Biotechnology Journal*, 28:63–70, 2025.

[18] Sagar Kora Venu and Sridhar Ravula. Evaluation of deep convolutional generative adversarial networks for data augmentation of chest x-ray images. *Future Internet*, 13(1):8, December 2020. ISSN 1999-5903. doi: 10.3390/fi13010008. URL http://dx.doi.org/10.3390/fi13010008.

[19] Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, Alan Karthikesalingam, and Sven Gowal. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30(4):1166–1173, 2024. doi: 10.1038/s41591-024-02838-6.

[20] Mohamed Loey, Florentin Smarandache, and Nour Eldeen M. Khalifa. Within the lack of chest covid-19 x-ray dataset: A novel detection model based on gan and deep transfer learning. *Symmetry*, 12(4):651, 2020. doi: 10.3390/sym12040651. URL https://doi.org/10.3390/sym12040651.

[21] Ankit Mishra. Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection, 2021. URL https://arxiv.org/abs/2101.04749. arXiv:2101.04749.

[22] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models, 2020. URL https://arxiv.org/abs/2002.09797.

[23] MaiFeng Ng and CarolAnne Hargreaves. Generative adversarial networks for the synthesis of chest x-ray images. *Engineering Proceedings*, 31(1):84, 2023. doi: 10.3390/ASEC2022-13954. URL https://doi.org/10.3390/ASEC2022-13954.

[24] Kai Packhäuser, Lukas Folle, Florian Thamm, and Andreas Maier. Generation of anonymous chest radiographs using latent diffusion models for training thoracic abnormality classification systems. In *IEEE Int. Symp. Biomed. Imaging*, pages 1–5. IEEE, 2023.

[25] Artur Parkhimchyk, Amirreza Naziri, and Laleh Seyyed-Kalantari. Exploring visual prompt tuning for demographic adaptation in foundation models for medical imaging. In *NeurIPS 2024 Workshop on Adaptive Foundation Models (AFM)*, 2024.

[26] Andrew J Peterson. Ai and the problem of knowledge collapse. *AI & SOCIETY*, 40:3249–3269, 2025. doi: 10.1007/s00146-024-02173-x.

[27] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. In *Adv. Neural Inform. Process. Syst.*, volume 32, 2019.

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695. IEEE, 2022.

[29] Olivier Salvado et al. Localisation of racial information in chest x-ray for deep learning diagnosis. In *IEEE Int. Symp. Biomed. Imaging*, pages 1–4. IEEE, 2024.

[30] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B A McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021.

[31] Junjie Shentu and Noura Al Moubayed. Cxr-irgen: An integrated vision and language model for the generation of clinically accurate chest x-ray image-report pairs. In *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pages 5212–5221. IEEE, 2024.

[32] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget, 2023.

[33] Vallijah Subasri, Negin Baghbanzadeh, Leo Anthony Celi, and Laleh Seyyed-Kalantari. Potential for near-term ai risks to evolve into existential threats in healthcare. *BMJ Health & Care Informatics*, 32(1), April 2025.

[34] Tobias Weber, Michael Ingrisch, Bernd Bischl, and David Rügamer. Cascaded latent diffusion models for high-resolution chest x-ray synthesis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 180–191. Springer, 2023.

[35] MinJae Woo et al. Subgroup evaluation to understand performance gaps in deep learning-based classification of regions of interest on mammography. *PLOS Digital Health*, 4(4), apr 2025. Publication date: 2025-04-08.

[36] Shawn Xu et al. Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders, 2023.

A Additional Implementation Details

A.1 Architecture

Our multi-conditional unit employs three specialized encoders to process heterogeneous input modalities, with outputs subsequently integrated through a transformer module.

Label Encoder (L_θ): We encode diagnostic labels and demographic information using a trainable multilayer perceptron. The input consists of 17-dimensional vectors containing disease labels (encoded as 0 for absent, 1 for present, and -1 for uncertain) along with patient demographic information including age, gender, and race. The MLP maps these inputs to a 768-dimensional representation, producing a single token ($n_L = 1$).

Image Encoder (I): For chest X-ray image encoding, we employ the pre-trained ELIXR-C image encoder [36], which was trained via contrastive learning to align CXR images with clinical text. We extract features from the layer preceding the final classification head, yielding 1376-dimensional representations. These features are then mapped to 768 dimensions using a trainable MLP, producing a single image token ($n_I = 1$). The ELIXR-C backbone remains frozen during training to preserve clinically relevant features and prevent this encoder from learning to pass superficial visual patterns from input image rather than extracting meaningful clinical cues.

Text Encoder (T): Radiology reports are processed using Bio-ClinicalBERT [3], a domain-specific language model pre-trained on clinical text. Due to context window limit constraints, we truncate reports to the first 128 tokens. The encoder outputs the final hidden states for all tokens, resulting in a 128×768 representation ($n_T = 128$). The Bio-ClinicalBERT parameters remain frozen to retain pre-trained clinical knowledge.

Integration Transformer (τ_θ): The outputs from all three encoders are concatenated and processed by a trainable transformer module with the following specifications: 768-dimensional embeddings, 8 attention heads, 3 layers, and a feed-forward dimension of 2048. This transformer learns to integrate the multi-modal representations into a unified conditioning signal $C \in \mathbb{R}^{130\times768}$. During training, only the label encoder MLP, image encoder projection MLP, integration transformer, and the U-Net backbone are optimized, while the pre-trained ELIXR-C and Bio-ClinicalBERT models remain frozen.

VAE and U-Net: We employ a conditional U-Net architecture following the same design as Rombach et al. [28]. For the VAE, we utilize pre-trained weights from Weber et al. [34], who specifically trained a VAE for chest X-ray images to create latent representations for U-Net processing. The conditioning signal C is integrated into the U-Net through cross-attention mechanisms in the U-Net blocks.

A.2 Hyperparameters

We train our generative model using a base learning rate of 5.0×10^{-5} with input images resized to 256×256 resolution and processed in the latent space at $64 \times 64 \times 3$ dimensions. The diffusion process uses 1000 timesteps with a linear noise schedule ranging from 0.0015 to 0.0295. To enable flexible conditioning during training, we implement classifier-free guidance with dropout probabilities of $p_L = 0.4$ for labels, $p_I = 0.5$ for image embeddings, and $p_T = 0.4$ for text. Sampling was performed with a CFG scale of 4 and 75 denoising steps.

A.3 Coverage Metric

We compute the *Coverage* metric [22]. Coverage quantifies generative diversity by reporting the proportion of real-image neighbourhoods that contain at least one generated sample.

Let $\mathcal{B}(\mathbf{X}_i, r_i)$ denote the hypersphere centred on a real sample $\mathbf{X}_i$ with radius $r_i = \mathrm{NND}_k(\mathbf{X}_i)$, defined as the distance to its k-th nearest real neighbour. Coverage is then given by

$$\text{Coverage} \;=\; \frac{1}{N}\sum_{i=1}^{N} 1\Big[\exists\, j \text{ s.t. } \mathbf{Y}_j \in \mathcal{B}\big(\mathbf{X}_i, \mathrm{NND}_k(\mathbf{X}_i)\big)\Big], \tag{4}$$

where $1[\cdot]$ denotes the indicator function.