CIPHER: Causal Intervention Pathways for Healthcare Equity and Robustness

Xinyu Jia[1], Weidong Guo[1], Wangyuan Zhao[1], Yi Guo[1], Zeju Li[1,†], Yuanyuan Wang[1,†]

College of Biomedical Engineering, Fudan University, Shanghai, China
{zejuli,yywang}@fudan.edu.cn

Abstract. Deep learning models for medical diagnosis frequently exhibit substantial performance disparities across sensitive subgroups (e.g., race, sex), even when average accuracy is high. While generative data augmentation offers a route to mitigate this, existing strategies are suboptimal; they typically address only one or two dependency channels between sensitive attributes and image features. We formalize the medical image formation process via a structural causal model, revealing that sensitive attributes actually influence image content through four distinct pathways—a structural complexity neglected by prior works. Based on this insight, we introduce CIPHER (**C**ausal **I**ntervention **P**athways for **H**ealthcare **E**quity and **R**obustness), a framework designed to systematically intervene on all four causal paths. To achieve this, CIPHER utilizes a diffusion backbone equipped with classifier-free guidance and null-text inversion. This technical design enables the faithful reconstruction of patient-specific anatomy while allowing for the precise, editable synthesis of counterfactuals required to break sensitive dependency chains. We tested CIPHER using chest X-ray and dermoscopy benchmarks across both standard and shifted data distributions. By employing a multi-pathway intervention strategy, our model reduced worst-group disparities by an average of 35.8% compared to disease-conditioned synthesis baselines, while also improving total diagnostic accuracy.

Keywords: Fairness · Generative Models · Causality.

1 Introduction and Motivation

Deep learning has achieved strong performance in medical imaging tasks such as chest X-ray (CXR) interpretation and dermoscopic lesion recognition [2,7,15,16]. Yet strong average accuracy does not guarantee fairness: clinically meaningful gaps often persist across sensitive subgroups defined by sex, age, race/ethnicity, or proxies such as skin tone, which hinders safe and equitable deployment [4,6,10, 20,28,31]. Sensitive attributes can also be implicitly encoded in medical images

† Corresponding authors.
Code is available at https://github.com/fdu-farm/CIPHER.

and exploited as shortcut cues, producing subgroup-specific errors that may worsen under distribution shift across sites, scanners, and protocols [5,8,9,25,32].

Achieving algorithmic fairness requires a rigorous understanding of exactly how sensitive factors propagate through the model's decision-making pipeline. Currently, the field lacks a comprehensive framework to map these complex dynamics. To bridge this gap and move beyond surface-level heuristics, we introduce a formal causality diagram that explicitly models these interactions (Fig. 1). Let S denote sensitive attribute, D the disease label, C the data availability or sampling mechanism, and Y the final observed or generated image. Under this framework, S influences Y via **four** distinct routes: population bias ($S \to Y$), manifestation bias ($S \to D \to Y$), prevalence bias ($S \to C \to Y$), and environment bias ($S \to D \to C \to Y$). Because each pathway induces fundamentally different type of model failure, treating synthetic data augmentation as a black-box remedy offers limited insight. Even if overall fairness metrics improve, it remains entirely unclear which specific dependency was corrected (e.g., data coverage, physical manifestation, or spurious shortcuts), ultimately failing to explain why these gains often collapse under distribution shift.

Despite the complex causal landscape of medical image formation, existing fairness interventions remain largely insufficient because they typically target only a single causal edge—primarily through basic attribute resampling [17, 26] or rudimentary disease-conditioned synthesis [12, 18, 19, 24, 27]. While generative tools such as inversion-based editing offer a promising avenue for creating instance-anchored counterfactuals that preserve patient-specific context [22, 23], prior work lacks a unified, causally grounded framework that integrates global distribution reshaping with local editing [3]. We argue that these existing synthesis strategies are fundamentally limited because they address only one or two of the four critical causal pathways, failing to provide a comprehensive solution to the multifaceted nature of subgroup disparities.

Therefore, we propose CIPHER (**C**ausal **I**ntervention **P**athways for **H**ealthcare **E**quity and **R**obustness), a diffusion-based generative framework grounded in the structural causality of medical image formation [3, 17]. Unlike existing heuristics, CIPHER provides a complete theoretical safeguard: by formally organizing control around (S, D, C, Y) and intervening on all relevant edges ($S \to Y$, $D \to Y$, $S \to C$, $D \to C$), it comprehensively blocks every path where S can bias the observed distribution. To implement this, our novel architecture leverages classifier-free guidance and null-text inversion to achieve highly faithful, instance-anchored synthesis [23]. Evaluations on CXR and dermoscopy show that mixing complementary paths maximizes accuracy and fairness, consistently reducing worst-group AUROC gaps across distribution shifts. Summary of contributions:

- We formalize the causal pathways through which sensitive attributes S bias observed images Y, laying a foundation for evaluating fairness interventions.
- We introduce CIPHER, a versatile diffusion-based framework that translates a theoretical causal graph over (S, D, C, Y) into four explicit, intervenable generation and editing pathways.

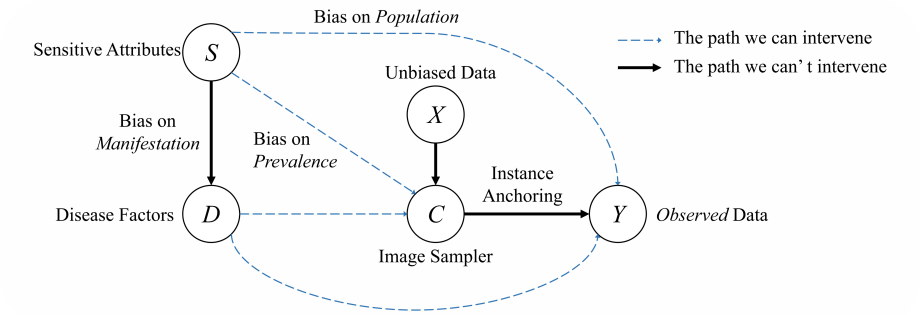

Fig. 1: Causal graph of medical image formation. The graph illustrates how sensitive attributes (S), disease labels (D), and sampling mechanisms (C) influence the image (Y) through distinct causal pathways and biases.

- We enable precise demographic subgroup reshaping and multi-attribute, instance-anchored counterfactual editing via null-text inversion.
- We demonstrate that complementary paths maximizes fairness and robustness, achieving an 8.5% to 61.6% reduction in worst-group AUROC gaps under distribution shift compared to existing approaches.

2 Analysis

Setup and notation. As illustrated in Fig. 1, our structural causal model (SCM) elucidates how subgroup unfairness emerges even in the absence of explicit access to S. Because the sensitive attribute fundamentally shapes the training distribution, we formalize this bias through four distinct causal pathways as follows:

$$S \to Y, \quad S \to D \to Y, \quad S \to C \to Y, \quad S \to D \to C \to Y. \tag{1}$$

We detail the underlying causal mechanisms for each pathway below:
1. $S \to Y$**: Population bias.** This represents a direct causal influence where sensitive attributes induce group-specific appearance or acquisition shifts. Consequently, the interventional distribution $P(Y|do(S))$ allowing models to leverage these cues as shortcuts. This results in subgroup-specific error patterns even when the clinical label is invariant [8, 9, 21].
2. $S \to D \to Y$**: Manifestation bias.** Here, S acts as a modifier of the disease's visual expression. Even with disease D is held constant, the conditional distribution $P(Y|D, S)$ varies due to subgroup-linked differences in anatomical context or severity. Models trained on dominant phenotypes fails to generalize to underrepresented manifestations, leading to systematic diagnostic gaps [4, 6, 20, 28].
3. $S \to C \to Y$**: Prevalence bias.** In this pathway, S influences the selection mechanism C, which governs data availability. This causal chain shifts the effective support of the observed distribution, creating a disparity where $P(Y|S, C =$

Table 1: CXR examples for causal pathways by which sensitive factors S can influence the observed image Y.

Causal pathway	**Phenomenon in CXR**
$S \to Y$	Base physiological differences (e.g., bony patterns, global intensity) that affect the background of an image regardless of disease.
$S \to D \to Y$	Group-specific presentation of a disease (e.g., how "Pneumothorax" looks different in a smaller chest cavity vs. a larger one).
$S \to C \to Y$	Either because a group is sampled less often or has a different natural base rate of the disease (e.g., White 56.7%, Asian $< 11.0\%$).
$S \to D \to C \to Y$	The complex interaction where the likelihood of a patient being "captured" depends on both their demographic (e.g. Older patients are twice as likely to have respiratory issues like edema).

1) is poorly sampled for minority groups. This imbalance biases generative and predictive models toward majority-group image statistics [10, 15, 16, 21].

4. $S \to D \to C \to Y$: Environment bias. This pathway represents intersectional selection bias. Formally, the selection mechanism $P(C|S, D)$ creates non-random missingness that varies across demographic-clinical "slices." Unlike marginal bias, this concentrates failures in specific intersectional strata, such as a specific disease manifestation unique to a vulnerable subgroup [28].

To ground these concepts, Table 1 provides real-world examples of these biases in Chest X-ray (CXR) imaging. Our causal analysis implies simply increasing the volume of synthetic data is insufficient; rather, because distinct disparities originate from specific structural dependencies, they necessitate targeted intervention targets. Accordingly, CIPHER explicitly models S, D, and C, via two complementary controls: (i) Global (S, D) Conditioning: Controls subgroup and disease composition to decouple S from D, mitigating $S \to Y$ shortcuts and isolating manifestation shifts in $P(Y|D, S)$. (ii) Instance-Anchored Counterfactuals: Employs local edits to generate "matched pairs" that bypass the selection mechanism C, populating sparse (S, D) strata to counteract prevalence and environment biases. This transforms augmentation from a black-box heuristic into a causal framework for addressing the structural sources of subgroup disparity.

3 Methodology

3.1 Preliminaries: Latent Diffusion and Null-Text Inversion

Latent diffusion. As shown in Fig. 2, we adopt a latent diffusion model (LDM) that runs diffusion in a compact latent space [12, 24, 27]. At its core, the model relies on a text-conditioned U-Net $\epsilon_\theta(\cdot)$ to estimate the noise injected into the latent space. By conditioning on the text embedding c, the network denoises the representation to recover the target latent state at any given step t.

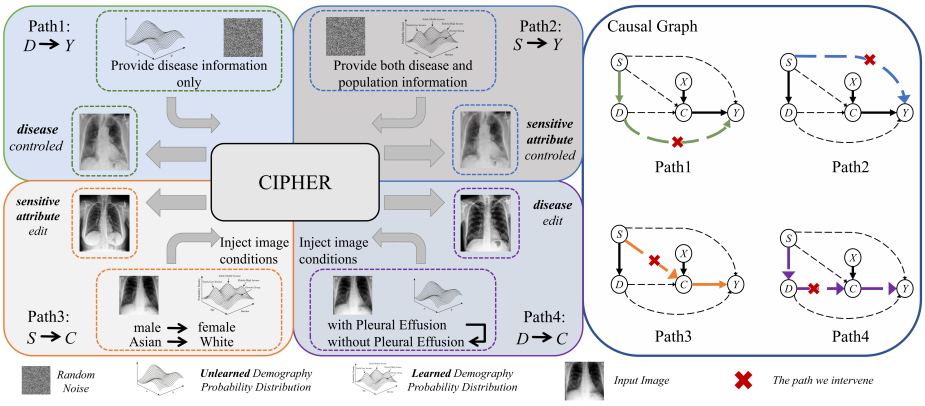

Fig. 2: Overview of our causal counterfactual generation framework.

Classifier-free guidance (CFG). We use classifier-free guidance:

$$\hat{\epsilon} = \epsilon_\theta(z_t, t, u) + s\big(\epsilon_\theta(z_t, t, c) - \epsilon_\theta(z_t, t, u)\big), \tag{2}$$

where u is the unconditional embedding and s is the guidance scale [13].

Null-text inversion for faithful, editable reconstruction. To realize our counterfactual editing function, $\mathcal{E}$, which executes a designated attribute transition (denoted by the operator $\rightarrow$), we employ null-text inversion [23]. This technique optimizes only the unconditional embeddings in CFG to improve image reconstruction while keeping c fixed. Given an input anchor image and its source prompt c (capturing the orignal attributes (d, s)), we perform DDIM inversion to obtain an initial trajectory [29], then optimize step-wise unconditional embeddings $\{u_t\}_{t=1}^{T}$ to minimize reconstruction error. To apply a counterfactual transition such as $(d, s) \rightarrow (d, s')$ during the editing phase, we replace c with an edited prompt c' (e.g., modifying disease/metadata tokens). By generating the new image with c' while reusing the optimized $\{u_t\}_{t=1}^{T}$, the function $\mathcal{E}$ successfully modifies the targeted attributes while preserving the unobserved image structure, drastically reducing off-target drift. We support global (unmasked) and localized (masked) edits under the same inversion trajectory (Fig. 2) [1, 11, 22].

3.2 Causal Framework and Intervenable Pathways

Causal Variables and the Generation Switch. We treat the textual representations of the sensitive and disease variables as intervenable—meaning we can inject, edit, block, or hold fixed their respective text embeddings, $c_S(\cdot)$ and $c_D(\cdot)$. Crucially, we do not intervene on intrinsic, dataset-implied constraints (such as the biased observational co-occurrence of S and D) or the instance-anchoring structural noise when editing a real image.

Four Complementary Generation and Editing Paths. Based on this framework, Fig. 2 instantiates into four parallel paths: These pathways operate at two

distinct rungs of the causal hierarchy, dictated by C: Paths 1 and 2 operate as global synthesis directly from noise, while Paths 3 and 4 perform instance-anchored counterfactual editing via diffusion inversion on real images.

Path 1: Disease-only generation (control D, block S). When generating from noise, we sample from the observational distribution by omitting the sensitive tokens entirely:

$$y \sim p_\theta(Y \mid c_D(d')) \,. \tag{3}$$

By conditioning solely on $c_D(d')$, we block explicit textual guidance for S. While this increases the broader appearance coverage under a given disease D, the generative process marginalizes over the biased observational prior.

Path 2: Sensitive-focused generation (control S with D held fixed). To perform an interventional cross-demographic synthesis, we apply a joint intervention on both variables by concatenating (i.e. $\oplus$) the conditions:

$$y \sim p_\theta(Y \mid c_D(d) \oplus c_S(s')) \,. \tag{4}$$

We hold the disease semantics $D{=}d$ fixed while systematically varying S to reshape subgroup composition. This forces the model to decouple the variables, allowing us to generate balanced datasets and actively probe subgroup-associated visual cues without altering the core clinical manifestation.

Path 3: Instance-anchored sensitive counterfactual editing (edit S, keep D fixed). Transitioning to counterfactual editing, we intervene on S for a specific observed anchor image $X = x$:

$$y \;=\; \mathcal{E}(x; (d, s) \to (d, s')). \tag{5}$$

Using diffusion inversion to capture the unobserved background context of the real instance, we edit only S-related tokens while holding D structurally unchanged. This yields strictly matched counterfactuals, effectively isolating the demographic features to audit the model's reliance on shortcut cues.

Path 4: instance-anchored disease counterfactual editing (edit D, keep S fixed). Similarly, we compute the counterfactual manifestation of the disease on a fixed patient background:

$$y \;=\; \mathcal{E}(x; (d, s) \to (d', s)). \tag{6}$$

We abduce the latent context of x and selectively swap the disease condition $c_D(d) \to c_D(d')$ during the denoising trajectory while keeping S strictly fixed. This ensures that the causal effect of the edit preserves the specific patient context and concentrates changes entirely on the physical evidence of the disease.

4 Experiments and Results

Datasets. We use CheXpert Plus as the in-distribution (InD) dataset and MIMIC-CXR as the out-of-distribution (OOD) testbed [15, 16]. For CXR, we apply subject-disjoint InD splits, and reserve MIMIC-CXR for OOD evaluation

only. We restrict images to the PA view, remove uncertain labels, and select 6 disease labels, yielding 14,138 InD training images and 8,347 OOD test images under the same criteria. We leverage report text to guide diffusion training [15]. For dermoscopy, we use the public MILK10K dataset [30], which includes demographic metadata (age, sex, and skin-tone proxies). We use all 10,480 images to fine-tune the diffusion model.

Evaluation. We evaluate performance across age and race subgroups: ages 21–90 are binned into 7 ten-year groups (21–30, . . . , 81–90), and CXR race analysis considers the four most populated categories—Asian, White, Black, and Hispanic.In addition, on the dermoscopy dataset, we further stratify test cases by the *MILK10K skin-tone level* labels for subgroup evaluation. For downstream diagnosis, we train a DenseNet121 [14] and report test-set mAUC. Fairness is assessed via subgroup AUROC per demographic attribute. We report the maximum AUROC gap across categories (Age/Sex/Race/Skin) and its relative reduction compared the Path1 (i.e. disease-conditioned synthesis) baseline, a widely adopted strategy for enhancing model performance.

4.1 Fine-Tune Diffusion Model

We fine-tune **Stable Diffusion v1-4** on each modality to adapt the generator to the medical imaging domain [12, 24, 27]. For all experiments, we run 30,000 optimization steps with a base learning rate of 1×10^{-4} and a cosine annealing schedule. Unless specified, we maintain a 1:1 real-to-synthetic ratio. Evaluating ratios from 1:1.5 to 4:1 confirms that 1:1 yields the optimal accuracy–fairness trade-off. We freeze all components except the U-Net and the text encoder.

For CXR, we fine-tune using cleaned medical report–image text pairs, where textual condition is derived from the processed radiology reports. For dermoscopy, we fine-tune using disease label–image text pairs, i.e., prompts constructed from diagnostic labels. In both modalities, we additionally inject metadata tokens (e.g., age/sex/race or skin-tone when available) into the textual condition to enable demographic-aware generation and editing in subsequent paths.

4.2 Ablation Experiments

To evaluate CIPHER for downstream diagnosis and disentangle the effects of different paths, we train classifiers under seven data settings: Real, a conventional augmentation baseline RandAugment, P1–P4 (Real augmented with synthetic samples generated by a *single* path), and CIPHER (our full method). In Fig. 3, each marker shows the mean with error bars across runs; moving right indicates a better performance and smaller gap (better fairness).

Upper Fig. 3 illustrates CXR performance and fairness for Age and Race under InD (CheXpert Plus) and OOD (MIMIC) shifts. **InD:** single-path variants can improve mAUC (e.g., P4 is highest), but gap reductions are limited. CIPHER yields the best overall trade-off, improving mAUC by 0.003 while reducing the Age gap by 0.019 (21.8%) and the Race gap by 0.013 (32.5%) relative to P1. **OOD:** RandAugment and single-path variants are less consistent under shift,

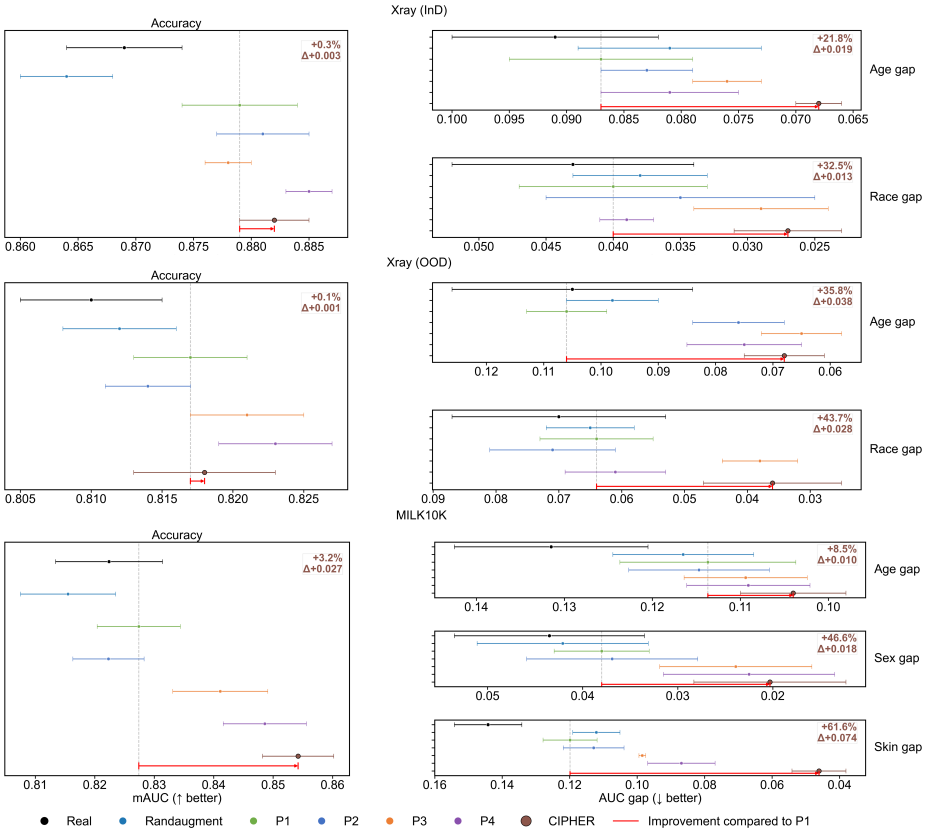

Fig. 3: Results on chest radiology and dermoscopy datasets: average AUC vs. fairness (AUC) gap across different paths for race/age (radiology, in-distribution & OOD) and sex/age/skin tone (dermoscopy).

often improving mAUC without reliably reducing gaps. CIPHER remains the most robust setting, improving mAUC by 0.001 and reducing the Age gap by 0.038 (35.8%) and the Race gap by 0.028 (43.7%).

Lower Fig. 3 reports MILK10K performance and fairness for Age, Sex, and Skin. Single-path variants provide partial gains, whereas CIPHER consistently dominates, achieving the highest mAUC (0.027) and the largest gap reductions: Age 0.010 (8.5%), Sex 0.023 (46.6%), and Skin 0.074 (61.6%).

We further compared CIPHER with MixUp, CutMix, JTT, and SELF across three datasets. CIPHER consistently yields the smallest fairness gaps across age, race, gender, and skin tone in both in-distribution and OOD settings, validating the advantage of our causality-driven framework over heuristic augmentation and debiasing methods.

5 Conclusion

We introduce CIPHER, a diffusion-based framework that optimizes the accuracy–fairness trade-off by leveraging a structural causal perspective of medical image formation. By organizing control around (S, D, C, Y) and intervening on controllable edges, CIPHER provides a unified approach to fairness-aware data augmentation. We show that synthesizing complementary causal pathways yields the most robust trade-off, simultaneously enhancing mean AUROC and narrowing subgroup disparities under distribution shift. Future research could explore the integration of more sophisticated generative architectures or the development of a fully unified synthesis pipeline based on the presented SCM.

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

References

1. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18392–18402 (2023)
2. Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., Yap, M.H.: Analysis of the ISIC image datasets: Usage, benchmarks and recommendations. Medical Image Analysis **75**, 102305 (2022)
3. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nature Communications **11**(1), 3673 (2020)
4. Chen, R.J., Hussain, S., Ren, J., et al.: Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature Biomedical Engineering **7**, 719–742 (2023)
5. D'Amour, A., Heller, K., Moldovan, D., et al.: Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research **23**(226), 1–61 (2022)
6. Daneshjou, R., Vodrahalli, K., Novoa, R.A., et al.: Disparities in dermatology AI performance on a diverse, curated clinical image set. Science Advances **8**(32), eabq6147 (2022)
7. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115–118 (2017)
8. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**, 665–673 (2020)
9. Gichoya, J.W., Banerjee, I., Bhimireddy, A.R., et al.: AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health **4**(6), e406–e414 (2022)
10. Glocker, B., Jones, C., Bernhardt, M., Winzeck, S.: Algorithmic encoding of protected characteristics in chest X-ray disease detection models. eBioMedicine **89**, 104467 (2023)
11. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross-attention control. arXiv preprint arXiv:2208.01626 (2022)

12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
13. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
14. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4700–4708 (2017)
15. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 590–597 (2019)
16. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019)
17. Jones, C., Castro, D.C., De Sousa Ribeiro, F., Oktay, O., McCradden, M.D., Glocker, B.: A causal perspective on dataset bias in machine learning for medical imaging. Nature Machine Intelligence **6**, 138–146 (2024)
18. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R.: Diffusion models in medical imaging: A comprehensive survey. Medical Image Analysis **88**, 102846 (2023)
19. Ktena, I., Wiles, O., Albuquerque, I., Rebuffi, S.A., Tanno, R., Guha Roy, A., Azizi, S., Belgrave, D., Kohli, P., Karthikesalingam, A., Cemgil, A.T., Gowal, S.: Generative models improve fairness of medical classifiers under distribution shifts. Nature Medicine **30**, 1166–1173 (2024)
20. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020)
21. Lotter, W.: Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. Nature Communications **15**, 7465 (2024)
22. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (ICLR) (2022)
23. Mokady, R., Hertz, A., Aberman, K., Papanastasiou, G., Mosseri, I., Irani, M.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6038–6047 (2023)
24. Nichol, A., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171 (2021)
25. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning. pp. 151–159 (2020)
26. Puyol-Antón, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., King, A.P.: Fairness in cardiac mr image analysis: an investigation of bias due to data imbalance in deep learning based segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 413–423. Springer (2021)

27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
28. Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A., Chen, I.Y., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. Nature Medicine **27**(12), 2176–2182 (2021)
29. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR) (2021)
30. Tschandl, P., Akay, B.N., Rosendahl, C., Rotemberg, V., Todorovska, V., Weber, J., Wolber, A.K., Müller, C., Kurtansky, N., Halpern, A., Weninger, W., Kittler, H.: MILK10k: A hierarchical multimodal imaging-learning toolkit for diagnosing pigmented and nonpigmented skin cancer and its simulators. Journal of Investigative Dermatology **146**(2), 357–364.e7 (2026)
31. Yang, Y., et al.: Demographic bias of expert-level vision-language foundation models in medical imaging. Science Advances **11**(13), eadq0305 (2025)
32. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine **15**(11), e1002683 (2018)