Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CompDiff enables fair and zero shot medical image generation across demographic intersections through compositional diffusion

Mahmoud K. Ibrahim

mahmoud.ibrahim@maastrichtuniversity.nl

Maastricht University

Bart Elen

Flemish Institute for Technological Research

Chang Sun

Maastricht University

Ahmad Jiblawi

Icahn School of Medicine at Mount Sinai

Mohammed M. Saleh

Massachusetts General Hospital

Maryam Ibrahim

Massachusetts General Hospital

Gökhan Ertaylan

Flemish Institute for Technological Research

Michel Dumontier

Maastricht University

Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11128264/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

CompDiff enables fair and zero shot medical image generation across demographic intersections through compositional diffusion

Mahmoud K. Ibrahim[1,2,3,*], Bart Elen[3], Chang Sun[1,2], Ahmad Jiblawi[4], Mohammed M. Saleh[5], Maryam K. Ibrahim[6], Gökhan Ertaylan[3] and Michel Dumontier[1,2]

[1]Institute of Data Science, Faculty of Science and Engineering, Maastricht University, Maastricht, The Netherlands

[2]Department of Advanced Computing Sciences, Faculty of Science and Engineering, Maastricht University, Maastricht, The Netherlands

[3]VITO, Belgium

[4]Icahn School of Medicine at Mount Sinai, The Mount Sinai Hospital, New York City, USA

[5]Department of Radiology, Division of Abdominal Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA

[6]Division of Gastroenterology, Massachusetts General Hospital, Boston, Massachusetts, USA

[*]Correspondence: mahmoud.ibrahim@maastrichtuniversity.nl
ORCID: 0000-0002-5099-7609

Code Models Dataset

Abstract

Medical image generators trained on imbalanced data can fail at demographic intersections absent from training. We introduce CompDiff, which encodes age, sex and race separately and composes supervised demographic tokens alongside clinical text. Across chest radiographs and fundus images, CompDiff improves overall and subgroup fidelity relative to prompt conditioning (RoentGen-v2) and loss reweighting (FairDiffusion). It generalises in zero-shot generation to 16 chest X-ray intersections excluded from training, achieving the lowest mean FID-RadImageNet in every intersection. In a blinded reader study of these unseen intersections, two radiologists gave CompDiff the highest mean scores among generators for anatomical realism and agreement with the clinical impression, and selected its images most often as the most realistic. Pretraining with CompDiff images improved downstream classification, while CompDiff audit cohorts reduced estimation error on rare intersections. These findings support compositional demographic conditioning for extending medical image synthesis to underserved populations.

Keywords: Diffusion models; Synthetic medical data; Algorithmic fairness; Chest radiography

1 Introduction

Audit studies of classifiers for chest radiographs have documented that models trained on demographically imbalanced imaging data can systematically underserve underrepresented patient subgroups, with the largest disparities affecting intersections such as Black women[1,2]. We use *cell* to denote a subgroup defined by a specific combination of demographic attributes, such as an age × sex × race group. Synthetic augmentation with text-to-image diffusion models is a widely proposed remedy: if a generator can synthesise realistic samples of an underrepresented subgroup, it can rebalance the training corpus of a downstream classifier[3–6]. This remedy, however, depends on the generator itself synthesising equally well across demographic groups.

When trained on imbalanced data, a diffusion model can achieve strong average fidelity while its quality on individual subgroups is left to chance (Figure 1); on the MIMIC-CXR training distribution used in this work, the rarest age × sex × race intersections contain only 47–92 training images, from 26–48 patients (Supplementary Figure 1a), and combined with rare pathologies these intersections are frequently absent altogether (Figure 1a). A training corpus may contain elderly patients, Asian patients, and female patients, yet have zero examples at the intersection of all three with a specific pathology. No amount of oversampling or reweighting can address groups that do not exist in the training data, and prompt conditioning supplies no mechanism that supervises how such a group is composed from the groups that do. We refer to this gap as the *imbalanced generator problem*. The stakes are not merely aesthetic: when synthetic images systematically have lower fidelity for a subgroup, using them to augment or rebalance a training set teaches the downstream classifier a distorted picture of exactly the patients the intervention was meant to protect, reintroducing or even amplifying the disparities the augmentation was intended to remove.

Existing generation methods address unequal data in several ways. FairDiffusion uses Fair Bayesian Perturbation to adapt the diffusion training loss across demographic groups[7]. FairGen combines resampling, regularisation for class diversity and preference optimisation aligned with physician judgements to improve demographic equity and clinically relevant synthesis[8]. RoentMod edits pathology in existing chest radiographs while preserving unrelated anatomy, using the resulting counterfactual images to identify and reduce shortcut learning[9]. ChexGen is a foundation model for chest radiography that supports synthesis guided by text, masks and bounding boxes, including demographic control and downstream fairness applications[10]. These studies establish complementary uses of synthetic medical images. Here we examine a specific design question: whether a dedicated demographic conditioner can generate intersections deliberately removed from training, and whether those images support downstream learning and fairness audits. FairGen, RoentMod and ChexGen provide context for this question; they were not retrained as experimental comparators in this study.

We test whether demographic attributes can be composed from components learned in observed cells, whether this representation improves fidelity and subgroup equity relative to prompt conditioning (RoentGen-v2) and loss reweighting (FairDiffusion), and whether the resulting cohorts improve downstream classification and the estimation of performance on rare intersections. The distinction between conditioning mechanisms is summarised in Table 1.

We propose **CompDiff**, which addresses the imbalanced generator problem at the *representation* level (Figure 1b). Our key idea is that demographic identity can be modelled compositionally: a rare intersection such as "80+ Asian female" can be *composed* from separate representations of age, sex and race learned from more abundant cells. CompDiff instantiates this idea in a dedicated Hierarchical Conditioner Network (HCN). It encodes age as a continuous value and sex and race as

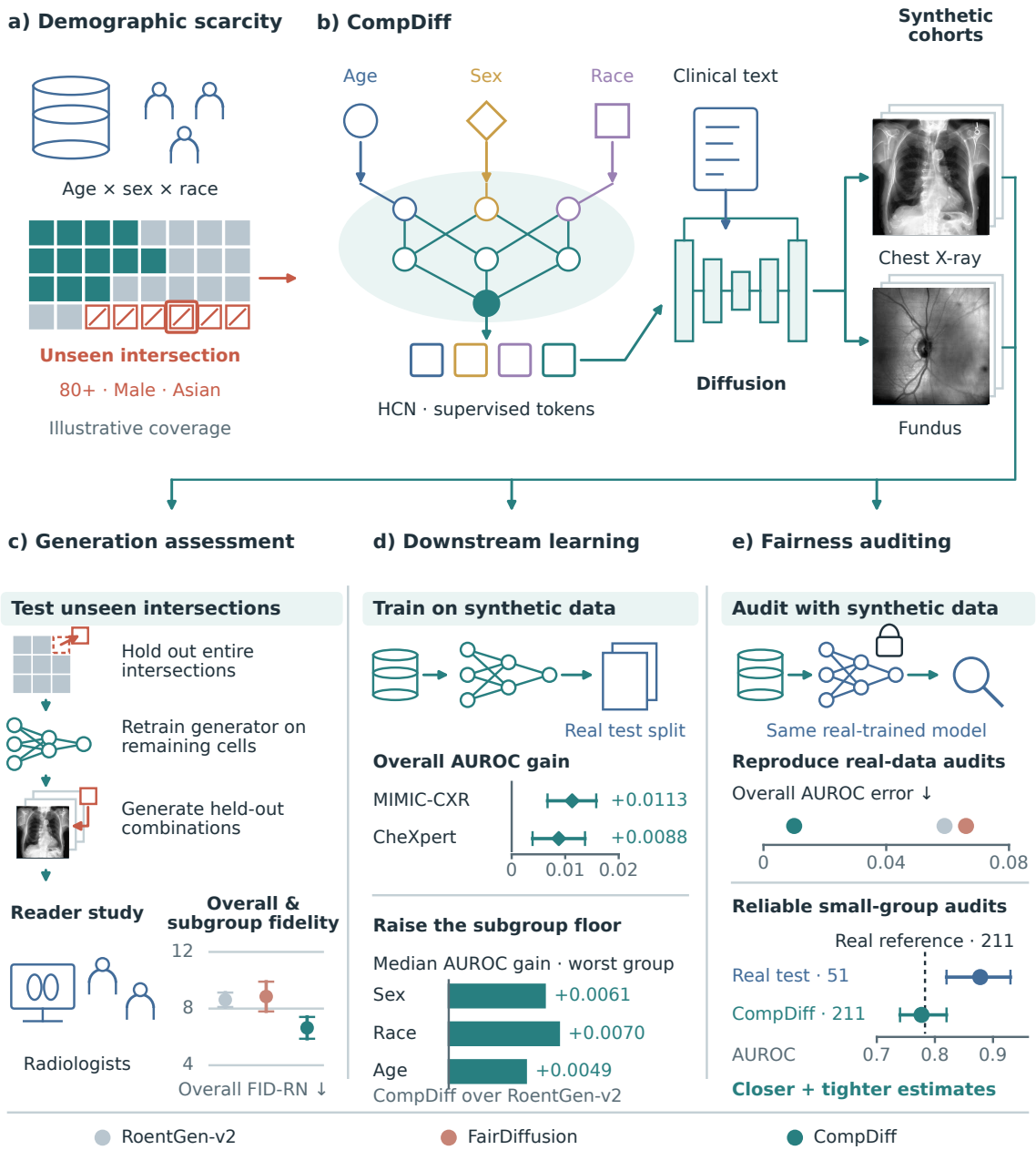


Figure 1: CompDiff: generating across demographic intersections to support learning and fairness auditing. **(a)** Uneven demographic coverage leaves some combinations of age, sex and race absent from training. **(b)** The Hierarchical Conditioner Network (HCN) composes supervised demographic tokens, which guide diffusion alongside separately encoded clinical text. Coverage and model diagrams are schematic. **(c)** Generation is assessed through a blinded radiologist study of intersections withheld from training and comparisons of overall and subgroup fidelity. The example plot shows overall chest X-ray Fréchet inception distance with RadImageNet features (FID-RN), with standard deviations across three training runs. **(d)** Synthetic training is evaluated for overall classification and for the weakest demographic groups. Diamonds show CompDiff's overall gains over RoentGen-v2 after training on synthetic images alone; bars show gains for the baseline's weakest groups after synthetic pretraining and fine-tuning on real images. Diamonds have pointwise 95% Welch confidence intervals from four classifier runs per generator and corpus, averaging four corpora equally. Bars show medians across 48 configurations evaluated on both test datasets. **(e)** Synthetic cohorts audit a fixed classifier in two ways: reproducing evaluations on real data overall and improving the accuracy and precision of estimates for small subgroups. Upper points show absolute error relative to real-data evaluation. Below, real and CompDiff estimates for Asian women aged 60–80 are compared with the dashed real reference; intervals are 95% bootstrap confidence intervals from 1,000 resamples. AUROC denotes area under the receiver operating characteristic curve; panel (d) averages eight diagnostic labels. Grey, orange and green denote RoentGen-v2, FairDiffusion and CompDiff; blue denotes real data.

Table 1: Conditioning and fairness mechanisms relevant to CompDiff. The table distinguishes the intervention and the task it supports; it does not report a matched performance comparison with every cited system. The fused demographic embedding and second text branch are controls from this study. The unstructured composer with a matched parameter count retains CompDiff's separate attribute encoders and supervised output tokens.

Method	Intervention	Evaluation focus
Standard latent diffusion[11]	Text-conditioned denoising	Image synthesis from prompts
FairDiffusion[7]	Adaptive weighting of the diffusion loss	Demographic equity in medical synthesis
FairGen[8]	Resampling, diversity regularisation and physician preferences	Equitable, clinically relevant medical synthesis
RoentMod[9]	Image-to-image modification of pathology	Counterfactual shortcut testing and correction
ChexGen[10]	Foundation model with text and spatial controls	Chest radiograph synthesis and downstream model development
Composable Diffusion[12]	Composition of conditional scores at sampling	Object and scene combinations
Fused demographic embedding	Demographic embeddings fused into one token, with auxiliary supervision	Conditioning control (Section 2.7)
Second text branch	Additional demographic text pathway	Conditioning control (Section 2.7)
CompDiff	Separate encoders and supervised tokens for age, sex and race, plus a composed token	Unseen demographic intersections, two imaging modalities, blinded chest X-ray reader study, downstream learning and auditing

categories, then produces one demographic token per attribute plus a composed token, concatenated to the CLIP[13] text embeddings as cross-attention context for the diffusion UNet. HCN is named for the pairwise hierarchy of the composer we instantiate. The resulting interface separates demographic control from the clinical text pathway (Figure 2).

We evaluate CompDiff on chest X-rays from MIMIC-CXR[14] and fundus images from FairGenMed[7], with matched RoentGen-v2 and FairDiffusion baselines. CompDiff attains the best value on 28 of 29 subgroup comparisons and the lowest mean FID-RadImageNet across three training runs in all 16 unseen chest X-ray intersections. A separate blinded radiologist study tests realism and clinical correspondence at these intersections. Downstream experiments assess synthetic training data on MIMIC-CXR and CheXpert[15], while separate audit experiments test how closely synthetic cohorts reproduce estimates from real data overall and on rare intersections. Ablations distinguish the contribution of separate encoding and supervision for each attribute from the network used to combine them. Figure 1c–e outlines the protocol for withholding intersections and distinguishes overall and subgroup benefits of synthetic cohorts for classifier training and fairness auditing.

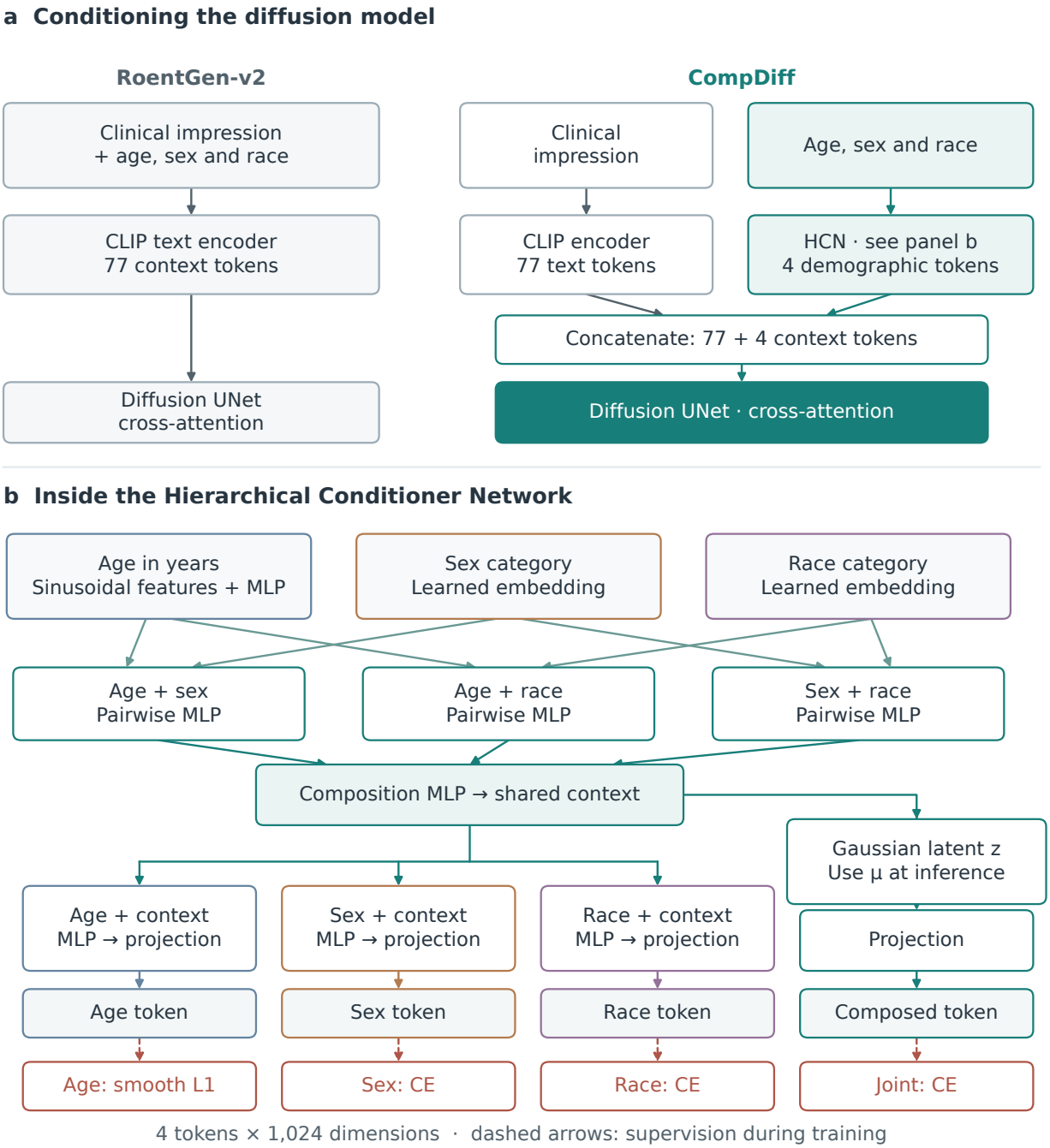


Figure 2: CompDiff architecture and demographic conditioning. a, RoentGen-v2 encodes clinical and demographic text together through CLIP's 77-token context. CompDiff encodes the clinical impression through CLIP and age, sex and race through the Hierarchical Conditioner Network (HCN). HCN supplies 4 additional context tokens, concatenated with the text tokens before diffusion UNet cross-attention. **b,** HCN encodes age continuously and sex and race with separate categorical embeddings. Three dedicated multilayer perceptrons (MLPs) combine the attribute pairs; a composition MLP fuses their outputs into a shared demographic context. Each attribute embedding is concatenated with this context and projected to its own token. A Gaussian latent supplies only the composed token: it is sampled during training and set to its mean at inference. The four projected tokens each have 1,024 dimensions. Auxiliary heads supervise age, sex and race from their respective tokens and the joint demographic label from the composed token. Dashed arrows mark this training-only supervision; CE denotes cross-entropy.

2 Results

We first compared the fidelity and conditioning alignment of CompDiff, RoentGen-v2 and FairDiffusion on chest radiographs and fundus images. Chest X-ray experiments then tested generation at unseen demographic intersections, radiologist assessment, classifier training and classifier auditing using synthetic cohorts. These experiments distinguish three uses of classifiers: measuring whether generated images express the requested labels, learning from synthetic training data, and evaluating a fixed model on synthetic test data.

2.1 Generation fidelity and alignment

CompDiff achieved the lowest mean Fréchet Inception Distance (FID) on both modalities (Figure 3). FID measures the distance between real and synthetic image feature distributions, with lower values indicating closer agreement. Values are means ± standard deviations across three independently trained runs per method and modality. On chest X-rays, CompDiff reached 74.7 ± 5.7, compared with 88.4 ± 4.7 for RoentGen-v2 and 85.6 ± 1.7 for FairDiffusion, a 15% reduction relative to RoentGen-v2. On fundus images, the corresponding values were 60.1 ± 4.3, 72.7 ± 3.7 and 64.2 ± 1.8, a 17% reduction relative to RoentGen-v2.

The chest ordering persisted with features from RadImageNet, pretrained on radiological images[16]: FID-RadImageNet was 6.64 for CompDiff, 8.62 for RoentGen-v2 and 8.85 for FairDiffusion. On fundus, FairDiffusion had the lowest FID-RadImageNet (5.51 versus 5.76 for CompDiff). We therefore distinguish the two feature spaces and use Inception FID as the primary fundus measure, since RadImageNet is out of domain for fundus photography.

Agreement with the requested findings and attributes. Pretrained classifiers assessed whether generated images expressed their conditioning labels. Chest disease alignment was the mean AUROC over five prompted findings: atelectasis, cardiomegaly, oedema, pneumothorax and pleural effusion. CompDiff scored 0.754, compared with 0.720 for FairDiffusion and 0.704 for RoentGen-v2. Fundus alignment used separate classifiers for glaucoma and cup-disc ratio. CompDiff had the highest average score (0.975 versus 0.917 and 0.937), with the best glaucoma AUROC in every training run (0.957 versus 0.930 and 0.916) and a larger margin in cup-disc alignment (0.994 versus 0.904 and 0.957).

Sex accuracy was near ceiling for all methods. CompDiff had lower race accuracy (0.949 versus 0.980 and 0.974) and higher age error (8.66 versus 6.50 and 8.04 years RMSE) than the prompt-conditioned baselines. Neither difference was statistically distinguishable across the three training runs ($p = 0.11$ for race and $p = 0.15$ for age, Welch's tests). These marginal alignment measures complement the intersectional fidelity and evaluations at unseen intersections below.

Distributional coverage and supporting checks. CompDiff also reached more of the real chest image distribution. The joint UMAP projection in Figure 4 shows greater overlap with the cloud of real images; this is a qualitative view of the features, rather than a direct display of FID. Quantitative metrics in the original RadImageNet feature space supported the same finding (Table 2). CompDiff had the highest Precision, the fraction of synthetic samples within the neighbourhoods of real images, and Coverage, the fraction of neighbourhoods of real images reached by synthetic samples[17,18]. RoentGen-v2 had the highest Recall, a related measure that also depends on the

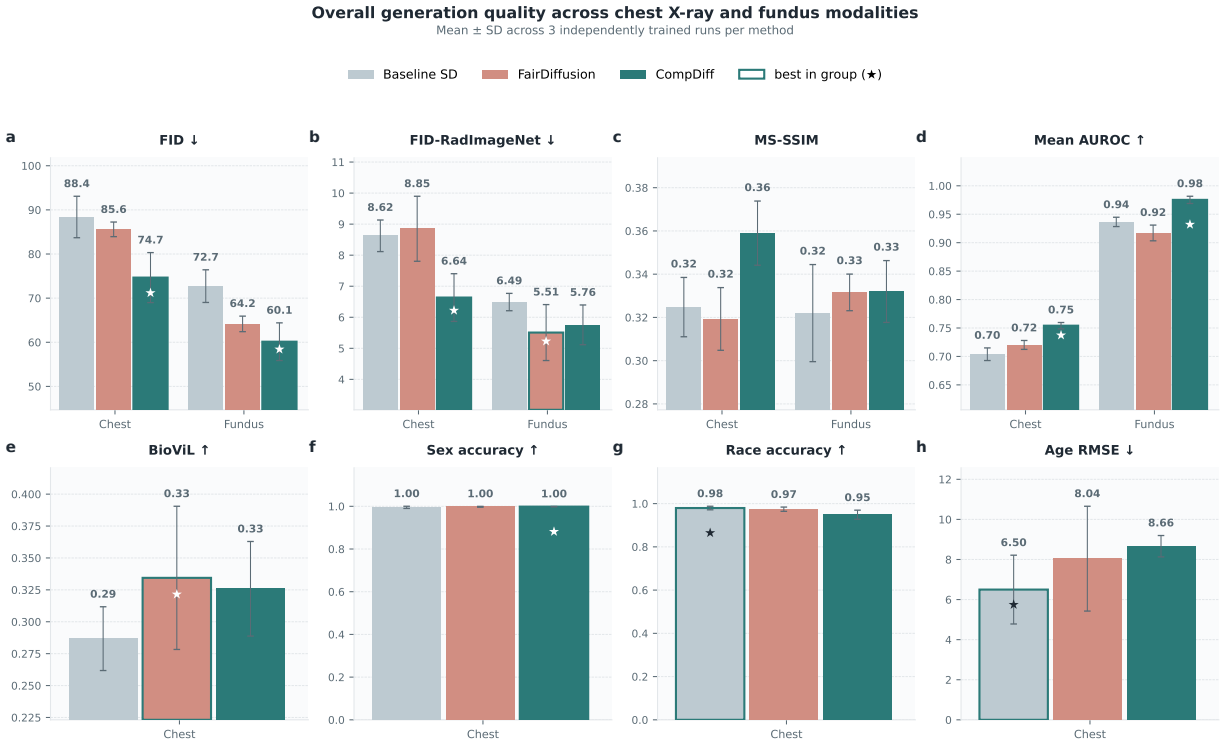


Figure 3: Overall generation quality across chest X-ray and fundus modalities. **a**–**d**, Metrics computed for both modalities, each panel showing a chest group and a fundus group: **a**, FID; **b**, FID-RadImageNet; **c**, MS-SSIM; **d**, mean AUROC. **e**–**h**, Agreement with the requested findings and demographic attributes for chest images: **e**, BioViL; **f**, sex accuracy; **g**, race accuracy; **h**, age RMSE. Each group contains the three methods (Baseline SD, FairDiffusion, CompDiff). A star and a green outline mark the best method within a group, and an upward arrow in a panel title indicates that higher is better and a downward arrow that lower is better; MS-SSIM carries neither, because neither tail is desirable and values are read against the defensible band 0.25–0.75. Chest and Fundus bars are the mean over three independently trained runs per method, evaluated on the corresponding evaluation split, with error bars giving the standard deviation across those runs. Demographic and vision–language classifiers trained on chest images are out of domain on fundus images and are therefore reported for chest only.

Table 2: Distributional fidelity and coverage on the full chest X-ray test set with its natural demographic distribution, computed in RadImageNet ResNet-50 feature space with $k = 10$ manifold neighbours. Higher is better for all four metrics. Precision[17] and Density[18] measure *fidelity* (how often synthetic samples fall inside the real manifold), with Density a refinement of Precision that is less sensitive to outliers. Recall[17] and Coverage[18] measure *diversity* (how much of the real manifold is reached by synthetic samples), with Coverage the replacement for Recall that is robust to mode collapse (uses manifold radii measured on real images rather than synthetic images, so it is not inflated by synthetic outliers).

Method	Precision ↑	Recall ↑	Density ↑	Coverage ↑
RoentGen-v2	0.329	**0.480**	0.072	0.084
FairDiffusion	0.344	0.234	0.070	0.099
CompDiff	**0.396**	0.387	**0.091**	**0.139**

spread of neighbourhoods of synthetic images.

Paired real–synthetic MS-SSIM fell within the reference range 0.25–0.75 for every method and did not distinguish them[5]. Chest BioViL similarity was nearly identical for CompDiff and FairDiffusion (0.326 and 0.334), both above RoentGen-v2 (0.287). Authenticity, which compares synthetic–real proximity

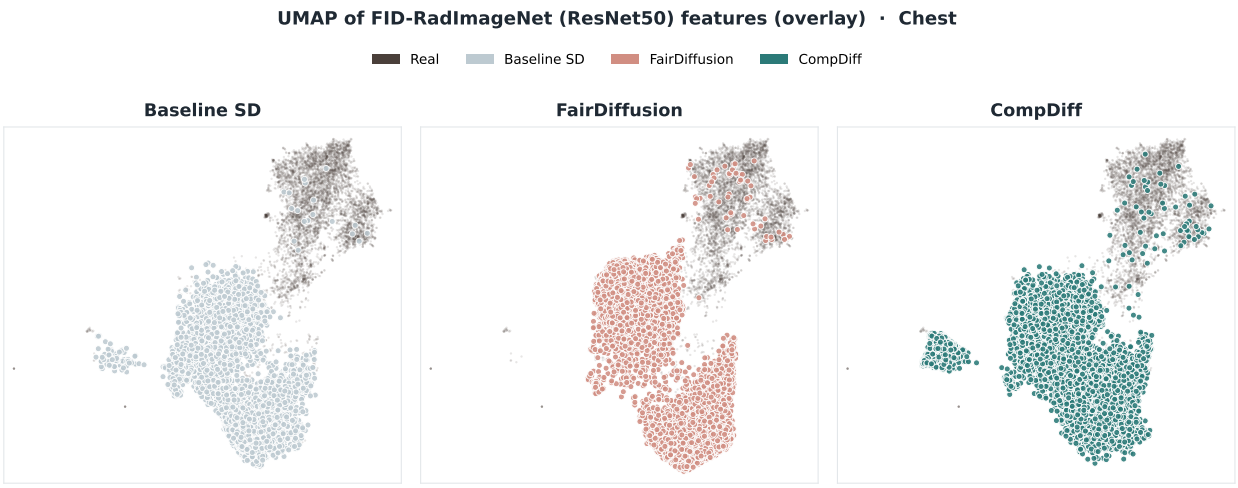


Figure 4: Joint UMAP projection of chest X-ray RadImageNet features. One UMAP was fitted to pooled real and synthetic ResNet-50 features, giving all panels shared axes. The same real cohort appears in warm grey beneath each generator's samples: RoentGen-v2 (Baseline SD, light grey), FairDiffusion (red) and CompDiff (green). CompDiff shows greater overlap with the cloud of real images. This projection is qualitative; Precision, Recall, Density and Coverage in Table 2 quantify the distributions in the original feature space.

with distances between neighbouring real images, was at least 0.76 for all three generators[19].

2.2 Subgroup generation quality

CompDiff's fidelity gains extended across sex, race and age groups on both modalities (Figure 5), measured with FID-RadImageNet for chest X-rays and Inception FID for fundus images. It achieved the lowest distance in 28 of 29 overall, equity-scaled and subgroup comparisons. The exception was the Black fundus stratum, where FairDiffusion's mean was lower by 1.1 FID.

Equity-scaled FID (ES-FID) captures overall fidelity together with variation between subgroups. CompDiff had the lowest ES-FID across sex, race, age and their joint intersection on both modalities. On chest X-rays, intersectional ES-FID was 12.1, compared with 14.2 for RoentGen-v2 and 14.3 for FairDiffusion. On fundus, the intersectional measure was dominated by dispersion at small cell sizes, so the measures for each attribute provide the more reliable comparison.

Comparisons within cells were limited by sample size. Among chest X-ray intersections with at least twenty images, CompDiff had the lowest FID-RadImageNet in all 16 cells, reducing cell distance by 14.9% relative to RoentGen-v2 and 18.1% relative to FairDiffusion (paired comparisons, $p < 0.0001$ for both). The methods could not be distinguished in the more sparsely sampled cells. Restricting the aggregate equity analysis to the adequately sampled cells preserved the method ordering.

2.3 Generation at unseen intersections

To test whether CompDiff could compose demographic intersections absent from training, we removed eight intersections from each of two MIMIC-CXR training variants. Holdout A excluded the eight rarest intersections; Holdout B excluded the eight next rarest. Each method was trained three times per holdout, giving 18 models. FID-RadImageNet compared generated images with real test images from the same withheld intersection, measuring fidelity within each demographic cell. Figure 6 summarises the models' mean distances across their eight withheld cells and the average differences between methods.

CompDiff achieved the lowest mean FID-RadImageNet across three runs in all 16 withheld intersections (Supplementary Figure 2; Supplementary Table 1). Averaged equally across cells and holdouts, the CompDiff minus RoentGen-v2 difference was -2.364, with a pointwise 95% confidence interval of $[-3.710, -1.018]$. The corresponding difference from FairDiffusion was -1.246 $[-2.569, +0.078]$. These intervals use independently trained models as the replicates. Both average differences remained in favour of CompDiff after omitting any single trained model. With Inception features, CompDiff retained the lowest mean across three runs in 12 of 16 cells (Supplementary Figures 3 and 4; Supplementary Note 1).

Complementary controls tested whether sample size or the choice of fidelity measure changed the comparisons within each cell. Because limited numbers of real and synthetic images can inflate FID (Supplementary Figure 5), we adjusted for this estimated bias. The corrected FID preserved the winning configuration in 41 of 42 evaluated cells, including both withheld intersections and those represented in training. Kernel Inception Distance (KID) measures distributional similarity using an unbiased estimator, avoiding the systematic bias from limited sample sizes that affects FID[20]. Computed from the same RadImageNet features, KID preserved the winning configuration in 38 of 42 cells. Normalisation to a common sample size supported the same comparisons. Among the 12 eligible withheld cells, corrected FID preserved the complete configuration ordering in all 12, and KID in 10. The four smallest Holdout A cells, with fewer than 120 images, were excluded (Supplementary Note 2).

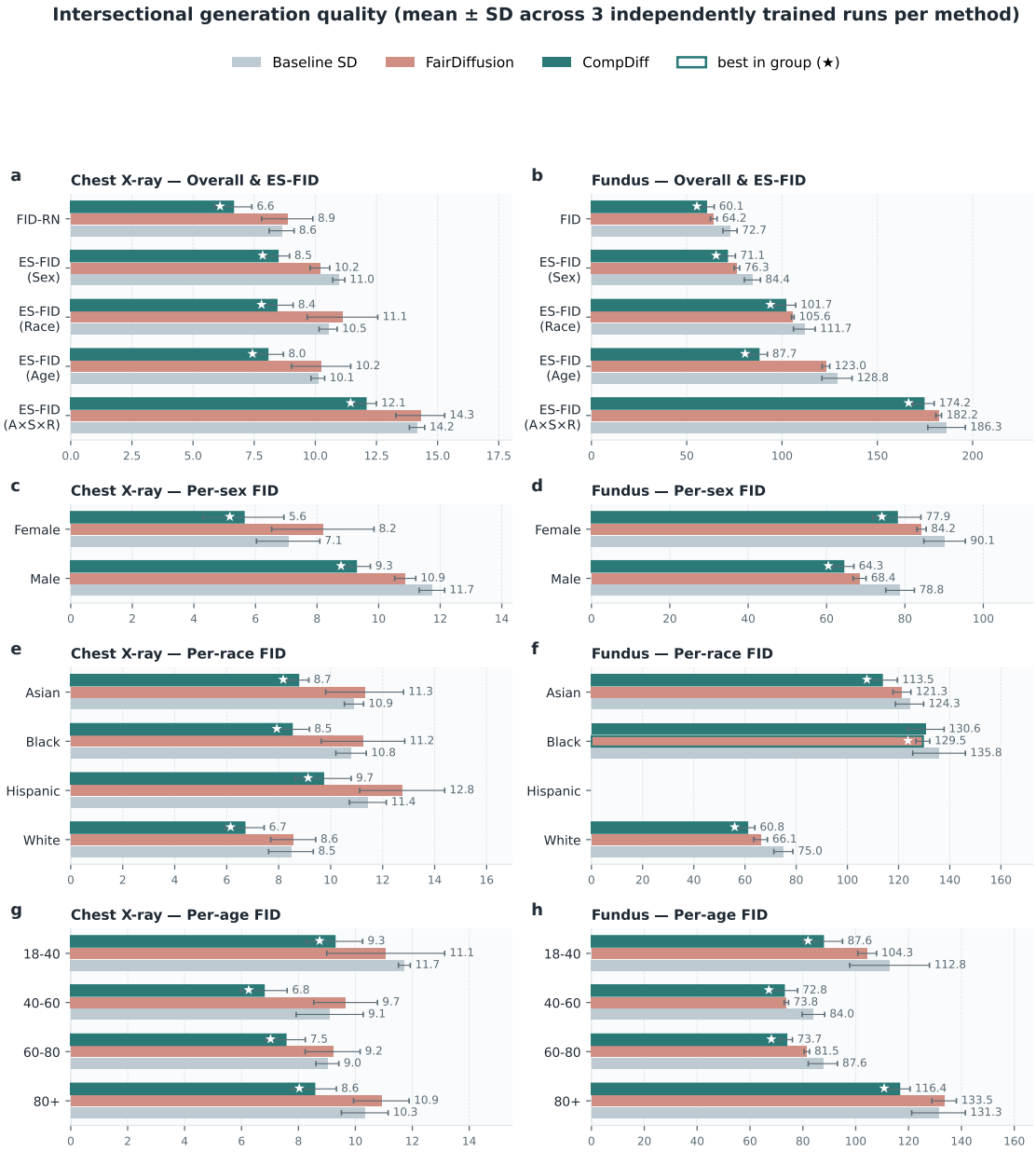


Figure 5: Intersectional generation quality (lower is better). Left column (**a**, **c**, **e**, **g**): chest X-ray; right column (**b**, **d**, **f**, **h**): fundus. **a**, **b**, Overall distance and equity-scaled distance (ES-FID) across sex, race, age, and the full age × sex × race intersection (A×S×R), where the intersectional ES-FID uses every demographic cell as the subgroup list. **c**, **d**, By sex. **e**, **f**, By race. **g**, **h**, Per age bin. Each row groups three methods (Baseline SD, FairDiffusion, CompDiff), and a star marks the best method per row. The chest column reports FID-RadImageNet and the fundus column reports Inception FID, following the rationale for the feature backbone of Section 4.4. Bars are means over three independently trained runs per method, with error bars giving the standard deviation across those runs. The fundus intersectional ES-FID is dominated by its dispersion term, because cell distances at the small fundus cell sizes are several times the overall distance; it is reported for completeness. The Hispanic row is omitted in panel **f** because the FairGenMed test set does not include Hispanic patients.

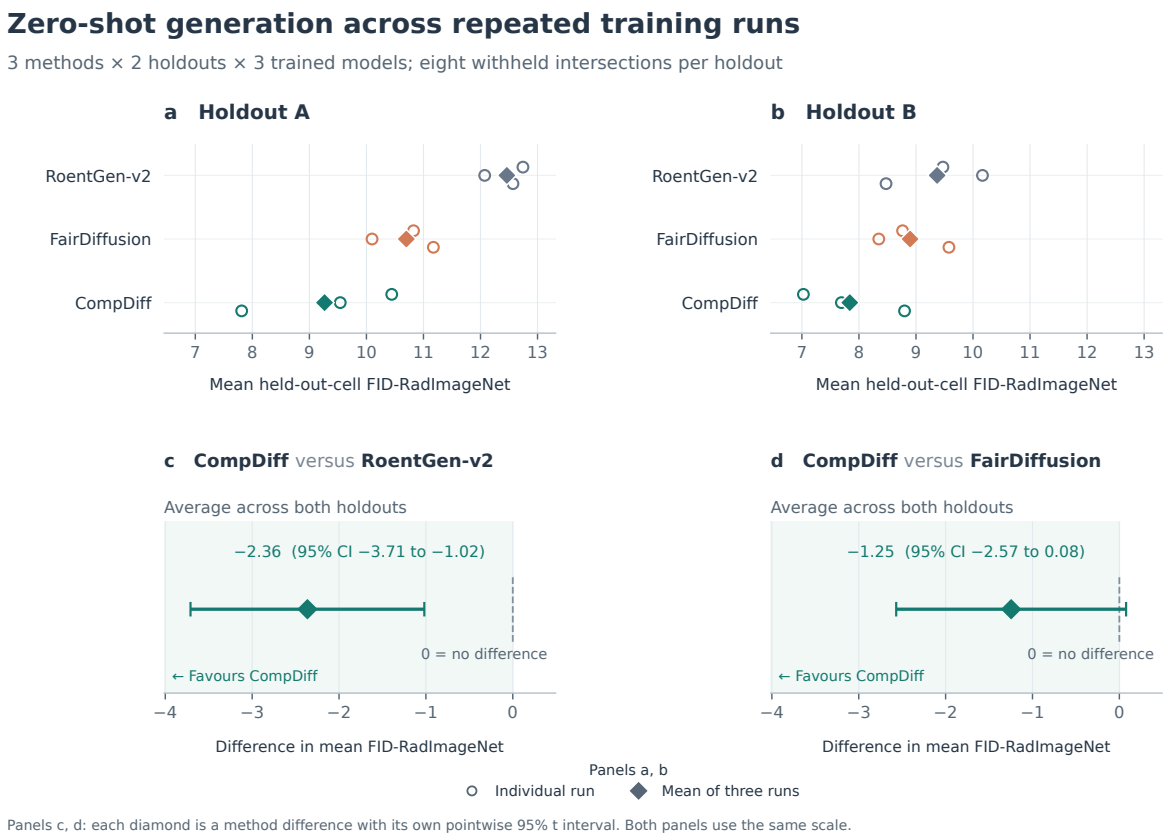


Figure 6: Zero-shot generation across repeated training runs. Each method was trained three times on each of two holdouts, each withholding eight demographic intersections. FID-RadImageNet measures the distance between real and synthetic feature distributions within an intersection; lower values indicate closer agreement. **a,b**, Mean distance across the eight withheld cells for each trained model (open circles), with the mean across three models shown as a filled diamond. Runs are not paired by seed number. **c,d**, CompDiff minus comparator differences, averaged equally across cells and holdouts. Each panel shows one difference and its pointwise 95% Welch–Satterthwaite confidence interval, with a separate zero reference and a common scale. Negative differences favour CompDiff. Each comparison uses twelve trained models; the experiment contains 18 models in total. Intervals describe variation across training, checkpoint selection and generation on the fixed test cohorts. Cell observations and means appear in Supplementary Figure 2.

Additional sampling did not close the gap between RoentGen-v2 and the CompDiff variants. Giving RoentGen-v2 a 4× generation budget and retaining its best quarter under a verifier still left either CompDiff or its unstructured composer variant as the best configuration in each of the 12 assessable withheld cells (Supplementary Note 3). The unstructured variant matched CompDiff's parameter count and retained its separate attribute encoders and supervised output tokens. It preserved fidelity within each cell on both holdouts, showing that generation at unseen intersections did not depend on the factorised composer (Supplementary Figure 6; Section 2.7). Together, these controls support the interface across the tested estimators, sampling budgets and composer choices.

Figure 7 shows samples with matched prompts from common and rare groups in both modalities, together with a withheld chest intersection (80+ Male Asian). Each column uses the same prompt across generators. The reader study below provides an independent assessment of realism and clinical correspondence on the unseen chest intersections.

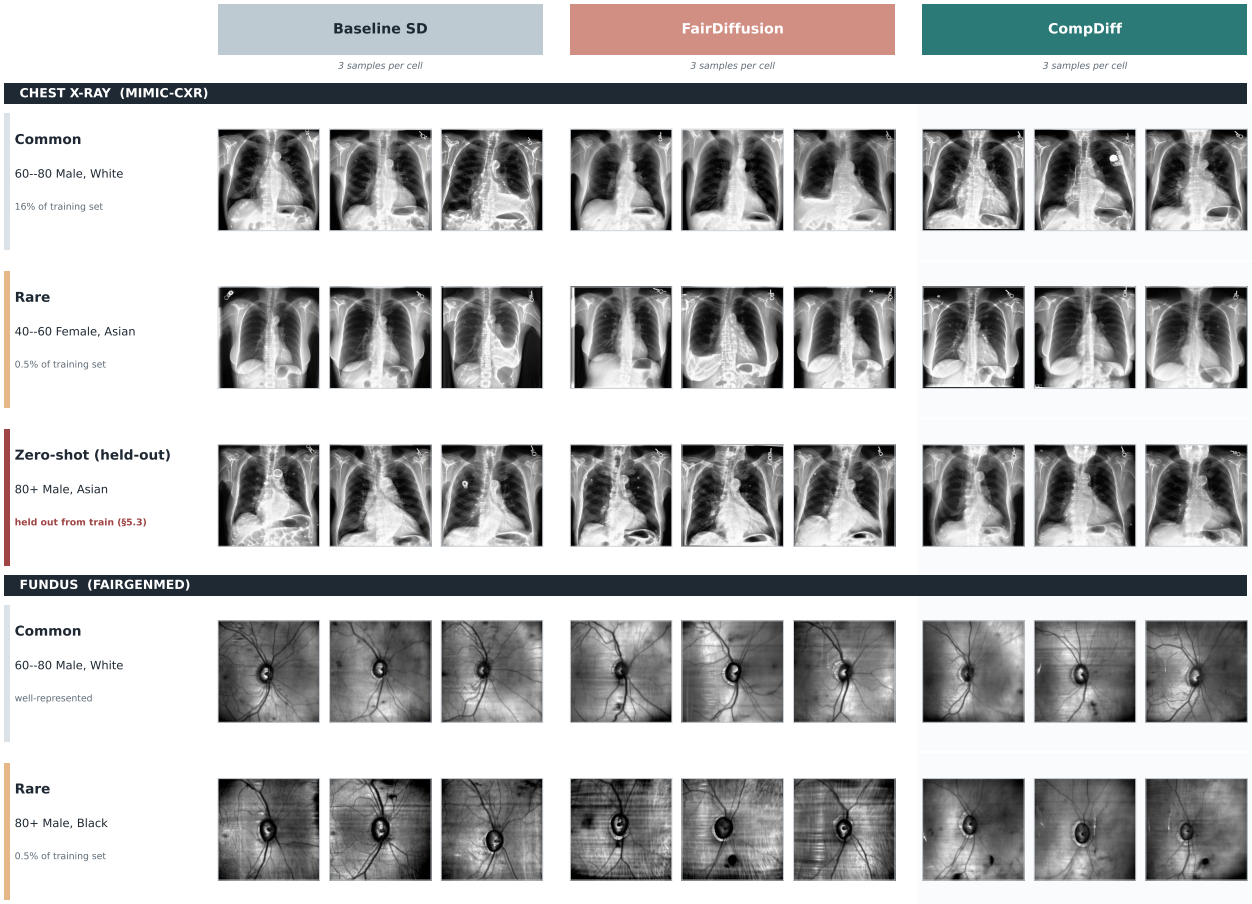


Figure 7: Qualitative samples across tiers of subgroup rarity on both modalities. **Chest X-ray (MIMIC-CXR), top three rows**: *Common* (60–80 Male, White; 16% of training), *Rare* (40–60 Female, Asian; 0.5% of training), and *Zero-shot (withheld)* (80+ Male, Asian; one of the eight intersections held out from training in the zero-shot protocol of Section 4.4). **Fundus (FairGenMed), bottom two rows**: *Common* (60–80 Male, White) and *Rare* (80+ Male, Black). The two Common chest/fundus rows, the Rare chest row, and the Rare fundus row are sampled from generators trained on the natural demographic distribution; the Zero-shot chest row is sampled from the generators of Section 4.4, trained with these intersections held out, so that row probes whether each generator can synthesise a demographic intersection it has never observed in training. Within each row, three samples with matched prompts are shown per method (Baseline SD, FairDiffusion, CompDiff); samples in the same column position across the three methods share the same clinical prompt, allowing visual comparison under the same clinical conditioning. The Zero-shot chest row prompts (all 80+ Male, Asian) are, left to right: a new large mass in the left upper lobe with rightward tracheal deviation concerning for malignancy; moderate cardiomegaly without pulmonary oedema or pleural effusion; and right basal atelectasis without pneumothorax.

2.4 Radiologist assessment of unseen intersections

In blinded assessment by two radiologists, CompDiff achieved the highest mean realism and prompt correspondence scores and was selected most often as the most realistic synthetic image from unseen demographic intersections (Table 3; protocol in Section 4.5; rating interface in Supplementary Figure 7). The readers completed 96 assessments of sets of three images spanning the 16 withheld intersections, covering 92 distinct sets. Mean prompt correspondence with the clinical impression (-2 to $+2$ scale) was $+0.47$ for CompDiff, compared with $+0.01$ for FairDiffusion and -0.05 for RoentGen-v2. Prompt correspondence was higher for CompDiff than for both RoentGen-v2 and FairDiffusion in the mixed-effects analysis, with Benjamini–Hochberg-adjusted p-values of 0.005 and 0.031, respectively.

Table 3: Blinded radiologist assessment of chest X-ray sources on the zero-shot withheld intersections, by two board-certified radiologists (96 rater–set instances over 92 distinct sets, 288 image ratings, across both withheld splits). Every rated image is of a demographic intersection absent from generator training. Realism is scored 1–5 and prompt correspondence from -2 to $+2$ (mean ± SD; unreadable images excluded from realism, unreadable and N/A from correspondence). "Most realistic" is the per-set forced-choice win rate: the fraction of rated sets containing that source in which it was picked as the most realistic image. Real images are a hidden upper anchor present in a subset of sets.

Source	**Realism ↑** (1–5)	**Prompt corr. ↑** (-2 to $+2$)	**Most realistic ↑** win rate
CompDiff	**4.17** ± 1.11	**+0.47** ± 1.72	**48%**
RoentGen-v2	2.36 ± 1.33	-0.05 ± 1.69	7%
FairDiffusion	4.13 ± 1.05	+0.01 ± 1.74	37%
Real (hidden anchor)	4.92 ± 0.29	+0.75 ± 1.86	92%

Anatomical realism ranged from 1 (clearly artefactual) to 5 (indistinguishable from a real scan). Mean scores were 4.17 for CompDiff, 4.13 for FairDiffusion and 2.36 for RoentGen-v2 (Section 4.9). Readers selected CompDiff as the most realistic image in 48.4% of the sets in which it appeared, compared with 36.8% for FairDiffusion and 6.7% for RoentGen-v2. The direct preference comparison with FairDiffusion also favoured CompDiff numerically, with 44 selections to 35 ($p = 0.37$; Supplementary Table 6). Hidden real images provided an upper reference, with mean realism 4.92 and 11 selections as most realistic in 12 appearances.

2.5 Downstream classification

We first compared the utility of images from all three generators for classifier training on chest X-rays and fundus images, evaluating the trained classifiers on real test images. Classifiers trained on CompDiff images achieved higher diagnostic discrimination and lower underdiagnosis rates on real chest X-rays across the overall, sex, race and age evaluations (Figure 8). Overall AUC was 0.72, compared with 0.69 for RoentGen-v2 and 0.68 for FairDiffusion; underdiagnosis rates were 0.40, 0.46 and 0.44, respectively. Underdiagnosis measures how often an image with disease is classified as having no finding. The sex, race and age panels summarise performance along each demographic attribute. On fundus, CompDiff achieved the highest AUC in 3 of 4 comparisons. The exception was the age evaluation, where RoentGen-v2 had higher AUC (0.66 versus 0.61), while CompDiff had the lowest equalised-odds difference (0.28, compared with 0.39 for RoentGen-v2 and 0.47 for FairDiffusion). Equalised-odds difference measures the larger of the gaps between groups in true positive and false positive rates, with lower values indicating smaller disparities.

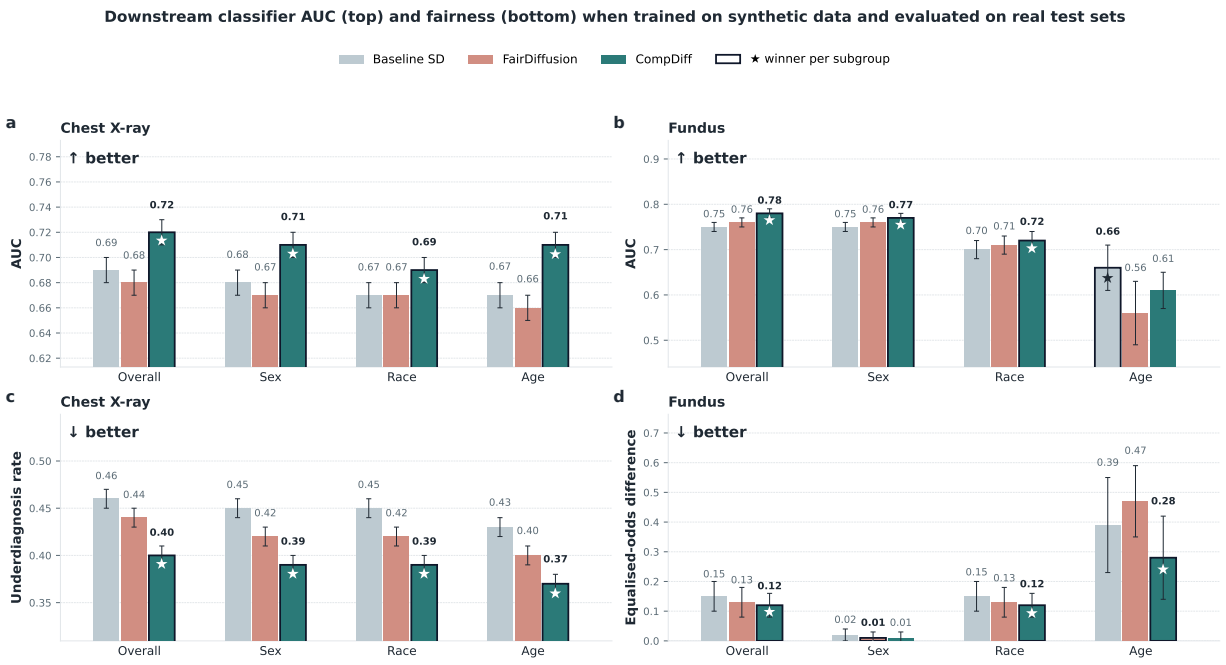


Figure 8: Downstream classifier performance when trained on synthetic data and evaluated on real test sets. **a**, **b**, AUC (higher is better) on chest X-ray (**a**) and fundus (**b**). **c**, **d**, Subgroup fairness (lower is better): underdiagnosis rate for chest (**c**) and equalised-odds difference for fundus (**d**). Each panel reports an overall result and separate summaries for sex, race and age and shows three methods (Baseline SD, FairDiffusion, CompDiff). Error bars are standard deviation across runs. The winning method per subgroup is outlined in dark slate and its label is bold. CompDiff is the winner on every chest X-ray (subgroup, metric) cell and on 7/8 fundus cells.

Gains across training configurations and diagnostic labels. A separate chest X-ray experiment compared CompDiff with RoentGen-v2 across independent classifier runs. The four synthetic corpora followed the natural demographic distribution at 1×, 2× and 3× the real training set size, or approximately balanced demographic intersections. When synthetic images were the sole source of chest training data, four classifier runs per corpus gave mean AUROC_8 values of 0.7962 for CompDiff versus 0.7849 for RoentGen-v2 on MIMIC-CXR, and 0.7363 versus 0.7275 on external CheXpert. AUROC_8 averages the same eight diagnostic labels on both datasets. Weighting the four corpora equally, the differences were +0.0113 and +0.0088, with pointwise 95% confidence intervals of $[0.0067, 0.0158]$ and $[0.0039, 0.0137]$, respectively (Figure 9a). CompDiff had the higher mean in every corpus regime on both datasets (Supplementary Figure 8).

The advantage extended to classifiers pretrained on synthetic images and then fine-tuned on real MIMIC-CXR data. Across the 48 matched configurations, CompDiff improved AUROC_8 in 40 on MIMIC-CXR and 40 on CheXpert. This included 23 of 24 configurations with pretraining from scratch and 17 of 24 with ImageNet initialisation on each dataset (complete sweep in Supplementary Figures 9 and 10). The comparisons used the matched AdamW recipe at 224×224 resolution (Table 4; Section 4.6).

Mean AUROC gains were positive for all 8 diagnostic labels on both test sets (Figure 9b). Pneumothorax showed the largest external gain, +0.0140 AUROC on CheXpert. An exploratory comparison of the test ROC curves made the gain more interpretable: at 90% specificity, mean pneumothorax sensitivity increased from 67.6% to 70.7% on MIMIC-CXR and from 39.8% to 42.6% on CheXpert (Figure 9c). Mean sensitivity gains were positive for all eight labels. Supplementary Table 13 also

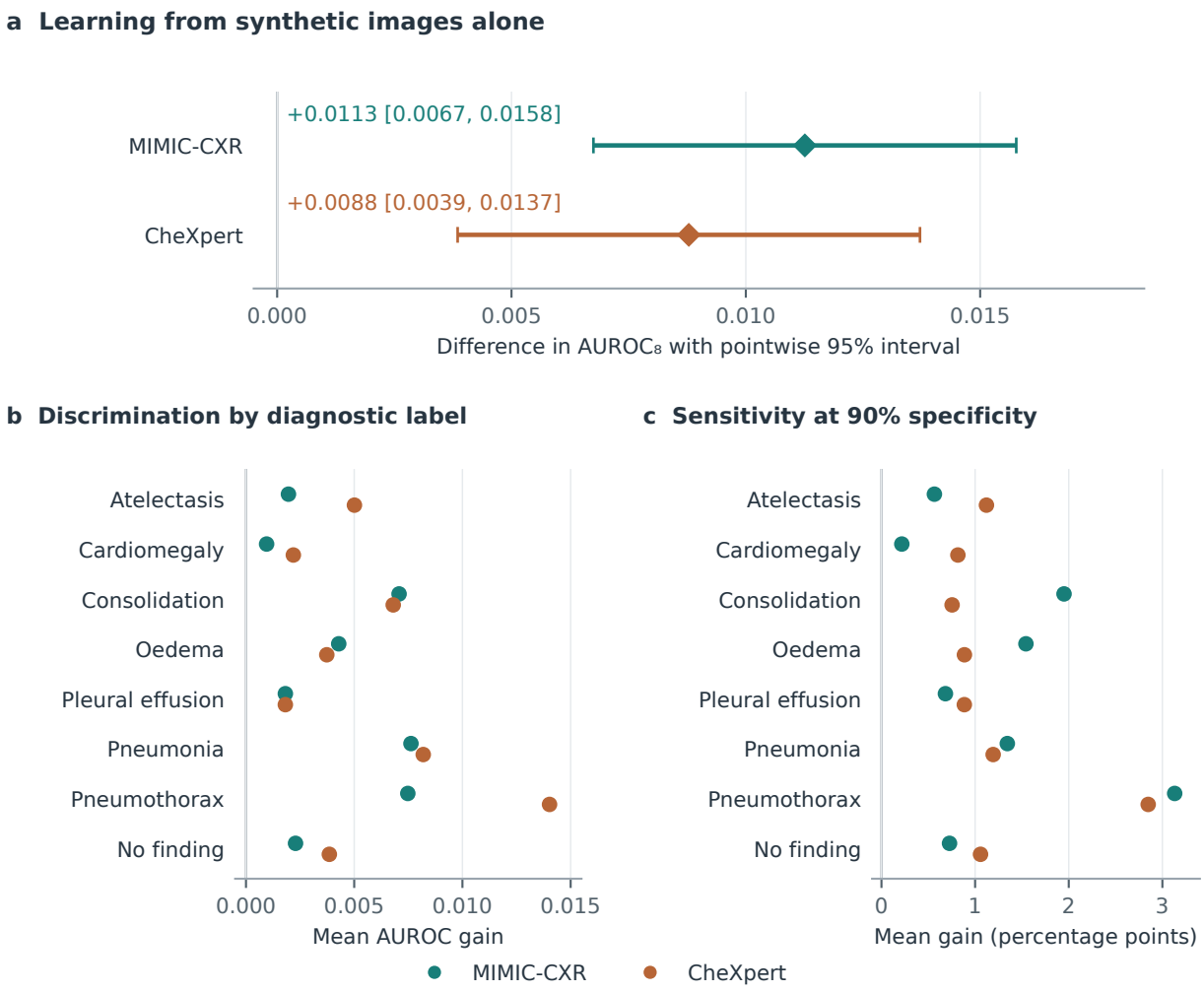


Figure 9: Diagnostic utility of CompDiff images relative to RoentGen-v2. All panels use the same eight diagnostic labels. **(a)** $AUROC_8$ difference for classifiers trained on synthetic chest images alone. Diamonds show the contrast averaged equally across four fixed corpus regimes; bars are pointwise 95% Welch confidence intervals based on four independent classifier runs per generator and regime, conditional on the corpora and real test cohorts. **(b)** Mean AUROC difference for each label across all 48 matched configurations with pretraining followed by fine-tuning. **(c)** Mean sensitivity difference at 90% specificity over the same configurations. Sensitivity is read separately from each test ROC for this exploratory comparison; it does not represent a validated clinical threshold. Panels (b) and (c) are descriptive averages over configurations that share training data and test cohorts. All differences are CompDiff minus RoentGen-v2; positive values favour CompDiff. Colours identify the test dataset. The same trained models are evaluated on both datasets.

reports precision–recall performance and the comparison at 95% specificity.

CompDiff raises the floor on underserved subgroups. The overall gains extended across demographic groups. Across the 48 training configurations, median $AUROC_8$ gains were positive for all 10 sex, race and age groups separately on MIMIC-CXR and CheXpert (Figure 10a). We then identified the subgroup with the lowest RoentGen-v2 AUROC within each demographic attribute and measured the CompDiff gain in that same subgroup. Median gains were positive for sex, race and age on both datasets (Figure 10b). Across the two datasets, these gains were positive in 84.4%, 87.5% and 71.9% of configuration evaluations, respectively. The improvements therefore reached the groups with the lowest baseline performance as well as the overall average. Subgroup sample sizes and patient resampling checks for the sparsely sampled groups are reported in Supplementary Note 8.

Table 4: Downstream classifier training recipes. Each recipe is evaluated for both CompDiff and the RoentGen-v2 baseline. The natural synthetic corpora match the demographic distribution of the MIMIC-CXR training split, while the balanced corpus is expanded to approximately uniform coverage of (age, sex, race) cells. The pretrain → real fine-tune family is additionally swept over six fractions of real data (1%, 10%, 30%, 50%, 70%, 100%), giving 48 matched configurations per evaluation dataset. Balanced-corpus counts are 414 980 for RoentGen-v2 and 415 072 for CompDiff.

Family / regime	**Pretraining corpus**	**Images**	**Initialisation**
Training on synthetic data alone (no real chest images)			
1× natural	–	62,094	ImageNet
2× natural	–	124,188	ImageNet
3× natural	–	186,282	ImageNet
Balanced	–	≈ 415,000	ImageNet
Pretrain → real fine-tune (FT fractions 1%–100%)			
1× natural	62,094 synthetic	62,094	Random; ImageNet
2× natural	124,188 synthetic	124,188	Random; ImageNet
3× natural	186,282 synthetic	186,282	Random; ImageNet
Balanced	≈ 415,000 synthetic	≈ 415,000	Random; ImageNet

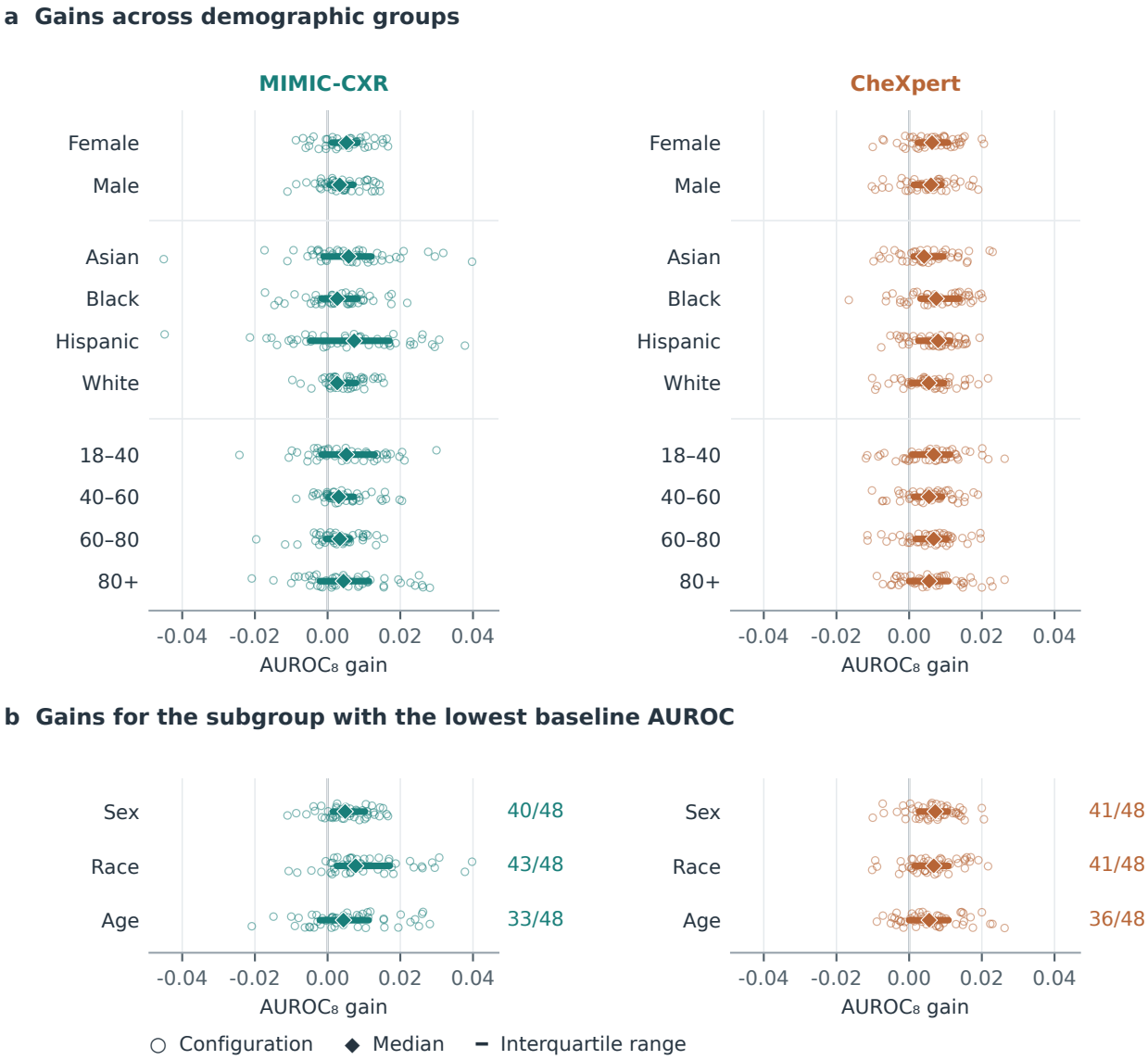


Figure 10: Subgroup gains across the complete downstream training sweep. **(a)** CompDiff minus RoentGen-v2 $AUROC_8$ for each sex, race and age group, shown separately for MIMIC-CXR and CheXpert. **(b)** Gain for the subgroup with the lowest RoentGen-v2 AUROC within each demographic attribute in each configuration, retaining its identity under CompDiff. Hollow circles show the 48 configuration differences; filled diamonds and thick bars show medians and interquartile ranges. These ranges describe variation across configurations, not confidence intervals. Counts in (b) give configurations with a positive gain. Every subgroup score averages the same eight labels, each with at least one positive and one negative image. Configurations share training data and test cohorts. Sample sizes and uncertainty checks for sparse groups are in Supplementary Tables 14 and 15.

2.6 Synthetic cohorts for fairness auditing

We tested two uses of synthetic cohorts for classifier evaluation: reproducing audit results from real data at overall and subgroup levels, and estimating performance on rare intersections where real test data are sparse. Each audit used a single DenseNet-121 trained on real MIMIC-CXR and held fixed during evaluation. Results are estimates from a single classifier, without replication across training seeds. The two designs are shown in Figures 11 and 12; protocols are given in Section 4.6.

Agreement with audits using real data. Synthetic cohorts matched the real test split in size and demographic distribution. CompDiff reproduced the real underdiagnosis rate most closely, with an absolute error of 0.097, compared with 0.235 for RoentGen-v2 and 0.284 for FairDiffusion. Its overall AUROC_8 error was also lower: 0.010, compared with 0.059 and 0.066, respectively. Averaging all 14 MIMIC-CXR labels, rather than the eight shared with CheXpert, gave AUROC_{14} errors of 0.027, 0.077 and 0.085. Figure 11a reports these overall comparisons.

Agreement extended to the subgroup audits in Figure 11b,c. Mean absolute AUROC error across marginal sex, race and age groups was 0.027 for CompDiff, compared with 0.069 for RoentGen-v2 and 0.083 for FairDiffusion. The largest subgroup error was also smallest for CompDiff: 0.075, compared with 0.128 and 0.135, respectively. CompDiff had the lowest error in every comparison of individual attributes and attribute pairs, although the difference from FairDiffusion for Age×Race was within rounding (0.077 versus 0.077).

Estimation on rare intersections. To estimate the same 14-label macro AUROC within each cell, we used a larger, independent real reference split, excluded from generator training, validation and model selection, and from classifier training. Each synthetic cohort matched this reference in demographic distribution and cell size. Comparisons included the 27 intersections for which all cohorts yielded an AUROC estimate, grouped by reference cohort size.

On the 9 rarest cells, mean absolute estimation error was 0.023 for CompDiff, 0.030 for FairDiffusion and 0.037 for RoentGen-v2, compared with 0.074 for the small real test cohort (Figure 12a). CompDiff reduced error by 69% relative to the real test cohort, which contained only 6–47 images per cell on this tier. The advantage depended on cell size: real test data gave the lowest error on the 9 most abundant cells (0.010 versus 0.038 for CompDiff), while FairDiffusion performed best on the middle tier (0.016 versus 0.028).

Synthetic cohorts also gave more precise estimates on rare cells. With 1000 bootstrap iterations, mean 95% confidence interval width on the rarest tier was 0.110 for CompDiff, compared with 0.167 for the small real test cohort, 0.122 for RoentGen-v2 and 0.127 for FairDiffusion (Figure 12b,c). The real reference cohort had the narrowest intervals, with a mean width of 0.095.

The 60–80 Female Asian cell illustrates the difference. With 51 real test images, estimated AUROC was 0.878 with a 95% confidence interval of $[0.82, 0.93]$, excluding the reference estimate of 0.783 from 211 images. CompDiff gave an estimate of 0.777 with a narrower interval, $[0.74, 0.82]$, that included the reference estimate.

These findings support synthetic cohorts as a complement to audits using real data for sparsely represented intersections. Real cohorts remain necessary for final validation and performed best on abundant cells.

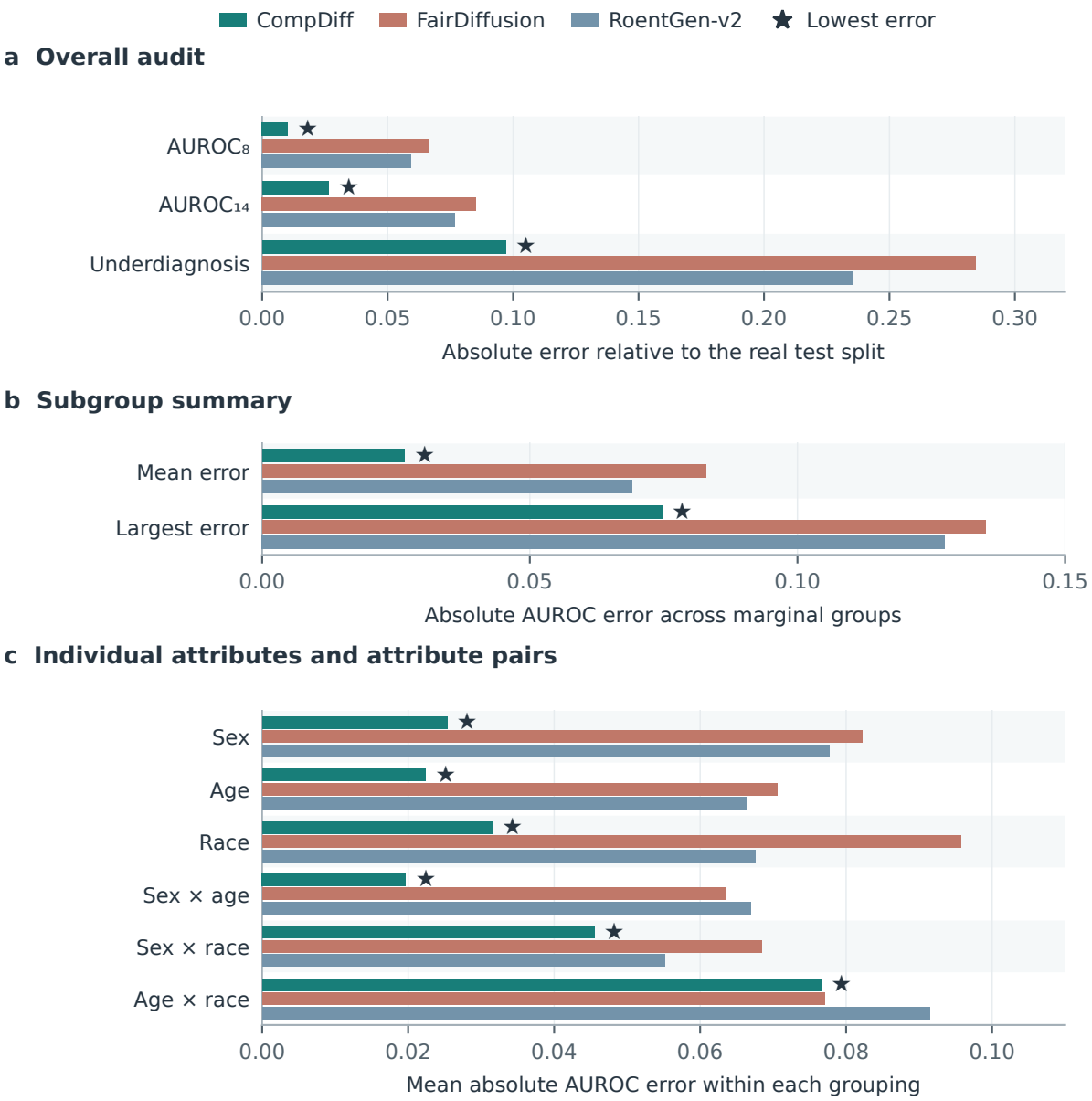


Figure 11: Agreement between synthetic and real test cohorts when auditing a fixed DenseNet-121 trained on real data. Each synthetic cohort matches the real test split in size and demographic distribution. Bars show absolute differences from the real test result; lower values indicate closer agreement. **a,** Overall errors for AUROC_8, AUROC_{14} and underdiagnosis rate. **b,** Mean and largest absolute AUROC errors across marginal sex, race and age groups. **c,** Mean absolute AUROC error within each attribute and attribute pair. Colours identify the generator. A star marks the lowest observed error in each comparison, selected before rounding. All comparisons use a single fixed classifier.

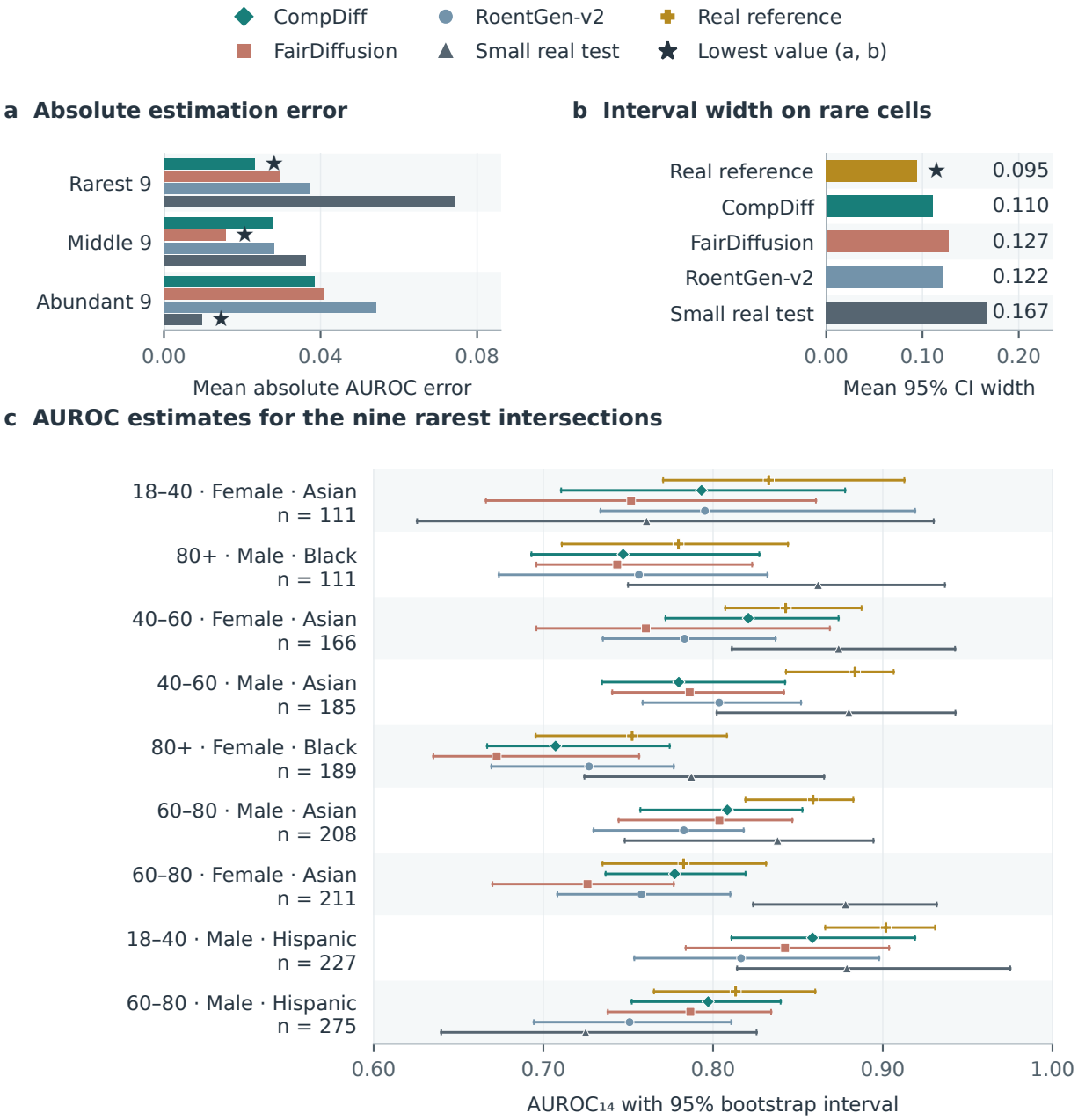


Figure 12: Accuracy and precision of classifier evaluation on rare demographic intersections. A single DenseNet-121 trained on real data is evaluated using macro AUROC over 14 labels. The independent real reference split is excluded from training and model selection. Each synthetic cohort matches its demographic distribution and cell sizes; the small real test cohort contains fewer images. **a,** Mean absolute difference from the reference AUROC across three rarity tiers of 9, 9 and 9 cells, ordered by reference cohort size. Grouped bars show the four estimators, with a star marking the lowest observed error in each tier. Lower error indicates greater accuracy relative to the reference. **b,** Mean 95% confidence-interval width on the 9 rarest cells. A star identifies the narrowest mean interval among all five cohorts, including the real reference. Narrower intervals indicate greater precision. **c,** AUROC estimates and 95% bootstrap intervals for each of those cells, with the real reference marked by a plus. The displayed n is the reference cell size. Intervals use 1000 bootstrap iterations. The same nine cells contribute to panels b and c. These are estimates for one classifier; panels a and b do not quantify variation across classifier training runs.

Table 5: Ablations, grouped by design question and ordered by effect size. All values are validation metrics at each run's own selected checkpoint, chosen by the rule used throughout this paper (Eq. 9: lowest validation FID among checkpoints meeting the controllability floors). Where a run has no eligible checkpoint we report its checkpoint with the lowest age RMSE, since in those runs the age floor is unreachable over the entire trajectory and so disqualifies every checkpoint. Rows marked ‡ were measured under the configuration in which age is conditioned through the text prompt and only sex and race pass through the conditioner; they are comparable among themselves and to the ‡ reference row. *Floors* states whether the reported checkpoint meets the floors, and in each ablated row the value that decides its verdict is set in bold. FID-RN abbreviates FID-RadImageNet; AUROC is the XRV disease classifier AUROC on the synthetic images (prompt alignment, Section 4.4), not a downstream classifier.

Variant	FID↓	FID-RN↓	Sex↑	Race↑	Age↓	AUROC↑	Floors
Where demographics enter							
Demographics in the prompt‡	**94.5**	8.90	0.99	0.99	5.79	0.75	✓
No demographic conditioning‡	70.9	6.30	**0.52**	0.68	17.40	0.78	×
Second text branch‡	**140.0**	9.60	0.75	0.50	20.70	0.74	×
Route through timestep modulation	72.0	6.84	0.99	0.96	8.76	0.77	✓
Where supervision lands							
No output supervision‡	80.3	7.33	**0.51**	0.69	18.10	0.70	×
Token dropout instead of supervision‡	80.1	7.03	**0.51**	0.71	16.60	0.72	×
Supervision before projection‡	79.8	–	**0.54**	0.68	17.90	0.71	×
Single composed token	123.7	17.23	0.98	0.90	**12.41**	0.63	×
How attributes are represented							
Fused demographic embedding‡	80.6	5.70	**0.50**	0.70	17.20	0.70	×
Binned age embedding‡	74.8	5.50	0.96	0.87	**15.31**	0.76	×
How attributes are combined							
Factorised composer (matched pair)	71.2	6.34	1.00	0.96	8.98	0.76	✓
Unstructured composer (matched pair)	80.5	5.28	0.99	0.96	8.51	0.74	✓
Transformer composer	113.5	14.54	0.95	0.94	**11.33**	0.67	×
Loss terms							
No KL regularisation‡	77.1	7.54	0.99	0.95	8.16	**0.73**	✓
No variational latent‡	77.6	–	1.00	0.96	**10.10**	0.74	×
No compositional consistency‡	**88.0**	9.15	0.97	0.94	8.00	0.75	×
Strong compositional consistency‡	**97.1**	11.59	0.99	0.94	10.80	0.72	×
Reference configurations							
Age in prompt; sex and race in conditioner‡	75.5	7.19	0.99	0.96	8.75	0.76	✓
CompDiff (final)	71.6	6.06	1.00	0.96	8.96	0.75	✓

2.7 Conditioning interface ablations

Chest X-ray ablations tested which parts of CompDiff supported image fidelity while preserving demographic control (Table 5). Control was assessed with pretrained demographic predictors: eligible checkpoints required sex accuracy of at least 0.98, race accuracy of at least 0.92 and age RMSE no greater than 10 years. We compared fidelity at each variant's selected checkpoint and used matched composer controls to separate the contribution of the interface from its internal wiring.

Separating demographics from clinical text. Encoding demographics in the prompt preserved sex and race accuracy (0.99 and 0.99) but gave poor fidelity (FID 94.5). Removing demographics improved FID to 70.9 while reducing sex accuracy to chance (0.52). Fewer than 0.02% of training prompts exceeded CLIP's 77-token limit, suggesting that this trade-off involved competition within the text representation rather than widespread truncation (Supplementary Figure 11; Supplementary Note 9). A second text branch did not resolve it, reaching FID 140.0 and sex accuracy 0.75. The dedicated conditioner provided control with substantially better fidelity. Adding a route through

timestep modulation alongside its cross-attention tokens also met the control thresholds, at FID 72.0 versus 71.6 for the final model; cross-attention alone was sufficient in this comparison.

Preserving each attribute at the output. Removing output supervision reduced sex accuracy to 0.51; token dropout did not recover it (0.51). Applying supervision before projection also failed (0.54), consistent with demographic information being lost before it reached the UNet. Separate output tokens mattered: a single composed token failed the age and race thresholds even at its closest checkpoint (12.41 years and 0.90). The attribute encoders also contributed. A fused demographic embedding receiving the same supervision reached only chance sex accuracy (0.50), while discretising age worsened age RMSE to 15.31 years and race accuracy to 0.87. Together, these controls support separate encoders for age, sex and race, separate output tokens and supervision after projection.

Composer choice and convergence. Replacing pairwise composition with an unstructured MLP while retaining that interface preserved both control and fidelity. The matched composers differed in parameter count by only 0.069%. At their selected checkpoints, neither was uniformly better: the unstructured composer had higher FID but lower FID-RadImageNet and age RMSE. At a common training step, differences on the test set remained within the measured variation between independent runs of the final configuration. The unstructured control also preserved fidelity within each cell on both holdout splits, with differences varying in direction (Supplementary Note 4). Thus, pairwise factorisation was not required for generation at unseen intersections in these experiments.

Factorisation did improve convergence: its runs first met all three control thresholds at steps 7,500–10,000, compared with step 11,000 for the unstructured control. Composer choice still required care. The transformer variant failed the age threshold (11.33 years); its matched control using timestep modulation met it (8.76 years). The evidence therefore supports more than one effective composer, without implying that every architecture is interchangeable.

Regularisation and the variational latent. Removing KL regularisation retained age control but reduced disease alignment, whereas removing the variational latent also worsened FID to 77.6 and age RMSE to 10.1 years. The stochastic latent therefore contributed beyond its KL penalty. Removing compositional consistency worsened FID to 88.0, and strengthening it worsened FID further to 97.1. The default weight $\lambda_{\text{comp}} = 0.1$ gave the better balance among the tested settings.

3 Discussion

CompDiff improved generation for demographic intersections that were poorly represented or absent in training, and made the resulting synthetic cohorts more useful for classifier training and fairness auditing. Across chest radiographs and fundus images, the gains extended beyond average fidelity to subgroup generation quality and clinical alignment. With three training runs per method and holdout, CompDiff had the lowest mean FID-RadImageNet in all 16 unseen chest intersections. The average differences favoured CompDiff over both RoentGen-v2 and FairDiffusion and remained favourable when any single trained model was omitted. A separate blinded radiologist assessment supported the realism and clinical correspondence of CompDiff images at unseen intersections. These findings address a practical constraint in medical imaging: the groups for which additional data are most needed may also be those a conventional generator represents least well.

For classifier development, CompDiff improved what classifiers could learn from synthetic chest images. Repeated training supported an advantage when synthetic images were the sole chest training source, and fine-tuning on real images extended the gains across diagnostic labels on both MIMIC-CXR and external CheXpert. Higher pneumothorax sensitivity at a common specificity helps interpret the discrimination gain, although clinical operating thresholds would require independent validation. Improvements also reached the demographic groups with the lowest baseline performance, supporting the use of synthetic pretraining to improve performance for underserved groups. Mixing real and synthetic images during training did not show a consistent advantage (Supplementary Table 11). These findings support the matched training setting studied here and motivate testing which training choices sustain the benefit.

The auditing experiments suggest a second use: constructing evaluation cohorts for intersections too small to support reliable estimates from the available real test split. CompDiff reproduced overall and subgroup audit measurements more closely than the two baselines. On the rarest intersections, its cohorts also gave lower AUROC estimation error and narrower confidence intervals than the small real test cohort. Accuracy and precision are separate findings here: a narrow interval alone would not show that an estimate was close to the real reference. Each synthetic cohort matched a larger, independent real reference in size and demographic distribution, so the experiment evaluates the practical value of generating additional samples. The advantage was concentrated in rare cells; real test data remained the better estimator on abundant cells. Synthetic audits can therefore complement evaluation on real data, with their accuracy checked against an independent real reference where available.

The chest ablations help explain these benefits. Demographic control depended on preserving each attribute through a dedicated encoder, a separate output token and supervision at the point where the UNet received that token. A fused demographic embedding failed despite receiving the same auxiliary supervision. By contrast, an unstructured composer retained fidelity and control, including on unseen intersections, when it preserved the separate encoding and supervision of each attribute. HCN's pairwise factorisation improved convergence but was not necessary for the observed compositional capability. This separates the supported design principle from the particular composer used to implement it. HCN was trained jointly with the unfrozen UNet and text encoder; its use with other backbones or training regimes remains to be evaluated.

This contribution complements existing approaches to synthetic medical data. FairGen and FairDiffusion intervene in how training favours outputs or demographic groups[7,8]; RoentMod uses counterfactual pathology edits to address shortcuts[9], and ChexGen supports broader controllable chest radiograph synthesis[10]. CompDiff instead tests a demographic conditioning interface on intersections deliberately removed from training. It also differs from methods that compose conditional scores during sampling or regularise their independence[12,21]: attributes are combined inside one conditioner, which can learn their interactions without an independence assumption. The comparable unstructured composer is consistent with evidence that factorised representations can emerge during diffusion training[22]. At the data level, combining examples from parent groups offers another approach to augmentation of minority intersections[23]. Matched comparisons with FairGen, RoentMod and ChexGen remain open; the present study does not establish superiority over those systems.

Race requires particular care in interpreting these results. Here it denotes self-reported or administratively recorded categories supplied with the datasets, treated as sociotechnical labels rather than biological ground truth. Race alignment measures the recovery of image features associated with the dataset by a pretrained classifier. Such features may reflect acquisition sites and protocols, disease

prevalence, treatment patterns or other correlates that are not fully characterised[2]. A generator can reproduce those associations along with useful image content. Consequently, demographic alignment does not establish biological fidelity or fairness, and findings from synthetic cohorts require revalidation when transferred to populations with different acquisition and prevalence patterns.

Several limitations define the next steps. Three training runs per method and holdout provide limited precision for estimating variability. Their intervals are conditional on the two selected holdout tasks and fixed real cohorts, and combine training, checkpoint selection and generation variability. Excluding runs with different training and validation schedules preserved the direction of the zero-shot comparisons (Supplementary Note 1). The conditioner currently handles age, sex and race; richer clinical and acquisition attributes would test how well the interface scales beyond this setting. The downstream sweep shares synthetic corpora and real test cohorts across configurations; its configuration counts describe consistency within that sweep. Some subgroup labels occur in few patients, limiting precision even when their median gains are positive. Classifier audits used a single fixed model per experiment, so their conclusions require replication across classifier training runs. CheXpert provides evidence of external transfer for downstream training, but broader evaluation across institutions, scanners and labelling conventions is needed for both generation and auditing.

The reader study involved two radiologists with little overlap between their assigned sets. Treating readers as fixed effects supports inference across cases, not across the wider radiologist population; a larger panel with substantial overlap would allow assessment of inter-rater reliability and more precise subgroup judgements. Prompt correspondence from a single image also cannot resolve every statement about interval change. Finally, small reference cohorts affect fidelity estimation as well as classifier auditing. Rare cells may have both fewer and more homogeneous images, and FID estimated from finite samples depends on cohort size. The robustness analyses supported the method ordering within each cell, but do not justify ranking the intrinsic difficulty or fidelity of one demographic intersection against another.

The central implication is that preserving and supervising demographic attributes through a dedicated conditioning interface can make synthetic medical images more useful where real data are scarce. In the settings studied, that interface supported generation at unseen intersections, improved classifier training for underserved groups and enabled more accurate evaluation of rare subgroups.

4 Methods

4.1 Study design, datasets and ethics

We evaluated generation quality and downstream classifier performance on chest radiographs and fundus images. Chest radiographs were also used for the reader study at withheld intersections, the sweep of classifier training configurations and synthetic classifier audits. The generator comparisons used RoentGen-v2, FairDiffusion and CompDiff. Figure 1 summarises the experiments.

For chest imaging, we used postero-anterior (PA) MIMIC-CXR views with complete age, sex and race metadata[14]. Following the RoentGen-v2 protocol based on patient folders[5], patient folder `p19` supplied the test split (7,039 images), and `p10`–`p18` supplied training and validation (62,094 and 1,300 images). Test patients were disjoint from training and validation patients. Counts differ from the published RoentGen-v2 cohort because we required PA views and complete metadata. Supplementary Figure 1a shows the intersectional counts. FairGenMed[7] supplied $6,000$, $1,000$ and

3,000 scanning laser ophthalmoscopy fundus images for training, validation and testing, respectively. CheXpert provided external evaluation for the chest sweep of classifier training configurations[15].

The audit of rare intersections used a separate partition of MIMIC-CXR. We formed four demographically stratified splits and retrained each generator on the 31,694-image training partition. A larger real reference split of 28,174 images was excluded from generator and classifier training, validation, checkpoint selection, and hyperparameter and model selection. It supplied the reference AUROC estimates for this experiment (Section 4.6).

The study used de-identified datasets under their access conditions: MIMIC-CXR through the PhysioNet data use agreement for credentialed access and FairGenMed under its licence for research use. No new patient data were collected and no patients were contacted. The radiologist readers participated as co-authors, assessed de-identified real and synthetic images, and provided no personal data about themselves.

4.2 CompDiff architecture and objective

CompDiff uses a Hierarchical Conditioner Network (HCN) to encode age, sex and race separately from clinical text (Figure 2). CLIP encodes the clinical prompt as $E_{\text{text}} \in \mathbb{R}^{B\times 77\times d_{\text{ctx}}}$, where B is the batch size and d_{ctx} is the cross-attention dimension. HCN supplies $T = 4$ tokens, one per attribute plus a composed token. Concatenating them gives the UNet context $E_{\text{combined}} = [E_{\text{text}}, c] \in \mathbb{R}^{B\times 81\times d_{\text{ctx}}}$. Demographic attributes enter through HCN and are removed from the text prompt.

Attribute encoding and composition. For $V = \{\text{age}, \text{sex}, \text{race}\}$ and $K = |V| = 3$, each attribute has an embedding $e_v \in \mathbb{R}^{d_{\text{node}}}$. Sex and race use lookup tables with 2 categories (male, female) and 4 categories (White, Black, Asian, Hispanic/Other), respectively. Missing demographic attributes were excluded from the cohorts. Each encoder also has a learned null embedding for partial conditioning at inference. Age in years, a, is encoded continuously:

$$\begin{aligned} e_{\text{age}} &= \text{MLP}_{\text{age}}(\gamma(a)), \\ \gamma(a)_k &= \Big[\cos(a\omega_k),\ \sin(a\omega_k)\Big], \end{aligned} \tag{1}$$

Here $\omega_k = \eta^{-k/(d_\gamma/2)}$ for $k = 0, \ldots, d_\gamma/2 - 1$, with $\eta = 10^4$. A dedicated MLP combines each unordered attribute pair,

$$h_{v,w} = f_{v,w}([e_v, e_w]) \in \mathbb{R}^{d_{\text{node}}}, \tag{2}$$

where brackets denote concatenation. For three attributes, the pairs are age–sex, age–race and sex–race. Their representations are fused by a further MLP:

$$\begin{aligned} h_{\text{demographic}} &= g([\, h_{v,w} : \{v, w\} \subseteq V,\ v < w \,]) \\ &\in \mathbb{R}^{d_{\text{node}}}, \end{aligned} \tag{3}$$

The input dimension of g is $\binom{K}{2} d_{\text{node}}$ and its output dimension is d_{node}. The composed state parameterises a diagonal Gaussian through $(\mu, \log\sigma)$; we sample z by reparameterisation during training and use μ at inference. For each attribute, a dedicated MLP combines its embedding with the composed state,

$$u_v = q_v([\, e_v,\ h_{\text{demographic}} \,]) \in \mathbb{R}^{d_{\text{node}}}, \qquad v \in V, \tag{4}$$

and separate projections produce the context tokens:

$$\begin{aligned} c = [\, &\text{proj}_{\text{age}}(u_{\text{age}}),\ \text{proj}_{\text{sex}}(u_{\text{sex}}), \\ &\text{proj}_{\text{race}}(u_{\text{race}}),\ \text{proj}_{\text{cls}}(z) \,], \end{aligned} \tag{5}$$

Each projection is LayerNorm followed by a linear map. The variational latent appears only in the composed token; the attribute tokens are deterministic functions of u_v.

We used $d_{\text{node}} = 256$ and $d_{\text{ctx}} = 1024$. The age encoder has $d_\gamma = 128$ sinusoidal features over $[0, a_{\max}]$, where $a_{\max} = 100$ years, followed by $\text{Linear}(d_{\text{node}}) \to \text{SiLU} \to \text{Linear}(d_{\text{node}})$. The pairwise, composition and contextualisation MLPs use $\text{LayerNorm} \to \text{Linear}(2d_{\text{node}}) \to \text{SiLU} \to \text{Dropout}(0.1) \to \text{Linear}(d_{\text{node}})$. The variational heads use LayerNorm and a linear map, with $\log \sigma$ clamped to $[-5, 1]$.

Training objective. The joint training objective is

$$\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{comp}}\mathcal{L}_{\text{comp}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}. \tag{6}$$

The diffusion term is the standard loss for predicting noise[24], $\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0,\epsilon,t}\|\epsilon - \epsilon_\theta(x_t, t, E_{\text{combined}})\|_2^2$, over clean latent x_0, Gaussian noise ϵ and diffusion timestep t. The variational latent is regularised towards a standard normal:

$$\mathcal{L}_{\text{KL}} = \mathbb{E}\big[\,\text{KL}(\mathcal{N}(\mu, \sigma^2 I)\,\|\,\mathcal{N}(0, I))\,\big], \tag{7}$$

Compositional consistency provides an additive reference through $\mathcal{L}_{\text{comp}} = 1 - \cos(h_{\text{demographic}}, e_{\text{age}} + e_{\text{sex}} + e_{\text{race}})$. Auxiliary heads supervise the projected tokens consumed by the UNet:

$$\begin{aligned}\mathcal{L}_{\text{aux}} = \frac{1}{3 + \lambda_{\text{joint}}}\Big[\ &\text{CE}(\hat{y}_{\text{sex}}, y_{\text{sex}}) \\ &+ \text{CE}(\hat{y}_{\text{race}}, y_{\text{race}}) \\ &+ \kappa\,\text{SL}_1(\hat{a},\ a/a_{\max}) \\ &+ \lambda_{\text{joint}}\,\text{CE}(\hat{y}_{\text{joint}}, y_{\text{joint}})\ \Big],\end{aligned} \tag{8}$$

Sex and race use cross-entropy (CE) on their respective tokens. Age uses smooth-L_1 regression on $a/a_{\max} \in [0, 1]$, rescaled by κ. The joint head supervises the composed token on the age × sex × race label. Adult age bins are 18–40, 40–60, 60–80 and 80+, giving 32 adult intersections. The implementation includes an additional 0–18 bin and 40 joint outputs; the eight outputs for this extra bin receive no training signal because all included patients are adults.

Auxiliary heads use $\text{LayerNorm} \to \text{Linear}(512) \to \text{SiLU} \to \text{Dropout} \to \text{Linear}(C)$, with C the number of classes or 1 for age regression. We set $\lambda_{\text{KL}} = 0.005$, annealed linearly over 5,000 steps, $\lambda_{\text{comp}} = 0.1$, $\lambda_{\text{aux}} = 1.0$, $\lambda_{\text{joint}} = 0.5$ and $\kappa = 10$. HCN adds 6.32M parameters to the 1.21B trainable backbone parameters (0.52%). The auxiliary heads account for 2.13M and are discarded at inference, leaving a 0.35% parameter increase.

4.3 Training, sampling and checkpoint selection

All three generators used Stable Diffusion 2.1[11]. RoentGen-v2 encoded demographics in the clinical text prompt[5]; FairDiffusion added adaptive subgroup loss weighting through Fair Bayesian Perturbation[7]; CompDiff used HCN. For the standard chest X-ray and fundus comparisons, each generator was trained independently with three random seeds per modality. The figure label *Baseline SD* denotes RoentGen-v2.

Chest prompts followed `<AGE> year old <RACE> <SEX>. <IMPRESSION>`. Fundus prompts combined demographics with glaucoma status, vision (normal or mild, moderate or severe loss), cup-disc

ratio (normal, borderline or abnormal), and refraction (hyperopia, emmetropia or myopia). CompDiff removed demographics from the text and passed them to HCN; both baselines retained them in text.

The UNet and CLIP text encoder were unfrozen. Training used 512×512 images, bf16 mixed precision, gradient checkpointing and an exponential moving average of the weights. AdamW used a learning rate of 1×10^{-5} and a cosine schedule with 500 warm-up steps. Standard chest X-ray experiments used a batch size per GPU of 16 across six GPUs, giving an effective batch of 96 without gradient accumulation. Their 30,000-step budget corresponds to 2.88M sample presentations or approximately 46 epochs on the 62,094-image MIMIC-CXR training split.

For these chest experiments and the conditioning ablations, we evaluated 24 validation checkpoints from steps 1,000 to 30,000. Selection minimised validation FID subject to thresholds for demographic control:

$$\begin{aligned} t^\star &= \arg\min_{t \in \mathcal{A}} \ \mathrm{FID}_{\mathrm{val}}(t), \\ \mathcal{A} &= \{\, t \ : \ \mathrm{Acc}_{\mathrm{sex}}(t) \geq 0.98, \ \ \mathrm{Acc}_{\mathrm{race}}(t) \geq 0.92, \\ &\qquad \mathrm{RMSE}_{\mathrm{age}}(t) \leq 10 \text{ years} \,\}, \end{aligned} \tag{9}$$

The eligibility thresholds require sex and race classification accuracy of at least 0.98 and 0.92, respectively, and age RMSE no greater than 10 years. They prevent selecting a checkpoint with low FID that does not preserve the requested demographics. Full trajectories are in Supplementary Tables 8–10.

The zero-shot comparison across repeated training runs used three runs per method and holdout. All evaluated checkpoints were selected within the first 20,000 training steps. Checkpoint selection used mean validation FID-RadImageNet across the withheld cells at scheduled checkpoints of at least 5,000 steps. Supplementary Table 2 records the selected checkpoints and selection details.

Sampling used 75 denoising steps and classifier-free guidance scale 7.5 at 512×512 resolution. Corpora for classifier training matched the natural demographic distribution at $1\times$, $2\times$ or $3\times$ the MIMIC-CXR training size (62,094, 124,188 and 186,282 images), or used a separate corpus designed to balance demographic intersections. The latter expanded clinical impressions across all sex and race combinations, with age reweighting to approximately equalise the four age bins. The generated corpora contained 414 980 RoentGen-v2 images and 415 072 CompDiff images. Audit cohorts matched either the standard test split (7,039 images) or the separate real reference split (28,174 images), as described in Section 4.6.

4.4 Generation quality and unseen-intersection evaluation

Table 6 summarises the generation metrics and their directions. Distributional fidelity was assessed with Fréchet Inception Distance (FID)[25] and its RadImageNet counterpart[16], using Inception-v3 and RadImageNet ResNet-50 features, respectively. Lower distances indicate closer feature distributions. Inception FID was the primary fundus measure because RadImageNet was pretrained on radiological images. We also measured paired real–synthetic structural similarity with MS-SSIM[26], interpreted against the 0.25–0.75 range used by RoentGen-v2[5], and chest image semantic similarity with BioViL cosine similarity[27].

For chest distributional coverage, we extracted RadImageNet features from every available image in each cohort and fitted one joint UMAP to the pooled features. Real images form the common background in Figure 4. Quantitative Precision and Recall[17] and Density and Coverage[18] were computed in the original feature space. Precision measures how many synthetic images lie within

Table 6: Summary of all metrics of generation quality used in this paper, grouped by measurement role; procedures are given in Section 4.4. An upward arrow in the direction column indicates that higher values are better and a downward arrow that lower values are better. For MS-SSIM neither tail is desirable and we report whether the value falls inside the reference range 0.25–0.75[5]. Family (E) (ES-FID) is the extension for subgroup fairness used in Section 2.2.

Family	Metric	Direction	Property / axis	What it measures
(A) Similarity between real and synthetic images	FID[25]	↓	Fidelity (population)	Distributional distance, Inception v3 embeddings (backbone trained on natural images)
	FID-RadImageNet[16]	↓	Fidelity (population)	Distributional distance, RadImageNet ResNet-50 embeddings (radiology backbone)
	MS-SSIM[5,26]	in $[0.25, 0.75]$	Fidelity (per image)	Multiscale structural similarity between paired real and synthetic images
	BioViL cosine sim.[27]	↑	Fidelity (per image)	Semantic alignment of individual images in a chest X-ray image–language space
(B) Manifold / precision-recall coverage	UMAP visualisation	qualitative	Fidelity + Diversity	2-D projection of joint real/synth RadImageNet features for visual coverage
	Precision[17]	↑	Fidelity	Fraction of synth samples inside the real k-NN manifold (quality of individual images)
	Density[18]	↑	Fidelity	Refinement less sensitive to outliers of Precision
	Recall[17]	↑	Diversity	Fraction of real samples inside the synth k-NN manifold (unstable; reported only for completeness)
	Coverage[18]	↑	Diversity	Replacement less sensitive to outliers for Recall (uses manifold radius measured on real images)
(C) Generalisation	Authenticity[19]	↑	Neighbour similarity	Fraction of synthetic samples passing the distance check against real neighbours; not proof of non-memorisation
(D) Conditioning alignment	Disease AUROC[28]	↑	Cond. fidelity (clinical)	Agreement with prompted findings: XRV disease AUROC for chest; glaucoma and cup-disc classifiers for fundus
	Sex accuracy[2]	↑	Cond. fidelity (demographic)	Chest X-ray sex classifier accuracy on synth
	Race accuracy[28]	↑	Cond. fidelity (demographic)	XRV race classifier accuracy on synth
	Age RMSE[28]	↓	Cond. fidelity (demographic)	XRV age regressor RMSE (years) on synth
(E) Fairness of generation quality	ES-FID[29,30]	↓	Subgroup equity	Overall FID plus the mean absolute deviation of subgroup FIDs from that overall value

the real nearest-neighbour manifold; Recall reverses those roles. Density and Coverage provide alternatives less sensitive to synthetic outliers. Coverage uses neighbourhood radii measured on real images. We used $k = 10$ neighbours for cohorts with the natural demographic distribution ($n \approx 7{,}000$) and $k = 20$ for withheld cells ($n = 43$–663). Authenticity[19] was the fraction of synthetic images whose distance to their nearest real reference image was at least that real image's nearest-neighbour distance within the real reference cohort. We treated this as a nearest-neighbour diagnostic, rather than proof of non-memorisation.

Conditioning alignment and subgroup fidelity. Pretrained classifiers assessed whether generated images expressed the conditioning labels. Chest disease alignment used TorchXRayVision's XRV DenseNet-121[28], with prompted pathology labels as targets for AUROC over 5 diseases: atelectasis, cardiomegaly, oedema, pneumothorax and pleural effusion. Fundus alignment used pretrained EfficientNet classifiers for glaucoma status and cup-disc ratio categories. Chest sex accuracy used the dedicated classifier of Glocker et al.[2]; race accuracy and age RMSE used the XRV race classifier and age regressor. Generated 512×512 images were downsampled by area interpolation to the 224×224 inputs expected by XRV. Demographic targets were the attributes supplied to each generator, through text for the baselines and HCN for CompDiff. Race accuracy measures agreement with labels associated with the dataset as recovered by the selected classifier; it is not a measure of biological identity.

We computed fidelity separately by sex, race, age and their joint intersection. Equity-scaled FID[7,29,30] combines overall FID with the mean absolute deviation of subgroup distances from that overall value:

$$\begin{aligned}\text{ES-FID}_{\mathcal{A}^i} = \text{FID} \cdot \Bigg(1 + \frac{1}{|\mathcal{A}^i| \cdot \text{FID}} \\ \times \sum_{j=1}^{|\mathcal{A}^i|} |\text{FID} - \text{FID}_{\mathcal{A}^i_j}| \Bigg)\end{aligned} \tag{10}$$

Here $\mathcal{A}^i$ is the set of subgroups for attribute i, or the full intersectional subgroup list. The measure equals overall FID when every subgroup distance equals that overall value. Chest subgroup comparisons used RadImageNet features and fundus comparisons used Inception features.

Unseen intersections. We removed eight demographic intersections from each of two MIMIC-CXR training variants, reallocating the affected patients 80% to test and 20% to validation (Supplementary Figure 1). Holdout A removed the eight rarest intersections, with natural training counts below 380: 18–40 Asian women and men, and 80+ Asian, Black and Hispanic women and men. Holdout B removed the eight next rarest intersections, with counts 381–782: 40–60 and 60–80 Asian women and men, and 18–40 and 60–80 Hispanic women and men.

These disjoint sets of target intersections had no training patients after reallocation; other intersections retained their natural counts. Training/validation/test sizes changed from 62,094/1,300/7,039 to 60,830/1,564/8,039 for Holdout A and 57,737/2,199/10,497 for Holdout B. Each method was trained independently three times per holdout and used to generate the withheld intersections. FID-RadImageNet within each cell compared each synthetic cohort with real test images from the same intersection. The average across these cell distances gives equal weight to each intersection and differs from FID computed on pooled images.

4.5 Radiologist reader study

Two board-certified radiologists assessed images from the 16 withheld chest intersections. Each set contained three candidate radiographs for the same clinical impression and intersection, produced by one selected model per method on the relevant holdout split. The readers completed 96 set assessments over 92 distinct sets, with 4 sets assessed by both readers, giving 288 image ratings.

Images appeared in randomised order. Readers were blinded to source and to the presence or proportion of real images. A hidden real image from the same intersection replaced one generator in 12 assessments: RoentGen-v2 in 6, CompDiff in 5 and FairDiffusion in 1. Consequently, CompDiff, FairDiffusion and RoentGen-v2 appeared in 91, 95 and 90 assessments, respectively.

Readers scored anatomical realism from 1 (clearly artefactual) to 5 (indistinguishable from a real scan), and correspondence with the clinical impression from -2 (complete mismatch) to $+2$ (perfect correspondence), following Chambon et al.[3]. They also selected the most realistic image in each set. Readers could flag an image as unreadable or mark correspondence as not assessable, including when impressions described interval changes that could not be judged from a single image. One unreadable rating was excluded from both outcomes and 10 further ratings from correspondence, leaving 287 realism and 277 correspondence ratings. Forced-choice rates were calculated among sets containing the source. Supplementary Figure 7 shows the viewer.

Statistical models, source comparisons and sensitivity analyses are specified in Section 4.9.

4.6 Downstream training and classifier auditing

Classifier training on synthetic data. The chest training sweep compared CompDiff with RoentGen-v2 under matched classifier recipes (Table 4). Training on synthetic data alone used no real chest images. In the second family, classifiers were pretrained on synthetic images and fine-tuned on real MIMIC-CXR data. The four synthetic corpora described in Section 4.3 were evaluated with random or ImageNet initialisation and fractions of real data of 1%, 10%, 30%, 50%, 70% and 100%, giving 48 pretraining/fine-tuning configurations per evaluation dataset. Recipes using synthetic data alone used ImageNet initialisation.

For each of the four fixed corpora, we also evaluated four independent classifier runs per generator, giving 32 models evaluated on both test sets. Additional experiments repeated pretraining followed by fine-tuning on all real training images and training on mixed real and synthetic images. The repeated-training experiments reported here used 256×256 training inputs, batch size 128, an initial learning rate of 0.001 and a 50-epoch limit, with AdamW (Supplementary Table 11). The four runs shared these settings; their saved test evaluations used the common 224×224 preprocessing.

Classifiers were DenseNet-121 models[31]. The original fine-tuning sweep used AdamW with a cosine schedule, 224×224 classifier inputs, batch size 32 and a maximum of 100 epochs. The generated images had 512×512 resolution and were resized for classifier input. Batch size, centre cropping, random horizontal flips, small rotations, weight decay and early stopping were held constant between generators within each recipe. Evaluation used the real MIMIC-CXR test split and the external CheXpert test set.

AUROC_8 is the macro-average over a fixed set of eight diagnostic labels: atelectasis, cardiomegaly, consolidation, oedema, pleural effusion, pneumonia, pneumothorax and no finding. We recomputed all repeated-training and subgroup scores using this same label set and the saved prediction column

order. For the exploratory sensitivity analysis, each test ROC supplied the maximum true positive rate at a false positive rate no greater than 0.10 or 0.05. Thresholds were read separately from each classifier's test curve, rather than validated as operating thresholds. AUROC_8 was used for the downstream sweep; AUROC_{14} in the MIMIC audits averages all 14 available pathology labels.

We reported subgroup AUROCs, equity-scaled AUROC (ES-AUC), equalised-odds difference and underdiagnosis. ES-AUC discounts overall AUROC by variation between groups: the implemented score divides overall AUROC by one plus the mean absolute deviation, weighted by group size, of subgroup AUROCs from the overall value. Equalised-odds difference is the larger of the gaps in true positive and false positive rates between groups, computed with `fairlearn`[32]. Underdiagnosis is the false positive rate for *No Finding*: a diseased image classified as having no finding[1]. Subgroup AUROC_8 required both positive and negative images for every label; all ten sex, race and age groups met this rule on each test set. We retained the same eight labels in every group rather than changing the average when a label was sparse. Supplementary Table 14 reports the smallest positive count and the sensitivity check requiring at least ten positive images for every label. For the analysis of the subgroup with the lowest baseline performance, each matched configuration identifies the subgroup with the lowest RoentGen-v2 AUROC within sex, race or age; its identity is held fixed when computing the CompDiff gain.

Agreement with audits using real data. Each audit held a single DenseNet-121 trained on real MIMIC-CXR fixed during evaluation. Substitution of entire cohorts compared the real test split with a synthetic cohort from each generator, matched in size and demographic distribution. For each overall or subgroup metric, the audit error was its absolute difference from the corresponding value on the real test set. Outcomes included AUROC, equity-scaled AUROC, equalised-odds difference and underdiagnosis.

Estimation on rare intersections. For rare cells, the larger independent real reference split supplied AUROC estimates for each cell. The small real test split and a synthetic cohort from each generator were evaluated as estimators of these values, with each synthetic cohort matched to the reference split's size and demographic distribution. The small test split contained 6–18 images in its most sparsely populated intersections. Absolute cell estimation errors were averaged within three tiers defined by reference cohort size: the rarest 9, middle 9 and most abundant 9 cells. The reference split was excluded from training and model selection as described in Section 4.1.

4.7 Conditioning and composer ablations

Conditioning and loss controls are specified in Table 5, including the comparable reference configuration for each group of variants. Validation metrics were reported at the checkpoint selected by the shared rule based on thresholds for demographic control. If no checkpoint met the thresholds, the table reports the checkpoint with the lowest age RMSE.

The unstructured composer retained separate attribute encoders, separate output tokens, supervision of the output tokens and the same losses, replacing pairwise composition with MLPs matched in depth and parameter count that receive the concatenated attribute embeddings. Composer sizes were 2,896,640 and 2,898,632, differing by 0.069%. The unstructured output is split into attribute tokens whose association with each attribute is established by supervision. Both composers were also retrained on each holdout split. The factorised checkpoints were the CompDiff seed-42 models

in Supplementary Table 2; the unstructured control used the checkpoint with the lowest validation FID on Holdout A, also matching the factorised arm in training step, and the checkpoint with the lowest validation FID-RadImageNet on Holdout B. Further comparisons used a common step with three generation seeds or each arm's best validation checkpoint under either criterion. Steps to the thresholds for demographic control were recorded separately. Supporting composer comparisons are reported in Supplementary Note 4.

Supplementary analysis of demographic text. For the supporting CLIP analysis, we tokenised all 62,094 training prompts (32,319 unique clinical impressions), with and without demographics, to measure truncation at CLIP's 77-token limit. A separate probe of the frozen encoder used 36,562 clinical impressions and 24 demographic prefixes: young, middle-aged or elderly, combined with the two sex and four race labels in the form `<AGE> <RACE> <SEX> patient.`. For each impression, cosine perturbation relative to the embedding without demographic text was averaged over prefixes. These combinations using descriptive age terms differ from the 32 age-bin intersections. Linear probes measured recoverability of demographic labels from embeddings without demographic text. Clinical density was the fraction of words recognised as clinical entities. The visualisation used LOWESS with span 0.20, a power-law fit in log–log space, and means within quartiles of word count with 95% bootstrap intervals from 500 resamples. Results are reported in Supplementary Note 9.

4.8 Fidelity robustness and sampling controls

Controls for the fidelity estimator and sampling budget used one selected model per configuration and holdout, including the factorised and unstructured composers. These controls were separate from the comparison across repeated training runs. Fidelity robustness analyses rescored their images using KID-RadImageNet[20] and corrections for sample size. Holding generator, intersection and image pool fixed, we subsampled equal numbers of synthetic and real images from large cells and measured change relative to $s = 600$. Following Chong and Forsyth[33], FID was fitted against inverse candidate size $1/k$ and jointly against candidate and reference sizes $(1/k, 1/m)$. We compared local slopes with those fitted over a wide range of sample sizes, validated extrapolation from 150 images against withheld targets from large cells, and normalised cell scores to a common sample size using the fitted curves describing the dependence on sample size. Ranking and agreement on the winning configuration were assessed relative to the uncorrected FID-RadImageNet ordering of those configurations, with Spearman correlations within each cell. Corrected values were sensitivity analyses; primary values remained uncorrected. Withheld and well represented cells were compared within split and generator.

For post hoc selection, a verifier combined XRV demographic agreement with a nearest-neighbour realism term. Its real reference half was disjoint from the half used to evaluate selected images. Selecting the best quarter gave a 4× sampling budget. We compared selected and random samples using bootstrap intervals, with Mann–Whitney tests between withheld and well represented groups. Comparisons included 2× and 4× budgets and an oracle selector with access to evaluation references. Selection varied prompt and sampling noise jointly. Cells below 120 images were excluded, including 4 Holdout A cells. Supporting results are in Supplementary Notes 2 and 3.

4.9 Statistical analysis and reporting

Difference tests were two-sided with a nominal significance level of $\alpha = 0.05$; the equivalence analysis used two one-sided tests at the same level. Holm adjustment controlled family-wise error across the two primary zero-shot FID-RadImageNet comparisons. The reader study controlled the false discovery rate at $q = 0.05$ across its eight source contrasts using Benjamini–Hochberg adjustment. Other comparisons were exploratory, with adjustment for tests within demographic cells as specified below. Confidence intervals were pointwise 95% intervals without multiplicity adjustment. We report p-values with their comparisons and identify adjusted values explicitly. The reader and downstream reanalyses and zero-shot sensitivity analyses were specified after inspection of the existing results.

Replication and comparisons across trained models. Generation summaries are means and standard deviations across three independently trained models per method and modality, with fundus generation repeats first averaged within each model. Zero-shot comparisons used three models per method and holdout; the eight cell scores from a model were repeated outcomes. Classifier comparisons used four independent runs per generator and fixed corpus, giving 32 classifiers evaluated on both datasets. The datasets did not provide additional independent training replicates. Equal seed or run numbers were not treated as matched pairs.

For both experiments, we compared CompDiff with comparator b within each stratum g and weighted strata equally: $G = 2$ holdouts for zero-shot generation and $G = 4$ corpora for classifier training. Writing $\bar{Y}_{ga}$, s^2_{ga} and n_{ga} for the mean, variance and number of independent runs for method a in stratum g,

$$\begin{aligned} \widehat{\Delta}_b &= \frac{1}{G}\sum_{g=1}^{G}(\bar{Y}_{g,\mathrm{CD}} - \bar{Y}_{g,b}), \\ V_b &= \sum_{g,a} q_{ga}, \qquad q_{ga} = \frac{s^2_{ga}}{G^2 n_{ga}}, \\ \nu_b &= \frac{V_b^2}{\sum_{g,a} q_{ga}^2/(n_{ga}-1)}, \qquad a \in \{\mathrm{CD}, b\}. \end{aligned} \tag{11}$$

Welch–Satterthwaite intervals were $\widehat{\Delta}_b \pm t_{0.975,\nu_b}\sqrt{V_b}$; zero-shot tests used $\widehat{\Delta}_b/\sqrt{V_b}$ with ν_b degrees of freedom. This approach allows different variances between methods and strata. Each zero-shot contrast used twelve trained models; classifier contrasts used all 32 classifiers. Inference conditions on the fixed holdouts, generated corpora and real test cohorts. It assumes independent, approximately normal run means; three or four runs per group do not permit a reliable assessment of normality.

Zero-shot sensitivity checks used a model with common variance; fixed effects for each combination of cell and method with CR1 covariance clustered by trained model and a t distribution with degrees of freedom equal to the number of clusters minus one; and a percentile bootstrap resampling complete runs within method and holdout, enumerating 531,441 resamples. Permuting method labels for complete runs within holdouts enumerated 400 assignments per comparison. This requires exchangeability, a stronger assumption than equality of means with unequal variances. Additional checks omitted each model in turn, restricted CompDiff or all methods to seeds 43 and 44, and repeated scoring with Inception features. Classifier sensitivity checks omitted each run in turn and included all available models descriptively. These checks examine sensitivity to modelling choices; they do not establish normality or replace the primary analysis.

Reader ratings and preferences. Ratings were analysed separately for realism and correspondence with the mixed-effects model

$$y_{isr} = \mu + \beta_s + \gamma_r + u_i + \varepsilon_{isr}, \tag{12}$$

where y_{isr} is the score from reader r for source s in set i, β_s and γ_r are fixed source and reader effects, $u_i \sim \mathcal{N}(0, \sigma^2_{\text{set}})$ is a random set intercept, and $\varepsilon_{isr} \sim \mathcal{N}(0, \sigma^2)$ is the residual. Models were fitted by restricted maximum likelihood. Source contrasts were estimated within sets. The two readers were treated as fixed because the panel was too small to estimate a variance across the reader population reliably.

The realism model included 287 ratings from 92 image sets; the correspondence model included 277 ratings from 89 sets. The model accounts for shared image sets and reader calibration. Its Gaussian random effects and residuals are working assumptions for the bounded rating scales, rather than empirically established normal distributions. We assessed sensitivity using sign-flip permutation tests on differences within reader and set (10,000 draws), Wilcoxon signed-rank tests, ordinary least squares with standard errors clustered by image set, and a model treating reader as a variance component. The paired checks assume symmetry of the differences under the null; their sample sizes are reported with the contrasts in Supplementary Table 7.

The contrast family comprised all three generator pairs for each rating outcome and CompDiff against each baseline for forced choice. Rating comparisons used Wald tests and intervals. Forced-choice comparisons used exact binomial tests against equal preference among assessments containing both sources, conditional on one of them being selected: $n = 50$ for CompDiff versus RoentGen-v2 and $n = 79$ for CompDiff versus FairDiffusion. Preferences for a third source did not enter these tests. Realism equivalence between CompDiff and FairDiffusion used two one-sided t-tests on 89 paired assessments with a ± 0.5 margin, chosen as a conventional half-point difference rather than a clinically established threshold. All eight source contrasts and the sensitivity results are reported in Supplementary Note 5.

Configuration summaries and patient resampling. The 48 fine-tuning configurations shared training data, pretraining weights and test cohorts. Label means, subgroup medians, interquartile ranges and counts of positive differences describe this fixed sweep. For the three MIMIC-CXR groups with fewer than ten positive images for at least one label, 1,000 bootstrap draws sampled patients with replacement within each group, retaining all images and model predictions for each sampled patient. Each draw recomputed the median of the 48 matched configuration differences. Draws lacking either class for any of the eight labels were undefined. Supplementary Table 15 reports patient counts, the number of valid draws and their 2.5th–97.5th percentile range. These ranges describe sensitivity to patient sampling conditional on valid label coverage and fixed models; they do not establish nominal confidence coverage in these sparse groups.

Other comparisons. Welch's t-tests compared race accuracy and age error across three training runs per method. Tests within individual demographic cells required at least two available runs per method and used Welch's test with Benjamini–Hochberg adjustment at $q = 0.05$ across cells, separately for each comparator. Standard-distribution intersectional fidelity comparisons paired method means within the sixteen cells with at least twenty images, using paired t-tests and Wilcoxon signed-rank checks. Pairing controls for cell identity; these cells share trained generators and are not independent training replicates. Approximate normality concerns the paired differences for the t-test,

and symmetry concerns those differences for the signed-rank check. Composer comparisons used one training run per composer and regime and were interpreted against the final configuration's variation across training runs. Sampling-control comparisons used bootstrap intervals and Mann–Whitney tests as described in Section 4.8. Audit summaries describe a single classifier; rare-cell AUROC intervals used 1,000 bootstrap iterations at 95% confidence, with estimation errors and interval widths summarised by rarity tier. Tier means do not quantify variation across training seeds.

AI tools assisted literature checking, manuscript restructuring, language revision and statistical reanalysis of existing experimental outputs. No additional training or image generation was performed for the reanalysis. The authors retain responsibility for the scientific content and its verification.

Data availability

The source datasets are distributed by their respective custodians under their access and use conditions. MIMIC-CXR[14] is available to credentialed users through PhysioNet (`https://physionet.org/content/mimic-cxr/2.1.0/`) after completion of the required training and signature of its data use agreement. FairGenMed[7] is distributed by its authors for non-commercial research use under a CC BY-NC-ND 4.0 licence (`https://huggingface.co/datasets/harvardairobotics/FairGenMed`). CheXpert[15], used for external classifier evaluation, is described by Stanford at `https://stanfordmlgroup.github.io/competitions/chexpert/`, which provides the dataset access route. The synthetic training corpora used in the classifier experiments, a natural corpus of 62,094 chest X-ray images and a demographically balanced corpus of 415 072 images, each with its generation prompt, demographic conditioning and finding labels, are deposited on Zenodo under a CC BY 4.0 licence, the natural corpus at `https://doi.org/10.5281/zenodo.23688264` and the balanced corpus at `https://doi.org/10.5281/zenodo.23688328`, and are mirrored on the Hugging Face Hub at 🤗`https://huggingface.co/datasets/mahmoudibra98/compdiff-chest-xray-synthetic`. Synthetic fundus images are not deposited because the CC BY-NC-ND 4.0 licence of FairGenMed does not allow sharing derived material; they, and any other synthetic cohort, can be regenerated from the released weights. The de-identified per image reader study ratings are available from the corresponding author on reasonable request.

Code availability

The CompDiff implementation is archived on Zenodo as the tagged release v2.2-npjdm at `https://doi.org/10.5281/zenodo.23339597` and maintained at `https://github.com/mahmoudibrahim98/CompDiff`. Pretrained model weights are released on the Hugging Face Hub: the chest X-ray model at 🤗`https://huggingface.co/mahmoudibra98/compdiff-chest-xray` and the fundus model at 🤗`https://huggingface.co/mahmoudibra98/compdiff-fundus`.

Acknowledgements

This work was funded by the European Union under the Horizon Europe grant 101095435. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions

Mahmoud K. Ibrahim designed the study, developed the model, ran all experiments, produced the figures, and wrote the manuscript. C.S., M.D., B.E. and G.E. provided supervision and reviewed the manuscript. Maryam K. Ibrahim helped design and manage the radiologist reader study, which A.J. and M.M.S. conducted.

Competing interests

The authors declare no competing interests.

References

[1] Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In: *Proc. Mach. Learn. Res.* **149,** 232–243 (2021).

[2] Glocker, B., Jones, C., Bernhardt, M. & Winzeck, S. Algorithmic encoding of protected characteristics in chest X-ray disease detection models. *eBioMedicine* **89,** 104467 (2023).

[3] Chambon, P., Bluethgen, C., Langlotz, C. P. & Chaudhari, A. RoentGen: Vision-language foundation model for chest X-ray generation. arXiv:2211.12737 (2022).

[4] Bluethgen, C. et al. A vision–language foundation model for the generation of realistic chest X-ray images. *Nat. Biomed. Eng.* **9,** 494–506 (2025).

[5] Moroianu, S. L. et al. Improving performance, robustness, and fairness of radiographic AI models with finely-controllable synthetic data. arXiv:2508.16783 (2025).

[6] Ktena, I. et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.* **30,** 1166–1173 (2024).

[7] Luo, Y. et al. FairDiffusion: Enhancing equity in latent diffusion models via fair Bayesian perturbation. *Sci. Adv.* **11,** eads4593 (2025). doi:10.1126/sciadv.ads4593.

[8] Li, Z. et al. FairGen: preference-aligned diffusion for demographically equitable medical image synthesis. *npj Digit. Med.* (2026). doi:10.1038/s41746-026-02868-z.

[9] Cooke, L. H. et al. RoentMod: a synthetic chest X-ray modification model to identify and correct image interpretation model shortcuts. *npj Digit. Med.* **9,** 324 (2026). doi:10.1038/s41746-026-02497-6.

[10] Ji, Y. et al. A Generative Foundation Model for Chest Radiography. *NEJM AI* **3,** AIoa2500799 (2026). doi:10.1056/AIoa2500799.

[11] 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).

[12] Liu, N., Li, S., Du, Y., Torralba, A. & Tenenbaum, J. B. Compositional visual generation with composable diffusion models. In: *European Conference on Computer Vision (ECCV)* (2022).

[13] Radford, A. et al. Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning (ICML)*, pp. 8748–8763 (2021).

[14] Johnson, A. E. W. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6,** 317 (2019).

[15] Irvin, J. et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* **33,** 590–597 (2019).

[16] Mei, S. et al. RadImageNet: An open radiologic deep learning research dataset. *Radiol. Artif. Intell.* **4,** e210315 (2022).

[17] Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J. & Aila, T. Improved precision and recall metric for assessing generative models. In: *Adv. Neural Inf. Process. Syst.* **32,** 3927–3936 (2019).

[18] Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y. & Yoo, J. Reliable fidelity and diversity metrics for generative models. In: *International Conference on Machine Learning (ICML)*, pp. 7176–7185 (2020).

[19] Alaa, A., van Breugel, B., Saveliev, E. & van der Schaar, M. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In: *International Conference on Machine Learning (ICML)*, pp. 290–306 (2022).

[20] Bińkowski, M., Sutherland, D. J., Arbel, M. & Gretton, A. Demystifying MMD GANs. In: *International Conference on Learning Representations (ICLR)* (2018). arXiv:1801.01401.

[21] Gaudi, S., Sreekumar, G. & Boddeti, V. CoInD: Enabling logical compositions in diffusion models. arXiv:2503.01145 (2025).

[22] Liang, Q., Liu, Z., Ostrow, M. & Fiete, I. How diffusion models learn to factorize and compose. In: *Adv. Neural Inf. Process. Syst.* **37** (2024).

[23] Maheshwari, G., Bellet, A., Denis, P. & Keller, M. Synthetic data generation for intersectional fairness by leveraging hierarchical group structure. arXiv:2405.14521 (2024).

[24] Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. In: *Adv. Neural Inf. Process. Syst.* **33,** 6840–6851 (2020).

[25] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: *Adv. Neural Inf. Process. Syst.* **30,** 6629–6640 (2017).

[26] Wang, Z., Simoncelli, E. P. & Bovik, A. C. Multi-scale structural similarity for image quality assessment. In: *Proc. Asilomar Conf. Signals, Syst. Comput.*, vol. 2, pp. 1398–1402 (2003).

[27] Boecking, B. et al. Making the most of text semantics to improve biomedical vision-language processing. In: *European Conference on Computer Vision (ECCV)* (2022).

[28] Cohen, J. et al. TorchXRayVision: A library of chest X-ray datasets and models. arXiv:2111.00595 (2021).

[29] Luo, Y., Tian, Y., Shi, M., Elze, T. & Wang, M. FairVision: equitable deep learning for eye disease screening via fair identity scaling. arXiv:2310.02492 (2023).

[30] Tian, Y. et al. FairSeg: A large-scale medical image segmentation dataset for fairness learning with fair error-bound scaling. In: *International Conference on Learning Representations (ICLR)* (2024).

[31] 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).

[32] Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. In: *Adv. Neural Inf. Process. Syst.* **29,** 3315–3323 (2016).

[33] Chong, M. J. & Forsyth, D. Effectively unbiased FID and Inception Score and where to find them. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020). doi:10.1109/CVPR42600.2020.00611.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.pdf