Supplementary Information

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

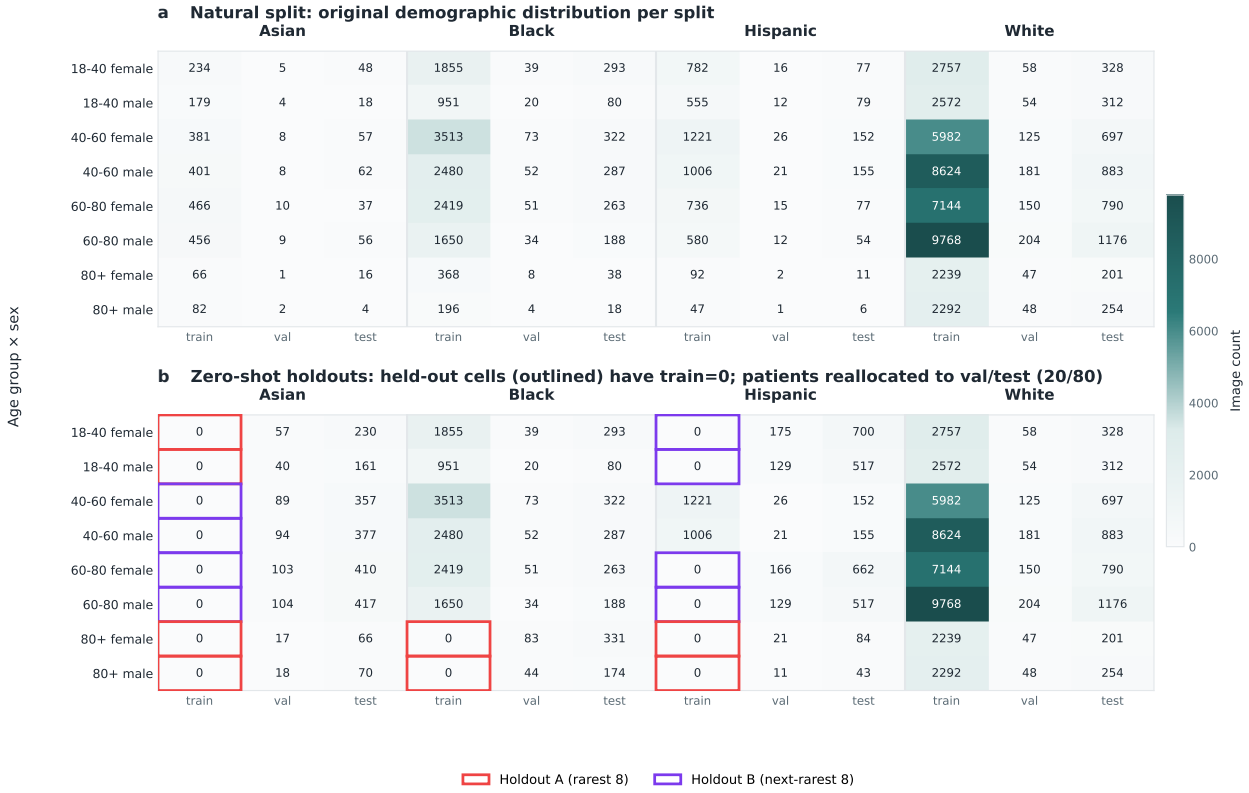
[5]Department of Radiology, Division of Abdominal Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA

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

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

Manuscript version 2.2

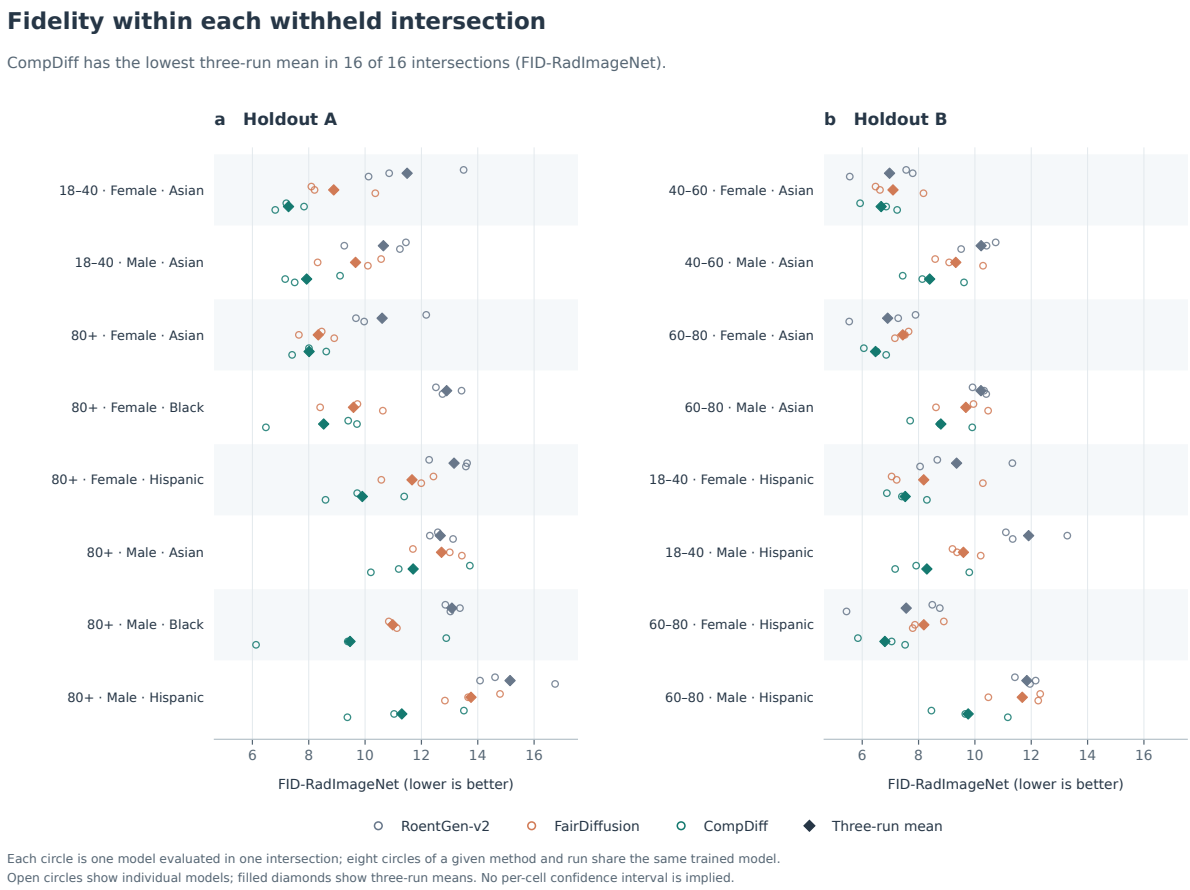
This document contains supporting results and illustrations. The main Methods describes the experimental design and the statistical procedures; Supplementary Notes 1, 5 and 8 report the supporting zero-shot, reader-study and downstream results, respectively, and Note 9 the prompt tokenisation and text-encoder probes. Figure and table numbers in this document form separate supplementary sequences; references to sections of the main paper and displays open the accompanying `main.pdf`.



Supplementary Figure 1: Two zero-shot withheld splits of the MIMIC-CXR distribution. Rows: age-bin × sex. Columns: race (outer) × split (inner). **a**, Natural distribution: image counts per age × sex × race × split intersection before any manipulation. **b**, Distribution after reallocation under the two zero-shot splits: each split removes eight disjoint three-way intersectional cells from train (count shown as 0) and reallocates the affected patients, 80% to test and 20% to validation, inflating the val/test counts for the withheld cells. Red outlines mark the eight cells of Holdout A (the rarest 8, natural training count below 380); violet outlines mark the eight cells of Holdout B (the next rarest 8, natural training count 381–782). The two splits are disjoint, so each cell falls into exactly one category.

1 Zero-shot results across repeated training runs and sensitivity analyses

The comparison across repeated training runs comprised 18 selected models: three methods, two holdouts and three runs per method and holdout. CompDiff had the lowest mean FID-RadImageNet across three runs in all 16 withheld intersections (Supplementary Figure 2). Supplementary Table 1 gives the cell means and standard deviations in both feature spaces.



Supplementary Figure 2: FID-RadImageNet within each withheld intersection across repeated training runs. Panels a and b show the eight target intersections in Holdouts A and B, respectively. Colours identify the generators. Open circles show individual trained models and filled diamonds show the mean across three models for that method and intersection. The eight cell observations contributed by one model share that fitted generator and are not independent training replicates. Lower distances indicate closer real–synthetic feature distributions. CompDiff has the lowest mean across three runs in all 16 cells; these descriptive cell means do not imply that every individual run wins. No cell confidence interval is shown. Numerical means and standard deviations are in Supplementary Table 1.

Supplementary Table 1: Fidelity at each withheld intersection across independently trained models.

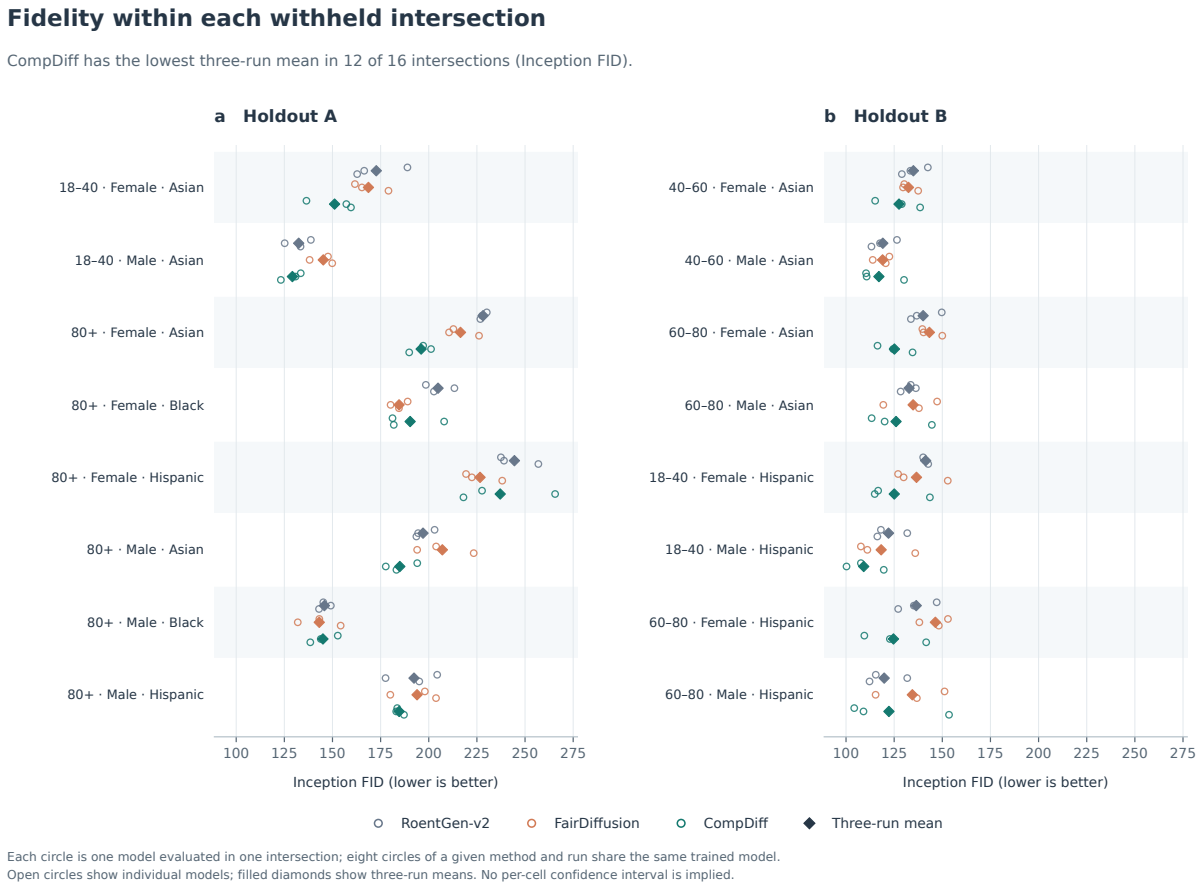
Holdout	Intersection	CompDiff	FairDiffusion	RoentGen-v2
FID-RadImageNet				
A	18–40 Female Asian	**7.28 ± 0.52**	8.89 ± 1.28	11.49 ± 1.77
A	18–40 Male Asian	**7.93 ± 1.04**	9.66 ± 1.19	10.65 ± 1.21
A	80+ Female Asian	**8.01 ± 0.61**	8.34 ± 0.64	10.60 ± 1.36
A	80+ Female Black	**8.53 ± 1.78**	9.59 ± 1.12	12.90 ± 0.47
A	80+ Female Hispanic	**9.90 ± 1.41**	11.67 ± 0.97	13.16 ± 0.77
A	80+ Male Asian	**11.71 ± 1.81**	12.71 ± 0.91	12.67 ± 0.42
A	80+ Male Black	**9.47 ± 3.38**	10.98 ± 0.14	13.08 ± 0.26
A	80+ Male Hispanic	**11.30 ± 2.08**	13.76 ± 0.98	15.15 ± 1.41
B	18–40 Female Hispanic	**7.53 ± 0.72**	8.18 ± 1.82	9.35 ± 1.74
B	18–40 Male Hispanic	**8.29 ± 1.36**	9.59 ± 0.54	11.91 ± 1.20
B	40–60 Female Asian	**6.67 ± 0.68**	7.09 ± 0.94	6.97 ± 1.23
B	40–60 Male Asian	**8.39 ± 1.11**	9.32 ± 0.87	10.22 ± 0.64
B	60–80 Female Asian	**6.47 ± 0.40**	7.44 ± 0.25	6.90 ± 1.22
B	60–80 Female Hispanic	**6.80 ± 0.86**	8.19 ± 0.62	7.56 ± 1.84
B	60–80 Male Asian	**8.79 ± 1.10**	9.68 ± 0.95	10.21 ± 0.26
B	60–80 Male Hispanic	**9.76 ± 1.36**	11.68 ± 1.04	11.84 ± 0.38
Inception FID				
A	18–40 Female Asian	**151.09 ± 12.71**	168.67 ± 9.19	172.74 ± 14.13
A	18–40 Male Asian	**129.18 ± 5.36**	145.22 ± 6.24	132.45 ± 6.84
A	80+ Female Asian	**196.04 ± 5.74**	216.48 ± 8.41	228.21 ± 1.72
A	80+ Female Black	190.33 ± 15.30	**184.58 ± 4.42**	204.83 ± 7.65
A	80+ Female Hispanic	237.10 ± 25.16	**226.64 ± 10.09**	244.48 ± 10.80
A	80+ Male Asian	**184.97 ± 8.31**	207.05 ± 14.94	197.02 ± 5.22
A	80+ Male Black	145.05 ± 7.20	**143.13 ± 11.14**	145.79 ± 3.11
A	80+ Male Hispanic	**184.62 ± 2.19**	193.92 ± 12.37	192.34 ± 13.61
B	18–40 Female Hispanic	**125.04 ± 16.06**	136.59 ± 14.14	141.35 ± 1.31
B	18–40 Male Hispanic	**109.19 ± 9.75**	118.25 ± 15.42	122.06 ± 8.49
B	40–60 Female Asian	**127.52 ± 11.74**	132.44 ± 4.37	134.97 ± 6.88
B	40–60 Male Asian	**117.08 ± 11.30**	118.99 ± 4.54	119.10 ± 6.74
B	60–80 Female Asian	**125.13 ± 9.12**	143.29 ± 5.81	140.04 ± 8.51
B	60–80 Female Hispanic	**124.63 ± 16.14**	146.40 ± 7.58	136.46 ± 10.06
B	60–80 Male Asian	**125.98 ± 16.46**	134.84 ± 14.21	132.74 ± 4.02
B	60–80 Male Hispanic	122.28 ± 27.14	134.43 ± 18.02	**119.82 ± 10.50**

Values are mean ± sample standard deviation across 3 trained models per method and holdout. Bold indicates the lowest mean within an intersection and feature space. Distances compare synthetic and real images within the same cell; lower values indicate closer feature distributions.

With Inception features, CompDiff had the lowest mean across three runs in 12 of 16 cells. Its average differences were -10.575 against RoentGen-v2, with a pointwise 95% interval of $[-24.672, +3.523]$, and -9.732 against FairDiffusion, with an interval of $[-23.686, +4.223]$. Both intervals included zero (Supplementary Figure 3); the respective unadjusted two-sided p-values were 0.1006 and 0.1350. The cell results in Supplementary Figure 4 show where the feature backbones differ in their rankings; changing the backbone preserved the direction of the average effects, but not every cell's ordering.



Supplementary Figure 3: Zero-shot comparison across repeated training runs using Inception-v3 features. This uses the same models and generated cohorts as Figure 6, changing only the feature backbone. **a,b**, Each open circle is one model's mean FID across its eight withheld intersections; filled diamonds show means across three independently trained models. **c,d**, CompDiff minus comparator differences averaged equally across cells and holdouts, with pointwise 95% Welch–Satterthwaite confidence intervals. Negative differences favour CompDiff. Each difference has its own zero reference, and both panels use the same scale. Both intervals for the average differences include zero. Results for individual cells appear in Supplementary Figure 4.



Supplementary Figure 4: Inception FID within each withheld intersection across repeated training runs. This figure uses the same generated cohorts and conventions as Supplementary Figure 2. Open circles show individual model scores and filled diamonds their means across three runs, separated by method within each cell. CompDiff has the lowest mean in 12 of 16 cells. The corresponding average effects and intervals across trained models are shown in Supplementary Figure 3; the cell observations are repeated outcomes from those same models, not additional independent replicates.

For the average FID-RadImageNet differences, two-sided p-values were 0.0058 against RoentGen-v2 and 0.0604 against FairDiffusion; after Holm adjustment across these two comparisons, they were 0.0116 and 0.0604, respectively.

Across alternative variance models and subsets of runs, both average FID-RadImageNet differences favoured CompDiff. Intervals for RoentGen-v2 excluded zero throughout. For FairDiffusion, the primary interval and both subset intervals included zero, whereas intervals based on a common variance, covariance clustered by training run or the percentile bootstrap excluded it. These checks do not replace the primary analysis. Permutations of method labels for entire runs gave two-sided $p = 0.010$ against RoentGen-v2 and $p = 0.035$ against FairDiffusion under the stronger exchangeability assumption. Omitting any single trained model preserved the direction of both average differences, with estimates ranging from -2.658 to -2.001 against RoentGen-v2 and -1.540 to -0.882 against FairDiffusion.

Supplementary Table 2 records every selected checkpoint. All selected test manifests passed checks for full split coverage and finite required metrics. Excluding the two CompDiff seed-42 models, or restricting every method to the other two runs, preserved the direction of both FID-RadImageNet contrasts. The analysis targets the fixed cells and real test cohorts; it does not estimate uncertainty from resampling patients or choosing different holdout tasks.

Supplementary Table 2: Selected checkpoints for the zero-shot comparison across repeated training runs.

Holdout	Generator	Training seed	Selected step
A	CompDiff†	42	8,500
A	CompDiff	43	8,500
A	CompDiff	44	10,000
A	FairDiffusion	42	11,000
A	FairDiffusion	43	7,500
A	FairDiffusion	44	10,000
A	RoentGen-v2	42	17,500
A	RoentGen-v2	43	7,500
A	RoentGen-v2	44	10,000
B	CompDiff†	42	7,500
B	CompDiff	43	8,500
B	CompDiff	44	6,000
B	FairDiffusion	42	11,000
B	FairDiffusion	43	11,000
B	FairDiffusion	44	18,500
B	RoentGen-v2	42	11,000
B	RoentGen-v2	43	10,000
B	RoentGen-v2	44	10,000

All runs used effective batch 48 and a horizon for the cosine learning-rate schedule of 30,000 steps. Test evaluation had full split coverage for every selected model. † The two CompDiff seed-42 runs used a different policy for validation coverage. For Holdout B, step 7,500 was retained although the validation minimum was at step 6,000, for which no test generation was available. Exclusion analyses are described in Supplementary Note 1. Run paths, manifest hashes and cell measurements are retained with the analysis source.

The comparison and sensitivity procedures are specified in Section 4.9.

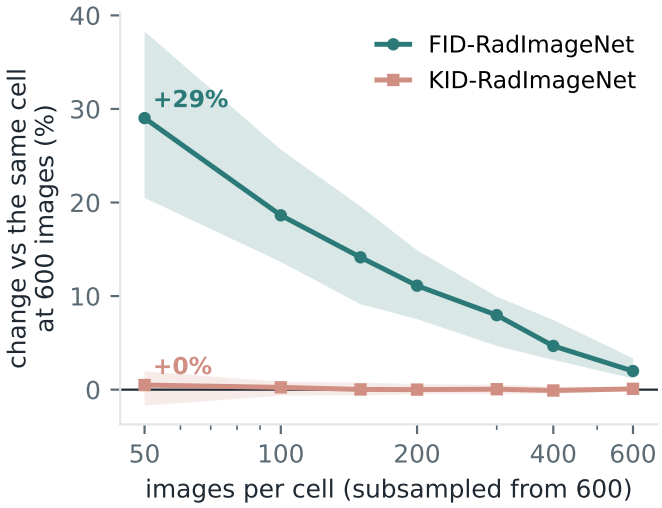
2 Robustness of the zero-shot ranking to the choice of fidelity estimator

Cell evaluation places FID in the regime where it is least well behaved. The estimator is biased upward at small sample sizes, and an intersectional cell is by construction the smallest sample in the study: the rarest cells of Holdout A contain fewer than one hundred images. Because rarity and sample size are the same quantity here, any table of cell fidelity risks confounding a property of the generator with a property of the estimator. The following controls assess this effect in one selected model per configuration and holdout, including both composer variants; they do not repeat the full training-run comparison in Supplementary Note 1.

The estimator is size-dependent at fixed model and fixed data

Relative to an anchor of $s = 600$, FID-RadImageNet inflates by 12.3% at $s = 150$ and by 25.1% at $s = 60$ (absolute: +0.92 and +1.89 FID-RN). At $s = 50$ the inflation is positive in all 32 cell×arm combinations examined, ranging from 14.6% to 49.1%. The kernel counterpart KID, which is an unbiased estimator[1], is flat on the identical subsamples (mean drift -0.4% and +0.5% on Holdouts A and B, positive in 18/32 combinations, i.e. scatter without direction). The contrast between the two metrics on the same images is direct evidence that the inflation is an artefact of the estimator rather

than a degradation of the samples, and Supplementary Figure 5 shows it directly.



Supplementary Figure 5: The fidelity estimator is size-dependent; the kernel estimator is not. Both metrics are read on the *same* subsamples: each cell's candidate and reference sets are reduced to a common size and the metric is expressed as a percentage change against that cell's own reading at 600 images. The generator, the cell and the image pool are fixed, so any vertical movement is attributable to the estimator alone. Lines are medians over the 32 cell × arm curves with the interquartile range shaded; both series share one axis, so the difference in magnitude can be read directly. FID-RadImageNet inflates as the cell shrinks, by a median 29% at 50 images, while KID-RadImageNet[1], which is unbiased, stays at zero on the identical images. The ranking consequences of this bias are given in Supplementary Table 3.

Extrapolation does not recover values for small cells

A natural remedy is to extrapolate FID to infinite sample size[2]. In this setting the remedy is only partial. Fitting FID-RN against $1/k$ on large cells is well conditioned in the aggregate ($R^2 = 0.94$ median, slope ≈ 66, coefficient of variation across cells 8.6%, so the bias is close to universal rather than specific to each cell), and the reference side carries 1.47× the slope of the candidate side, confirming that both arms of the comparison contribute. The fit is nonetheless misspecified near the sample sizes that matter: the local slope at $k \approx n$ is 3.2× the wide-range fitted slope (205 versus 66), because every cell here has k far below the feature dimension $d = 2048$. Validated against a withheld target on 12 large cell×arm combinations, extrapolation from a 150-image subsample recovers only 36% of the gap to the cell's own reading from the full sample and retains a systematic +0.73 FID-RN overestimate (relative MAE 14.1%, versus 20.8% for the uncorrected subsample; SD across runs 0.38). We therefore report corrected values as a robustness check only, and do not use them in place of the primary measurements.

Rankings within each cell are largely preserved across estimators

Supplementary Table 3 compares the arm rankings across estimators.

The conclusions drawn in Section 4.4 are comparisons *between arms within a cell*, evaluated at a common sample size. A size-dependent bias that is shared by all arms in a cell cancels in such a comparison, and the data confirm this. Across the 42 cells for which all arms are available, the arm ordering is identical under $1/k$-corrected FID in 42 cells and under the joint $(1/k, 1/m)$ correction

Supplementary Table 3: Robustness of the zero-shot arm ranking to the fidelity estimator. Each row rescores every cell with a different estimator and compares the resulting cell arm ordering against the ordering under the primary metric. Agreement is reported over all cells and over the withheld intersectional cells alone. Estimators: FID_∞ extrapolates to infinite sample size in the candidate dimension ($1/k$) or jointly in candidate and reference ($1/k, 1/m$); FID at common n rescores every cell at a shared sample size; KID is an unbiased kernel estimator. Ordering agreement counts require the complete arm ordering to match.

Estimator	Same ordering	Same winner	Withheld: same ordering	Median ρ
FID_∞ ($1/k$)	42/42	42/42	12/12	0.998
FID_∞ ($1/k$, $1/m$)	41/42	41/42	12/12	1.00
KID	33/42	38/42	10/12	0.968

in 41 cells; the winning arm is unchanged in all 42 cells under the former and in 41 under the latter. Restricted to the withheld intersectional cells large enough to subsample (the 12 cells with at least 120 images; the four smallest Holdout A cells fall below that floor and are excluded here exactly as in Supplementary Note 3), both the ordering and the winner are preserved in 12 of 12 cells, with median Spearman correlation within each cell $\rho = 1.00$. Under KID, an estimator with a different bias structure entirely, the winner is unchanged in 38 of 42 cells and the full ordering in 33. Restricting attention to the three generators of Figure 6 can only increase these agreement rates, since agreement on a total order implies agreement on each of its suborders.

Fidelity comparisons between cells are not supported by these data

The same analysis places a limit on what cell tables can be asked to show. Sample size explains only a median $R^2 = 0.07$ of the variation in FID-RN *across* cells, and rescoring every cell at a common sample size removes just 5.5% of that variation: the spread between cells is therefore not predominantly a size artefact either. But the direction of the remaining spread does not support a rare-cell deficit. Comparing withheld against well represented cells within each split×arm panel, the withheld cells read *better* by a median of -13.1% under the primary metric, -13.0% under joint extrapolation, -14.2% at common sample size and -15.9% under KID; the withheld group is the worse of the two in at most 1 of 8 panels under any estimator, and in 0 under KID.

We read this as a limitation of the comparison rather than as evidence that rare cells are generated better. A rare intersection may have a narrower and more homogeneous real reference distribution, which can be an easier target to match irrespective of generator behaviour, and the reference side of the estimator is precisely where the residual size sensitivity is largest. The defensible conclusion is negative and methodological: fidelity within each cell should be compared *between arms within a cell at matched sample size*, as in Section 4.4 and Figure 6, and differences between cells should not be interpreted as differences in generation quality.

3 Post hoc sample selection does not substitute for demographic conditioning

A conditioning architecture earns its complexity only if the same benefit cannot be obtained by sampling more and discarding the worse samples. We test this directly. The selection control is specified in Section 4.8.

Selection works, and it works for every arm. Over 48 withheld cell×arm combinations the improvement

is positive with a bootstrap confidence interval excluding zero in all of them, with a median relative gain of 13.6%. It is, however, neither specific to rare cells nor large: well represented cells improve by a statistically indistinguishable 12.6% (Mann–Whitney $p = 0.26$), two thirds of the gain is already present at a 2× budget (9.2%), and an oracle selector that is allowed to see the evaluation references reaches only 17.8%, of which the deployable verifier already captures 76%.

The direct comparison tests whether selection can replace the conditioner. Giving RoentGen-v2 a 4× selection budget and giving CompDiff none, the selected baseline attains a lower FID-RN in 1 of 12 withheld cells. Ranking all arm×mode configurations jointly, a configuration with a dedicated demographic conditioner is best in 12 of 12 cells, and no prompt-conditioned arm, with or without selection, is best in any cell. Sampling compute and demographic conditioning are therefore not interchangeable: a fourfold increase in sampling budget does not recover what routing demographics through a dedicated demographic conditioner provides, and the two compose rather than compete.

The best configuration varies between the factorised composer and the unstructured composer with a matched parameter count. Both preserve the separate attribute encoders and supervision of the output tokens. This agrees with the composer control: the benefit does not require pairwise factorisation, provided the encoders and supervised output tokens are retained.

Two limitations bound this experiment. Selection operates over the pool of samples actually generated, so it varies prompt and sampling noise jointly and does not isolate the contribution of the noise draw. And cells below 120 images do not admit a stable contrast between selected and random samples, which excludes 4 of the eight Holdout A withheld cells, the rarest in the study, from this analysis. The conclusions therefore apply to the intersections with sufficient samples for this selection analysis.

4 Zero-shot generalisation with an unstructured composer of matched size

The composer control of Section 4.7 supports the mechanism analysis in Section 2.7. It compares one selected model per composer and holdout.

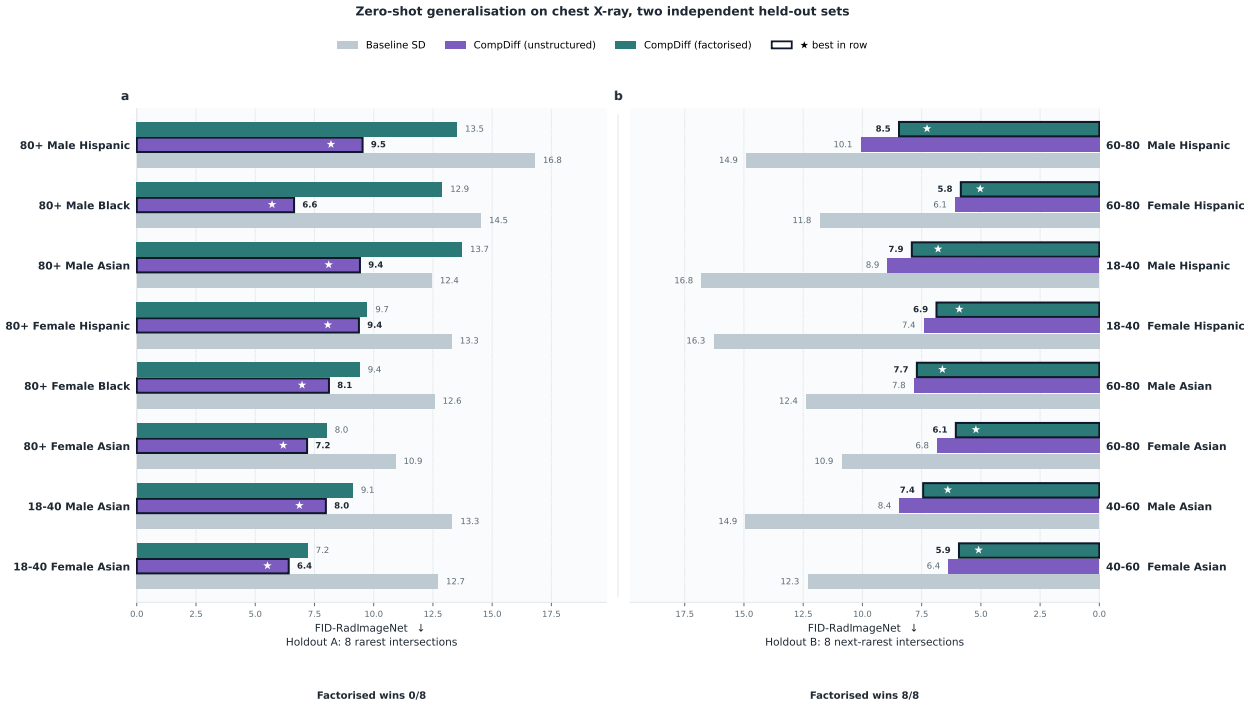
Supplementary Figure 6 plots FID-RadImageNet within each cell for the two composers on all 16 withheld cells, with the prompt-conditioned baseline retained to give the scale on which both composers sit. The pattern is the dissociation summarised in the main text: the unstructured composer attains the lowest value on every Holdout A cell and the factorised composer on every Holdout B cell, while both stay well below the baseline throughout. The gap between composers is comparable to the variability across independent training runs of the final configuration (FID-RadImageNet spread 2.09; Supplementary Table 4), so no effect of composer topology reproduces across the two splits in either direction. The same holds at the validation layer, where both arms are scored under an identical protocol: across the 3 pairings that Holdout B admits (matched by training step, and each arm's checkpoint with the best validation score under either selection criterion), the largest gap between composers in withheld mean FID-RadImageNet is 1.01. Supplementary Table 5 reports the one consistent difference, the convergence speed of demographic control. This is a comparison of composer variants; Figure 6 compares three independently trained models per method and holdout.

Supplementary Table 4: Paired composer control on the full data corpus, evaluated on the withheld *test* split at a matched step of 10,000 (all 7,039 test images, mean ± SD over three generation seeds). The two arms were trained together under an identical recipe and differ only in the composer. The last row is the spread across three independent training runs of the final configuration, the scale against which these differences should be read.

Composer	FID↓	FID-RN↓	Sex↑	Race↑	Age↓	AUROC↑
Factorised	62.56 ± 0.16	6.08	0.998	0.962	8.84 ± 0.06	0.78
Unstructured (param.-matched)	61.99 ± 0.15	7.05	0.995	0.955	10.85 ± 0.10	0.79
Spread across training runs	3.18	2.09	–	–	4.23	–

Supplementary Table 5: First checkpoint at which each full data run satisfies all three controllability floors of Eq. 9. The three factorised runs are independent training runs of the final configuration.

Run	All floors first met
Factorised, run 1	10,000
Factorised, run 2	7,500
Factorised, run 3 (matched pair)	8,500
Unstructured (parameter-matched)	11,000



Supplementary Figure 6: Fidelity of the two composers at unseen intersections. **a**, Holdout A, the eight rarest intersections. **b**, Holdout B, the eight next-rarest intersections. Both panels show FID-RadImageNet within each cell (lower is better) for the factorised composer (the seed-42 CompDiff models in Supplementary Table 2), the unstructured composer with a matched parameter count, and the RoentGen-v2 baseline as a scale reference; FairDiffusion is omitted because this figure answers a composer question. Every arm is scored at full coverage of its split (8,039 and 10,497 images). Checkpoint selection is specified in Section 4.7. The footer counts wins for the *factorised* arm against the other two arms shown, so 0/8 on Holdout A means the unstructured control attains the lowest value in all eight of those cells, and 8/8 on Holdout B the reverse; the sign of the composer difference does not reproduce across the two splits.

5 Reader study contrasts and sensitivity analyses

The reader protocol is described in Section 4.5, and the statistical analyses in Section 4.9. Supplementary Table 6 reports the complete contrast family; Supplementary Table 7 reports the sensitivity analyses.

Supplementary Table 6: Mixed-effects contrasts on the blinded reader study. Rating estimates are differences in mean score from Equation (12), with two-sided Wald tests and 95% intervals. Realism uses 287 ratings from 92 sets; correspondence uses 277 ratings from 89 sets. Both models include the two readers. p_{BH} is Benjamini–Hochberg-adjusted across the 8 comparisons. Forced-choice rows give selections for each source among assessments containing both. Two-sided exact binomial tests condition on selection of either compared source, giving 50 and 79 trials, respectively; preferences for a third source are excluded. Numerical p-values are shown to three significant digits.

Outcome	**Contrast**	Δ	**95% CI**	p	p_{BH}
Realism	CompDiff vs. RoentGen-v2	+1.81	[+1.48, +2.15]	4.20×10^{-26}	3.36×10^{-25}
	CompDiff vs. FairDiffusion	+0.04	[−0.29, +0.37]	0.812	0.812
	FairDiffusion vs. RoentGen-v2	+1.77	[+1.44, +2.10]	1.14×10^{-25}	4.56×10^{-25}
Prompt corr.	CompDiff vs. RoentGen-v2	+0.55	[+0.20, +0.91]	0.00236	0.00472
	CompDiff vs. FairDiffusion	+0.42	[+0.07, +0.77]	0.0191	0.0306
	FairDiffusion vs. RoentGen-v2	+0.13	[−0.21, +0.48]	0.45	0.514
Most realistic	CompDiff vs. RoentGen-v2	44–6	(preferences)	3.24×10^{-8}	8.65×10^{-8}
	CompDiff vs. FairDiffusion	44–35	(preferences)	0.368	0.491

Supplementary Table 7: Robustness of the reader study contrasts. Δ and the Wald column come from the mixed-effects model of Equation (12). The paired sample size counts reader–set assessments with both source ratings and applies to the permutation and Wilcoxon tests. The permutation column uses 10,000 sign-flip draws; the Wilcoxon column uses signed ranks. The Cluster column gives Wald tests from ordinary least squares with standard errors clustered by image set. The mixed and ordinary least-squares models use 287 realism ratings from 92 sets or 277 correspondence ratings from 89 sets. All tests are two-sided; p-values are unadjusted and shown to three significant digits.

Outcome	**Contrast**	Δ	n_{paired}	**Wald** p	**Permutation** p	**Wilcoxon** p	**Cluster** p
Realism	CompDiff vs. RoentGen-v2	+1.81	84	4.20×10^{-26}	1.00×10^{-4}	2.34×10^{-10}	7.34×10^{-21}
	CompDiff vs. FairDiffusion	+0.04	89	0.812	0.673	0.608	0.807
	FairDiffusion vs. RoentGen-v2	+1.77	89	1.14×10^{-25}	1.00×10^{-4}	2.97×10^{-10}	4.38×10^{-19}
Prompt corr.	CompDiff vs. RoentGen-v2	+0.55	80	0.00236	0.0132	0.012	0.0121
	CompDiff vs. FairDiffusion	+0.42	85	0.0191	0.0498	0.0458	0.011
	FairDiffusion vs. RoentGen-v2	+0.13	86	0.45	0.52	0.48	0.699

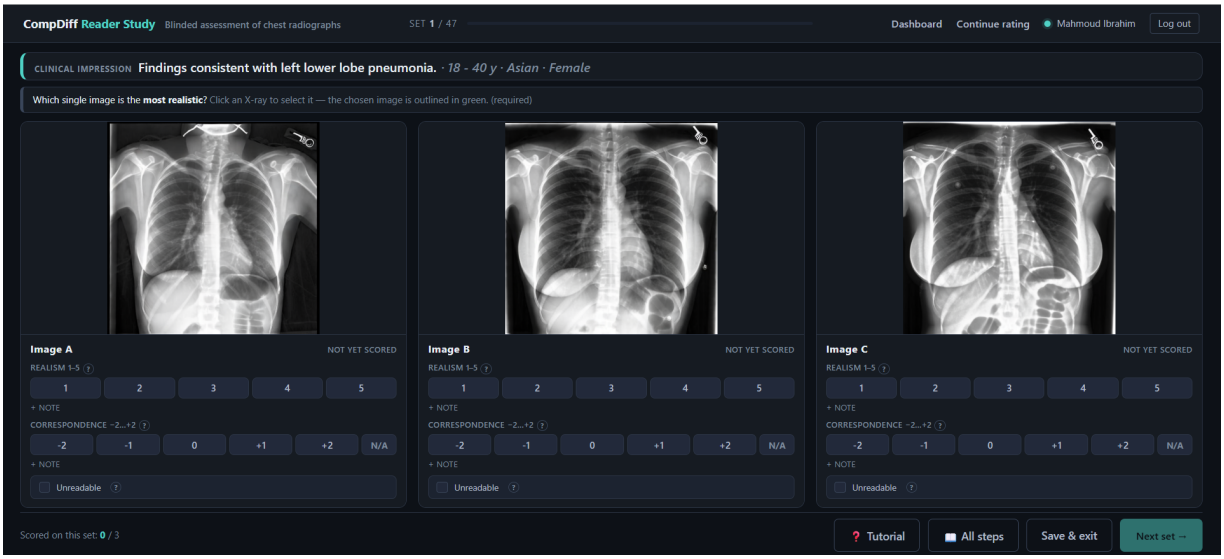
The source comparisons are consistent across the sensitivity analyses. The CompDiff–FairDiffusion correspondence contrast moves from $p = 0.019$ under the mixed-effects model to $p = 0.0498$ under the permutation test and is the weaker of the two correspondence results. Treating reader as a variance component reproduces every contrast to two decimal places.

The realism difference between CompDiff and FairDiffusion is +0.04 points (95% confidence interval $[-0.29, +0.37]$; $p = 0.81$). The equivalence test gives $p = 0.005$ on 89 paired sets, supporting equivalence within the analytical margin. For correspondence, the variance between sets is 1.62 against a residual of 1.41 (intraclass correlation 0.53). For realism, the set variance is estimated at zero against a residual of 1.32. These differences explain why the pooled correspondence standard deviations overstate the uncertainty of within-set comparisons and why the permutation and cluster-robust checks support the realism inference.

The 12 hidden real anchors occur in different sets. Their correspondence mean adjusted by the model is unstable; Table 3 therefore reports the raw mean. The realism and forced-choice anchor comparisons are unaffected.

6 Reader study rating interface

Supplementary Figure 7 reproduces the viewer used in the blinded reader protocol (Section 4.5).



Supplementary Figure 7: The blinded reader study rating interface (Section 2.4). The clinical impression is displayed at the top, followed by three chest X-rays displayed side by side in randomised order; each image carries an anatomical realism control (1–5) and a prompt correspondence control (−2 to +2), with a forced "most realistic" choice below the row and unreadable and N/A options for each image. Raters are blinded to which images are synthetic or real and in what proportion.

7 Validation trajectories

Supplementary Tables 8–10 report the complete validation trajectories, with the checkpoints selected by Equation (9) marked $\star$.

Overfitting within the shared budget. All three generators degrade well inside the shared budget. Each reaches its minimum validation FID within the first 11,000 steps (CompDiff at step 3,500, Baseline at 8,500, FairDiffusion at 11,000); by step 13,500 every generator exceeds an FID of 100, and by step 30,000 each sits at 1.4–2.2$\times$ its own minimum, with MS-SSIM roughly halved (from 0.34–0.36 at step 1,000 to 0.16–0.19) and disease AUROC below its peak during training for all three (Supplementary Tables 8–10). Because every trajectory is logged to the end of the budget, each selected checkpoint is a minimum over the complete trajectory rather than the best point of a truncated one.

Supplementary Table 8: Validation trajectory for the **Baseline SD** generator over the shared 30,000-step budget (training run with seed 11). This run is one of the three independently trained Baseline SD models whose selected checkpoints are averaged into the Baseline SD bars of the main text; the values tabulated here are those of this single run, not the means across three runs, and its selected checkpoint is marked ⋆. All three generators are scored at the identical 24 checkpoints. The selected checkpoint (step 8,500) is the checkpoint with the lowest FID among those meeting the demographic control floors of Section 4.3; rows marked † fail at least one floor and are therefore not eligible, regardless of their FID. Metrics follow Section 4.4 (FID-RIN = FID-RadImageNet; MS-SSIM as defined there, read against the 0.25–0.75 band; Sex/Race = prompt alignment accuracy; Age RMSE in years) and are computed on the validation split.

Step	FID↓	FID-RIN↓	MS-SSIM	AUROC↑	Sex↑	Race↑	Age RMSE↓
1,000	96.71	8.08	0.34	0.63	1.000	0.968	4.73
2,500	96.18	9.85	0.32	0.75	0.999	0.972	4.34
3,500	98.21	9.93	0.30	0.70	0.999	0.977	5.71
5,000	91.33	9.55	0.31	0.69	0.997	0.981	5.13
6,000	92.43	8.84	0.32	0.69	0.998	0.986	5.29
7,500	90.67	10.32	0.29	0.73	0.994	0.972	5.63
⋆ **8,500**	**89.99**	**8.63**	**0.31**	**0.72**	**0.989**	**0.972**	**7.70**
10,000	92.88	9.47	0.30	0.76	0.995	0.969	6.67
11,000	96.67	10.84	0.27	0.78	0.991	0.979	8.53
12,500	99.95	11.77	0.28	0.70	0.990	0.970	7.17
13,500	105.81	12.35	0.26	0.70	0.985	0.961	9.26
15,000	107.51	12.54	0.25	0.65	0.988	0.969	8.27
16,000	108.82	12.81	0.26	0.63	0.991	0.979	9.03
17,500†	109.97	12.71	0.24	0.63	0.976	0.946	12.20
18,500†	109.82	14.47	0.24	0.58	0.975	0.953	9.42
20,000†	140.94	17.59	0.18	0.62	0.984	0.962	10.18
21,000	143.33	18.68	0.16	0.56	0.995	0.981	8.03
22,500†	152.85	18.41	0.16	0.53	0.975	0.937	13.54
23,500†	149.92	18.70	0.15	0.61	0.978	0.924	14.63
25,000†	148.03	18.76	0.16	0.57	0.979	0.934	12.56
26,000†	150.76	18.77	0.16	0.56	0.983	0.930	11.65
27,500	125.83	14.92	0.21	0.65	0.985	0.969	8.90
28,500†	150.81	19.65	0.15	0.56	0.985	0.959	10.30
30,000†	150.03	18.23	0.16	0.64	0.978	0.923	10.50

Supplementary Table 9: Validation trajectory for the **FairDiffusion** generator over the shared 30,000-step budget (training run with seed 42). This run is one of the three independently trained FairDiffusion models whose selected checkpoints are averaged into the FairDiffusion bars of the main text; the values tabulated here are those of this single run, not the means across three runs, and its selected checkpoint is marked ⋆. All three generators are scored at the identical 24 checkpoints. The selected checkpoint (step 11,000) is the checkpoint with the lowest FID among those meeting the demographic control floors of Section 4.3; rows marked † fail at least one floor and are therefore not eligible, regardless of their FID. Metrics follow Section 4.4 (FID-RIN = FID-RadImageNet; MS-SSIM as defined there, read against the 0.25–0.75 band; Sex/Race = prompt alignment accuracy; Age RMSE in years) and are computed on the validation split.

Step	FID↓	FID-RIN↓	MS-SSIM	AUROC↑	Sex↑	Race↑	Age RMSE↓
1,000†	110.99	6.74	0.36	0.61	0.998	0.902	7.83
2,500	94.71	8.38	0.35	0.64	1.000	0.967	5.83
3,500	100.23	8.48	0.32	0.70	0.999	0.972	9.12
5,000	95.55	7.74	0.35	0.65	0.999	0.962	7.96
6,000	94.57	8.72	0.35	0.68	0.998	0.979	6.22
7,500	90.74	8.40	0.34	0.67	0.997	0.980	7.29
8,500	94.64	8.89	0.34	0.68	0.998	0.970	7.58
10,000	98.90	10.05	0.32	0.75	0.998	0.976	7.81
⋆ 11,000	**86.93**	**9.56**	**0.33**	**0.73**	**0.999**	**0.981**	**5.56**
12,500	98.14	9.85	0.32	0.71	0.998	0.977	7.79
13,500	100.64	11.24	0.31	0.74	0.999	0.976	5.55
15,000	99.00	11.74	0.29	0.72	0.999	0.982	6.07
16,000	104.32	13.95	0.27	0.72	0.996	0.977	7.52
17,500	115.92	15.21	0.25	0.57	1.000	0.987	6.32
18,500	97.70	12.22	0.28	0.73	0.996	0.982	8.12
20,000	112.69	14.30	0.25	0.64	0.994	0.983	7.20
21,000	123.85	15.63	0.22	0.62	0.985	0.963	9.40
22,500	117.35	14.81	0.24	0.61	0.988	0.979	9.05
23,500	119.03	15.18	0.23	0.65	0.991	0.971	9.30
25,000	119.25	15.54	0.23	0.65	0.996	0.973	7.09
26,000	132.83	18.18	0.20	0.59	0.991	0.960	8.22
27,500	115.29	14.25	0.24	0.65	0.993	0.981	7.67
28,500	150.26	20.06	0.17	0.53	0.992	0.965	7.76
30,000†	123.66	16.88	0.19	0.65	0.980	0.970	10.26

Supplementary Table 10: Validation trajectory for the **CompDiff** generator over the shared 30,000-step budget (training run with seed 42). This run is one of the three independently trained CompDiff models whose selected checkpoints are averaged into the CompDiff bars of the main text; the values tabulated here are those of this single run, not the means across three runs, and its selected checkpoint is marked ⋆. All three generators are scored at the identical 24 checkpoints. The selected checkpoint (step 10,000) is the checkpoint with the lowest FID among those meeting the demographic control floors of Section 4.3; rows marked † fail at least one floor and are therefore not eligible, regardless of their FID. Metrics follow Section 4.4 (FID-RIN = FID-RadImageNet; MS-SSIM as defined there, read against the 0.25–0.75 band; Sex/Race = prompt alignment accuracy; Age RMSE in years) and are computed on the validation split.

Step	FID↓	FID-RIN↓	MS-SSIM	AUROC↑	Sex↑	Race↑	Age RMSE↓
1,000†	78.39	5.89	0.35	0.64	0.552	0.663	17.75
2,500†	65.76	4.69	0.37	0.70	0.872	0.868	14.58
3,500†	63.53	4.79	0.37	0.73	0.983	0.911	14.45
5,000†	66.74	5.68	0.36	0.71	0.993	0.923	13.58
6,000†	69.01	6.01	0.36	0.77	0.992	0.921	13.23
7,500†	71.00	7.02	0.37	0.76	0.997	0.960	13.35
8,500†	75.17	7.33	0.35	0.71	0.998	0.961	10.39
⋆ 10,000	**71.63**	**6.06**	**0.38**	**0.75**	**0.999**	**0.956**	**8.96**
11,000	82.56	8.07	0.33	0.80	0.999	0.966	8.98
12,500	87.10	9.33	0.31	0.74	1.000	0.977	8.54
13,500	102.08	12.70	0.25	0.77	0.985	0.967	8.97
15,000	106.69	13.19	0.25	0.73	0.993	0.962	8.36
16,000	103.41	12.57	0.26	0.70	0.998	0.966	8.01
17,500	99.30	11.15	0.26	0.70	0.992	0.965	9.09
18,500	121.15	14.87	0.22	0.64	1.000	0.977	7.66
20,000†	115.68	13.82	0.23	0.62	0.983	0.975	10.28
21,000	110.41	14.20	0.23	0.62	0.996	0.974	9.31
22,500†	115.63	16.08	0.19	0.63	0.982	0.969	10.55
23,500	137.83	18.65	0.17	0.63	0.991	0.980	9.97
25,000	127.50	17.50	0.17	0.58	0.990	0.972	9.44
26,000†	135.04	17.93	0.16	0.61	0.980	0.975	10.02
27,500†	120.47	16.03	0.19	0.56	0.972	0.972	11.26
28,500†	116.95	16.42	0.18	0.67	0.988	0.987	10.95
30,000†	137.40	17.16	0.18	0.59	0.981	0.970	11.64

8 Downstream diagnostic and subgroup evidence

The analyses in Sections 4.6 and 4.9 use a consistent set of eight diagnostic labels throughout: atelectasis, cardiomegaly, consolidation, oedema, pleural effusion, pneumonia, pneumothorax and no finding. Repeated classifier runs measure variability in classifier training on fixed generated corpora; they do not represent independently trained generators.

Inference and sensitivity procedures are specified in Section 4.9.

Repeated training and the complete configuration sweep. For classifiers trained on synthetic images alone under the verified common AdamW recipe, CompDiff had the higher mean AUROC_8 in every corpus regime on both test datasets (Supplementary Figure 8). Omitting any single classifier run retained a positive average contrast on each dataset. Supplementary Table 11 compares the three training strategies using AdamW. Mixing real and synthetic images in one training stage did not yield a consistent advantage.

Supplementary Table 11: Repeated classifier training using AdamW and a consistent set of eight diagnostic labels.

Training	Test set	RoentGen-v2	CompDiff	Difference	95% interval
Synthetic alone	MIMIC-CXR	0.7849	0.7962	+0.0113	[0.0067, 0.0158]
Synthetic alone	CheXpert	0.7275	0.7363	+0.0088	[0.0039, 0.0137]
Pretrain, fine-tune	MIMIC-CXR	0.8577	0.8591	+0.0013	[−0.0026, 0.0053]
Pretrain, fine-tune	CheXpert	0.7810	0.7839	+0.0029	[−0.0007, 0.0066]
Mixed training	MIMIC-CXR	0.8457	0.8467	+0.0010	[−0.0006, 0.0026]
Mixed training	CheXpert	0.7588	0.7528	−0.0061	[−0.0096, −0.0025]

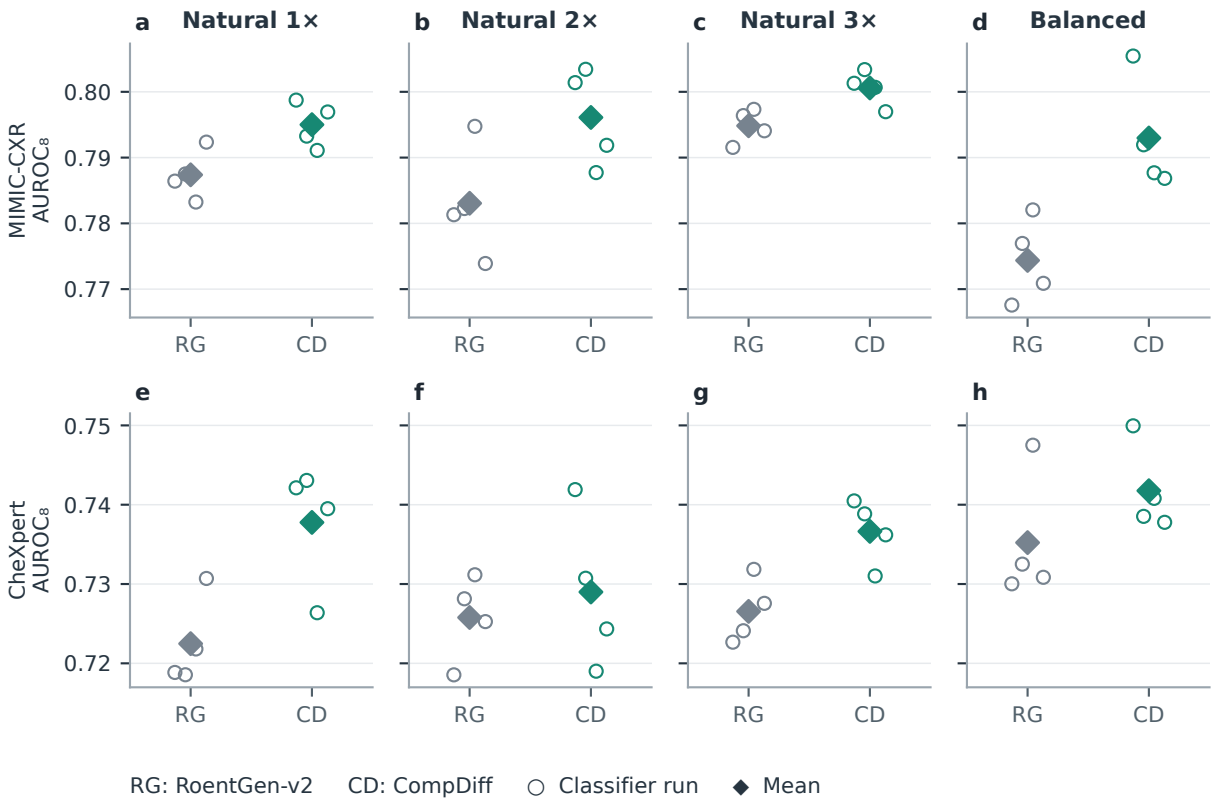
Each row averages four synthetic corpus regimes equally, with four independent classifier runs per generator and regime. The same classifiers are evaluated on both test sets. Intervals are pointwise Welch intervals across classifier runs, conditional on the corpora and test cohorts; no pairing by run index is assumed. Training uses 256×256 inputs, batch size 128, an initial learning rate of 0.001 and up to 50 epochs; test inputs are 224×224. Pretraining followed by fine-tuning uses ImageNet initialisation and all real training images; mixed training combines real and synthetic images in one stage. Differences are CompDiff minus RoentGen-v2. These additional comparisons are exploratory and intervals are not adjusted for multiplicity.

Including all five available models from the AdamW family, with the additional model that could inherit a different recipe, also retained positive average differences and positive corpus means on both datasets (Supplementary Table 12).

Supplementary Table 12: Sensitivity to including all available classifiers trained on synthetic images alone within the AdamW family.

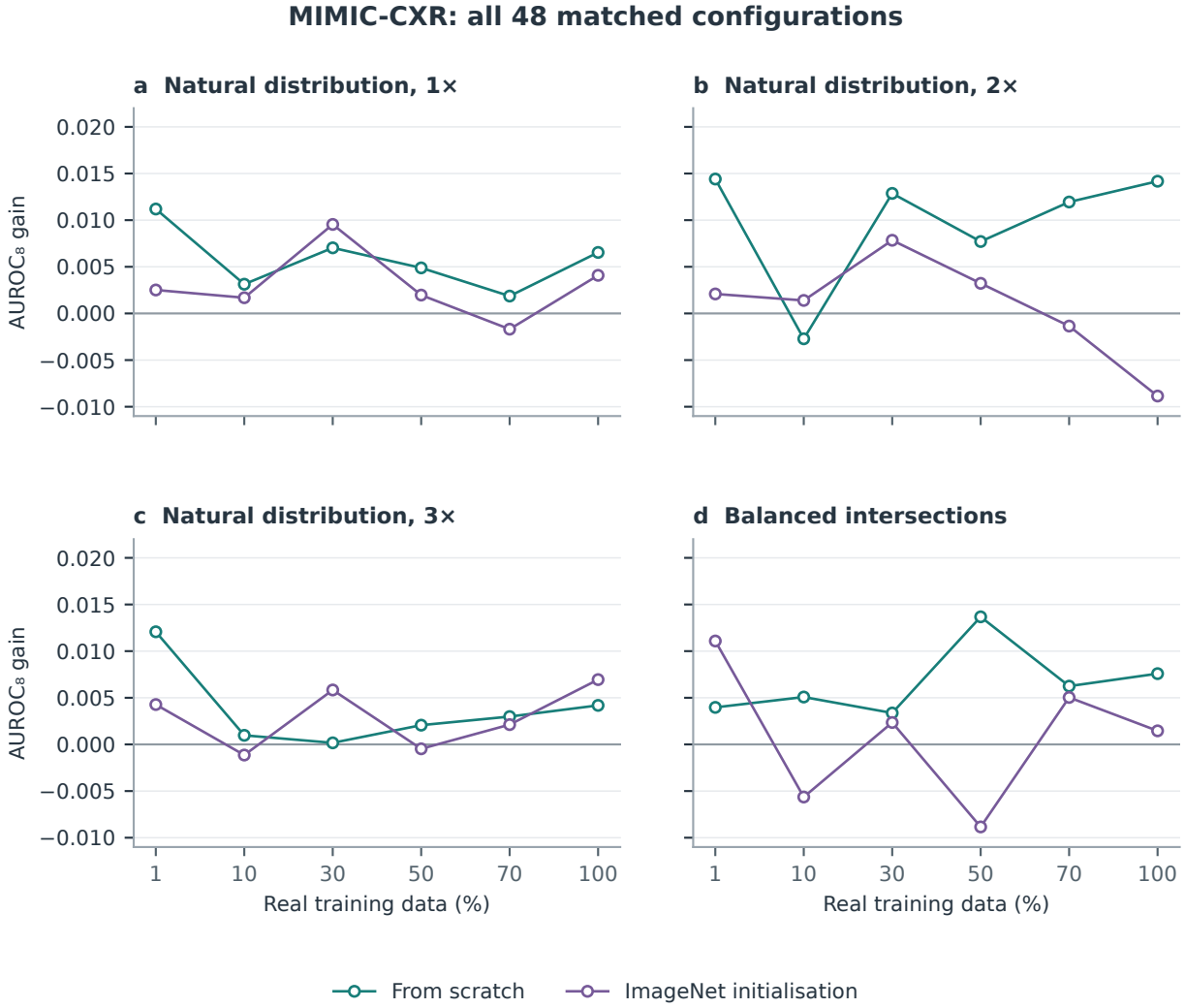
Test set	RoentGen-v2	CompDiff	Difference	Positive corpora / 4
MIMIC-CXR	0.7864	0.7949	+0.0086	4
CheXpert	0.7282	0.7365	+0.0083	4

Descriptive averages include all five available models per generator and corpus, including the additional model that may come from an earlier training recipe. The four corpora receive equal weight. These models are not treated as five exchangeable repeats of the same training protocol; the main intervals use the four runs with a verified common recipe. Including the additional models retains positive average differences and positive corpus means for CompDiff under the AdamW family on both test sets.

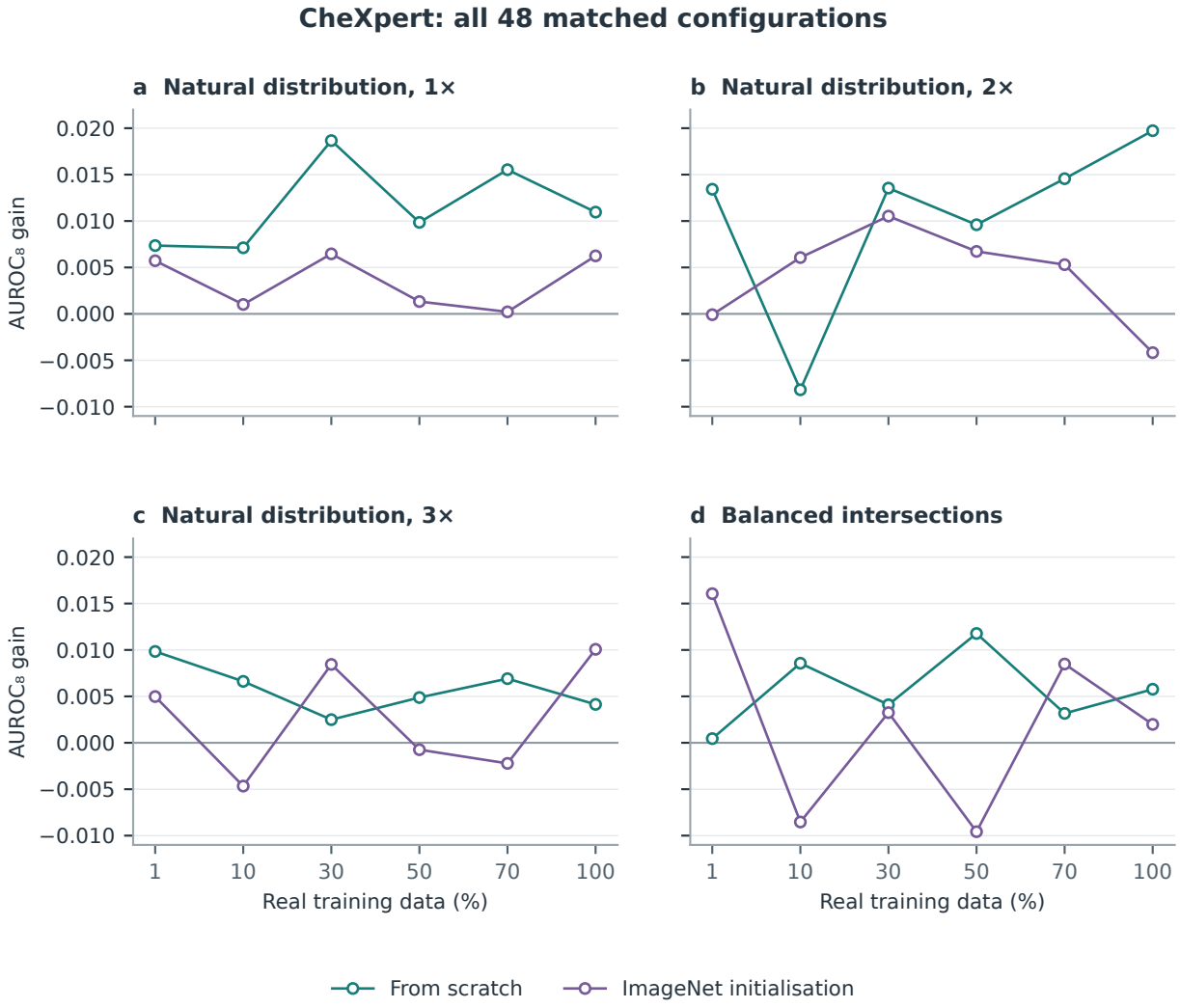


Supplementary Figure 8: Independent classifier runs using synthetic chest images alone. **a**–**d**, Evaluation on MIMIC-CXR for the natural 1×, 2× and 3× corpora and the balanced corpus. **e**–**h**, The same classifiers evaluated on CheXpert, in the same order. Each panel shows four classifier runs per generator for one fixed corpus regime under the AdamW recipe. Hollow circles are individual runs and filled diamonds are their means. RG denotes RoentGen-v2 and CD denotes CompDiff. $AUROC_8$ uses the same eight diagnostic labels as Figure 9. The same 32 trained classifiers are evaluated on MIMIC-CXR and CheXpert. Natural corpora contain one, two or three times the real training set size; the balanced corpus samples demographic intersections uniformly.

Supplementary Figures 9 and 10 show every configuration in the fine-tuning sweep. The mean CompDiff gain over the six real data fractions was positive for each combination of initialisation and synthetic corpus on both datasets. Mean AUROC and area under the precision–recall curve (AUPRC) gains were positive for all eight labels on both datasets (Supplementary Table 13). The exploratory sensitivity comparison at 95% specificity also retained positive mean gains for all eight labels, supporting the direction of the comparison at 90% specificity.



Supplementary Figure 9: Complete sweep of downstream training configurations on MIMIC-CXR. Panels show the four synthetic corpus regimes; lines distinguish initialisation from scratch and from ImageNet. Every point is a CompDiff minus RoentGen-v2 AUROC_8 difference at one fraction of real data used for fine-tuning, giving 48 matched configurations. Positive values favour CompDiff. All scores average the same eight diagnostic labels. Axes are shared with the corresponding figure for the other test dataset. Lines connect settings in the sweep and do not imply independent replications or interpolation between evaluated fractions.



Supplementary Figure 10: Complete sweep of downstream training configurations on CheXpert. Panels show the four synthetic corpus regimes; lines distinguish initialisation from scratch and from ImageNet. Every point is a CompDiff minus RoentGen-v2 AUROC_8 difference at one fraction of real data used for fine-tuning, giving 48 matched configurations. Positive values favour CompDiff. All scores average the same eight diagnostic labels. Axes are shared with the corresponding figure for the other test dataset. Lines connect settings in the sweep and do not imply independent replications or interpolation between evaluated fractions.

Supplementary Table 13: Diagnostic gains across the complete fine-tuning sweep.

Test set	Label	AUROC gain	AUPRC gain	Sens. at 90%	Sens. at 95%
MIMIC-CXR	Atelectasis	+0.0020	+0.0042	+0.57	+0.11
MIMIC-CXR	Cardiomegaly	+0.0010	+0.0091	+0.22	+0.32
MIMIC-CXR	Consolidation	+0.0071	+0.0037	+1.95	+2.12
MIMIC-CXR	Oedema	+0.0043	+0.0029	+1.54	+1.26
MIMIC-CXR	Pleural Effusion	+0.0018	+0.0074	+0.68	+1.02
MIMIC-CXR	Pneumonia	+0.0076	+0.0128	+1.34	+0.80
MIMIC-CXR	Pneumothorax	+0.0075	+0.0196	+3.13	+3.03
MIMIC-CXR	No Finding	+0.0023	+0.0024	+0.73	+0.75
CheXpert	Atelectasis	+0.0050	+0.0054	+1.12	+0.63
CheXpert	Cardiomegaly	+0.0022	+0.0079	+0.82	+0.96
CheXpert	Consolidation	+0.0068	+0.0039	+0.75	+0.35
CheXpert	Oedema	+0.0037	+0.0034	+0.89	+0.54
CheXpert	Pleural Effusion	+0.0018	+0.0063	+0.88	+1.23
CheXpert	Pneumonia	+0.0082	+0.0052	+1.19	+1.18
CheXpert	Pneumothorax	+0.0140	+0.0220	+2.85	+2.68
CheXpert	No Finding	+0.0039	+0.0085	+1.06	+1.13

Values are mean CompDiff minus RoentGen-v2 differences across all 48 matched configurations. Sensitivity differences are in percentage points, at the stated specificity. Sensitivity is the maximum true positive rate at a false positive rate no greater than 0.10 or 0.05 on each test ROC. The thresholds are selected separately on the test curves for this exploratory comparison; they are not validated clinical operating thresholds. Configurations share training data and test cohorts, so these descriptive averages do not represent 48 independent replications.

Subgroup coverage and sparse labels. All ten sex, race and age groups had positive median $AUROC_8$ gains separately on each dataset (Supplementary Table 14). Within sex, race and age, the subgroup with the lowest baseline AUROC improved in 40, 43 and 33 of 48 MIMIC-CXR configurations, respectively, and in 41, 41 and 36 of 48 CheXpert configurations. These counts describe consistency across the fixed sweep. They neither imply improvement in every configuration nor establish that disparities between groups narrowed.

Three MIMIC-CXR groups included a diagnostic label with fewer than ten positive images: Asian, Hispanic and age 80+. Requiring at least ten positives for every label excludes those groups; the median gain remains positive for all seven remaining MIMIC-CXR groups and all ten CheXpert groups. Patient resampling shows the uncertainty in the three sparse groups (Supplementary Table 15). In the Hispanic group, the least represented label occurred in only two patients; 137 of 1,000 resamples lacked the coverage needed to compute all eight AUROCs. The reported percentile ranges are therefore conditional on valid label coverage and fixed trained models. These results support the descriptive subgroup pattern while identifying where additional patients are needed for precise estimation.

Supplementary Table 14: Subgroup sample sizes and gains across training configurations.

Test set	Group	Images	Min. positive	Median gain	Interquartile range	Positive / 48
MIMIC-CXR	Female	3407	44	+0.0052	[+0.0012, +0.0082]	37
MIMIC-CXR	Male	3632	65	+0.0033	[+0.0006, +0.0071]	37
MIMIC-CXR	Asian	298	6	+0.0058	[−0.0011, +0.0120]	33
MIMIC-CXR	Black	1489	10	+0.0027	[−0.0016, +0.0081]	33
MIMIC-CXR	Hispanic	611	4	+0.0073	[−0.0046, +0.0170]	31
MIMIC-CXR	White	4641	71	+0.0027	[+0.0013, +0.0078]	42
MIMIC-CXR	18–40	1159	14	+0.0052	[−0.0016, +0.0130]	32
MIMIC-CXR	40–60	2460	37	+0.0031	[+0.0004, +0.0072]	37
MIMIC-CXR	60–80	2791	44	+0.0034	[−0.0004, +0.0061]	35
MIMIC-CXR	80+	629	7	+0.0044	[−0.0022, +0.0113]	33
CheXpert	Female	8009	337	+0.0063	[+0.0024, +0.0107]	41
CheXpert	Male	15120	601	+0.0060	[+0.0013, +0.0088]	38
CheXpert	Asian	3241	139	+0.0041	[+0.0012, +0.0094]	38
CheXpert	Black	1568	42	+0.0075	[+0.0031, +0.0137]	39
CheXpert	Hispanic	3151	128	+0.0080	[+0.0025, +0.0113]	39
CheXpert	White	15169	595	+0.0054	[+0.0007, +0.0096]	38
CheXpert	18–40	3780	100	+0.0068	[+0.0009, +0.0113]	38
CheXpert	40–60	7945	268	+0.0055	[+0.0014, +0.0091]	39
CheXpert	60–80	8957	382	+0.0067	[+0.0018, +0.0103]	40
CheXpert	80+	2428	116	+0.0055	[−0.0000, +0.0108]	36

Every subgroup score averages the same eight diagnostic labels, each with at least one positive and one negative image. Min. positive is the smallest positive image count among those labels. Median and interquartile range describe the 48 configuration differences within each dataset; they are not confidence intervals. Requiring at least ten positive images for every label excludes the MIMIC-CXR Asian, Hispanic and 80+ groups; all remaining subgroup median gains remain positive.

Supplementary Table 15: Patient resampling in the three sparsely sampled MIMIC-CXR groups.

Group	Images	Patients	Min. positive	Median gain	95% valid-draw range	Valid draws
Asian	298	155	5	+0.0058	[+0.0002, +0.0130]	994/1000
Hispanic	611	267	2	+0.0073	[−0.0006, +0.0138]	863/1000
80+	629	321	7	+0.0044	[−0.0009, +0.0113]	999/1000

The point estimate is the median CompDiff minus RoentGen-v2 AUROC gain across the fixed 48 configurations. Min. positive counts distinct patients with a positive image for the least represented label. Each of 1,000 draws resamples patients with replacement and retains all their images and all model predictions. Draws lacking either class for any label cannot yield the specified eight-label AUROC and are excluded. The range gives the 2.5th and 97.5th percentiles among valid draws, conditional on the fixed trained models and valid label coverage. It should not be interpreted as an unconditional 95% confidence interval, particularly for the Hispanic group, where some labels occur in only two patients.

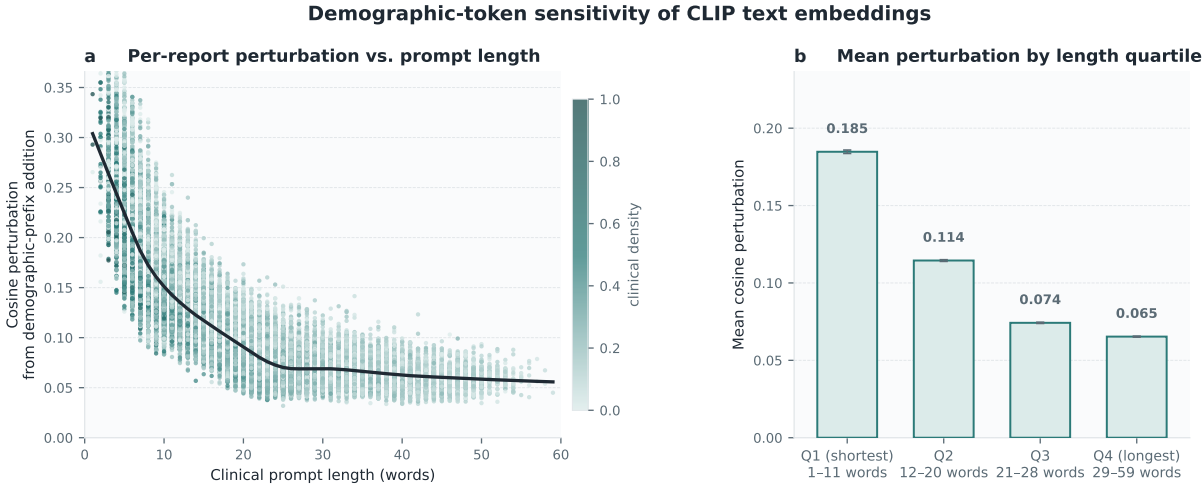
9 Demographic tokens in the prompt: truncation statistics and probe of the encoder

These analyses distinguish prompt truncation from changes within the shared text representation.

The analysis procedure is specified in Section 4.7.

Truncation is negligible. Tokenising the final training prompts of all 62,094 training images (32,319 unique clinical impressions) with Stable Diffusion 2.1's CLIP tokenizer, no clinical impression alone exceeds the 77-token context limit (median 33 tokens per unique impression, maximum 77), and prepending the demographic prefix, which costs exactly 7 tokens for every (age, sex, race) combination, pushes only 5 prompts ($< 0.02\%$) beyond it. Displacement of clinical tokens past the context limit therefore cannot explain the FID gap in the ablation. The prefix nonetheless accounts for a mean 29% (median 30%; 90th percentile 41%) of the tokens in a training prompt, because clinical impressions are short. We therefore examined how demographic text changes the encoder's shared representation.

Probe of the encoder. Two patterns are visible in Supplementary Figure 11. First, the perturbation falls steeply with clinical prompt length and then plateaus (Supplementary Figure 11a): a short impression of 1–11 words is perturbed 2.8× as much as a long one of 29–59 words (Supplementary Figure 11b), so when clinical content is scarce the demographic prefix dominates the representation that conditions the UNet. Second, this shift is not an artefact of demographic information leaking through clinical phrasing: linear probes trained on the embeddings without demographic text recover sex, race, and age only at levels close to chance (balanced accuracy 0.55, 0.26, 0.29 against chance 0.50, 0.25, 0.25). These patterns are consistent with demographic text altering the shared representation. The architectural ablations in Section 2.7 assess how separating the conditioning pathways affects image quality and demographic control.



Supplementary Figure 11: CLIP's text encoder is sensitive to demographic tokens, and this sensitivity falls steeply with clinical prompt length. For each of 36,562 MIMIC-CXR clinical impressions, the clinical text and the same text with each of 24 demographic prefixes (three age descriptors × two sexes × four race labels) are passed through Stable Diffusion 2.1's frozen CLIP text encoder; we report the cosine perturbation between embeddings with and without demographic text, averaged over the 24 demographic variants. **a**, Perturbation for each report against clinical prompt length on linear axes. Each point is one report, coloured by its clinical density (fraction of words that are recognised clinical entities); the dark line is a LOWESS smoother (span 0.20). The same relationship is well described by a power law in log–log space, $y \propto L^{-0.67}$ ($R^2_{\log\log} = 0.77$). **b**, Mean perturbation within quartiles of word count with bootstrap 95% CIs (500 resamples); impressions in the shortest quartile are perturbed 2.8× more than impressions in the longest quartile.

References

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

[2] 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.