WHY FAIR DATASET DISTILLATION FAILS ON COMPLEX DISTRIBUTIONS: A GEOMETRIC ANALYSIS AND A SPREAD-PRESERVING REMEDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Fair dataset distillation compresses a large training set into a small synthetic surrogate that should not amplify disparities across protected-attribute (PA) groups. Existing PA-aware methods differ in how they choose the per-class matching target used to align synthetic images with the real distribution, but agree on its form. Each reduces a class to a single point in feature space. We show that this shared form, not the choice of point, is a limitation. Any objective whose per-class target is a single vector depends on first moments alone, so it can match group means while discarding within-group geometry. Decomposing the squared Bures-Wasserstein distance between each group's real and synthetic feature distributions separates a centroid term, which such objectives control, from a covariance term, which none of them constrain. No relocation of the point can recover this covariance term. The residual grows with within-group variance, which is large for complex real-world distributions such as faces. We propose Fair Geometric Distillation (FGD), whose per-class target is instead a set of per-group coresets of real images approximating each group's distribution under a balanced budget, so the second moment of each group enters the objective. At one fixed regularisation setting, FGD attains the lowest per-group Wasserstein error on all six benchmarks and cuts CelebA DEO_M from 11.81 to 3.65 at 30 images per class (IPC), a 69% reduction over FairDD and 66% over the strongest prior fair method at that budget, and stays within 3.6 accuracy points of the no-fairness baseline at every budget. Gains are smaller where PA cells are sparse. Control experiments show that neither preventing variance collapse, nor balanced per-group sampling, nor coreset selection without distillation reproduces these gains individually. Progress on fair dataset distillation therefore depends on the form of the per-class target, not the choice of point within it.

1 INTRODUCTION

Dataset distillation (DD) replaces a large training set with a small synthetic one. Formally, it compresses $\mathcal{T} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ into a compact $\mathcal{S} = \{(\tilde{\boldsymbol{x}}_j, \tilde{y}_j)\}_{j=1}^{M}$ with $M \ll N$, such that models trained on $\mathcal{S}$ generalise comparably to those trained on $\mathcal{T}$ (Wang et al., 2018; Zhao & Bilen, 2023; Zhao et al., 2021; Cazenavette et al., 2022; Yu et al., 2024), which makes it attractive for continual learning, architecture search and privacy-aware data release (Sucholutsky & Schonlau, 2021; Yin et al., 2023). As DD reaches face and identity datasets, a new question arises whether $\mathcal{S}$ preserve fairness across protected-attribute (PA) groups (Kim et al., 2019; Sagawa et al., 2020; Jung et al., 2022; Mehrabi et al., 2021; Buolamwini & Gebru, 2018)? It does not. Synthetic samples over-represent majority groups, and downstream classifiers inherit the disparity and often worsen it (Zhou et al., 2025).

Two families of PA-aware objectives have been proposed. They differ in how the matching target is chosen, not in what kind of object it is. FairDD (Zhou et al., 2025) came first. Its *synchronised matching* replaces the single class-conditional target of distribution matching (DM) (Zhao & Bilen, 2023) and dataset condensation (DC) (Zhao et al., 2021) with equally-weighted per-group centroids, so that at the optimum the synthetic feature mean is the arithmetic mean of those centroids, independent of group sizes. COBRA (Moslemi et al., 2026) works differently and argues that bias can become worse when groups are imbalanced and far apart from each other in feature space. COBRA first

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

finds a single *barycenter* of the per-group representations, placed so that no group is favoured for being larger. Synthetic data is then matched to that point. So FairDD averages one loss per group whereas COBRA first averages the group representations and then uses one loss for the whole class. Both improve substantially on vanilla DM and DC.

However, they share a fundamental limitation as both ultimately give the distillation loop a single vector per class. This is why both methods are *variance-blind*. The limitation comes from the form of the target itself, rather than from the particular way the target is chosen. Choosing the point more fairly does not make the point contain information about spread. As formalised in Proposition 1, any two synthetic configurations with the same per-group centroid receive the same point-target loss, even if one reproduces the real within-group variation and the other collapses into a tight cluster (Jing et al., 2022; Ansuini et al., 2019). We make it precise with a Bures–Wasserstein decomposition of the per-group transport cost (Theorem 1, building on (Gelbrich, 1990; Jiang et al., 2020)). The squared 2-Wasserstein distance between synthetic and real per-group distributions splits into a centroid term, which point-target objectives control, and a covariance term, which captures the mismatch in within-group spread and is not directly controlled by these methods. This second term scales with the within-group variance $\mathrm{Tr}(\Sigma_{P_a})$. It is a floor below which their residual error cannot fall (Gordaliza et al., 2019; Graf & Luschgy, 2000).

The limitation becomes particularly important on real face datasets, where images within the same group can still be quite diverse. We measure this within-group variation on CelebA (Liu et al., 2015), bFFHQ (Karras et al., 2019), and UTKFace (Zhang et al., 2017). In these datasets, $\mathrm{Tr}(\Sigma_{P_a})$ is roughly an order of magnitude larger than on the synthetic colour-bias benchmarks, Coloured-MNIST (Arjovsky et al., 2019) and CIFAR10-S (Krizhevsky & Hinton, 2009), that have been the focus of prior PA-aware DD work (Cui et al., 2024).

This matters because a point target cannot represent this within-group diversity, no matter how carefully that point is chosen. We see this in the behaviour of FairDD and COBRA. FairDD tends to collapse the synthetic features into a nearly point-like cluster, while COBRA overshoots, with variance ratios between 1.04 and 1.13. Yet simply increasing the number of synthetic images does not necessarily solve the problem. On CelebA, increasing IPC from 10 to 50 actually increases DEO_M from 9.37 to 14.08 for FairDD and from 10.00 to 15.32 for COBRA instead of decreasing. Giving the distillation process more points does not by itself make the target capable of representing the geometry of the real distribution. Our approach addresses this by replacing the point target with a geometry-preserving set. We find that this gives a better match to the real feature geometry across all of the face datasets we test. Whether this improvement also translates into better downstream fairness depends on how many real images are available in each PA cell at a given synthetic budget, so we treat the downstream gains as dataset and budget dependent.

Contributions.

- **We identify a limitation of point-based targets.** We show theoretically that matching a point can only capture the mean of each group, not its spread (Theorem 1, Proposition 1). We also show that $\mathrm{Tr}(\mathbf{\Sigma}_{P_a})$ can be used before distillation to tell us when this limitation is likely to matter. Although FairDD and COBRA have the same optimal target in mean space under COBRA's default barycenter, their resulting variance ratios can differ by a factor of 38.
- **We introduce FGD, which matches distributions rather than points.** Instead of representing each group with a single point, we use class-conditional *geometrically fair coresets* which are small sets of *real* images that capture the geometry of each group (Mirzasoleiman et al., 2020; Feldman, 2020; Sorscher et al., 2022; Har-Peled & Mazumdar, 2004). These real coresets serve only as matching targets. Distillation then optimises synthetic counterparts that mimic the spread of each group's coreset, which allows the objective to account for within-group variation. Smoothed slot allocation and periodic refresh make the method more robust to sparse PA cells and changes in the feature extractor (Appendices F.1, F.2).
- **We evaluate FGD against both methods.** At a fixed $\lambda{=}0.25$, FGD attains the lowest DEO_M in 7 of 9 dataset$\times$IPC cells, halving COBRA's CelebA gap and cutting FairDD's by 47–69%, at 1.5–3.6 points of accuracy against the no-fairness baseline. Three controls fail to reproduce these gains, justifying that the target and not the optimiser is the key bottleneck (Section 5).

2 Background and Related Work

Dataset distillation and its fair variants. Wang et al. (2018) introduced DD as bilevel optimisation. The main families are gradient matching (Zhao et al., 2021; Zhao & Bilen, 2021), distribution matching (DM) (Zhao & Bilen, 2023; Wang et al., 2022), trajectory matching (Cazenavette et al., 2022), soft-label methods (Sucholutsky & Schonlau, 2021) and compact synthetic-data parameterisations (Kim et al., 2022), reviewed in (Yu et al., 2024). We build on DM, the base objective of both PA-aware methods we compare against, FairDD (Zhou et al., 2025) and COBRA (Moslemi et al., 2026). Both agree with us that the target rather than the loss weighting is the lever. They differ from us in what replaces it, as Section 1 sets out. Coreset selection carries formal data-reduction guarantees (Feldman, 2020; Har-Peled & Mazumdar, 2004), extended to fairness first for clustering (Huang et al., 2019). Fair Wasserstein Coresets (FWC) (Xiong et al., 2024) minimise a *squared* 2-Wasserstein objective under a partition matroid that enforces fair group representation. Unlike a centroid, an FWC coreset consists of *actual data points* from the support of P.

Model-level fairness pipelines (Kamiran & Calders, 2012; Zhang et al., 2018; Hardt et al., 2016; Zemel et al., 2013; Dwork et al., 2012; Mehrabi et al., 2021) are complementary, since PA-aware distillation intervenes at the *data* level. Appendix L covers these and adjacent work on condensation bias (Cui et al., 2024) and long-tailed distillation (Zhao et al., 2025).

Group fairness metrics. We use the metrics from FairDD: the difference of equalised odds, in maximum and average forms,

$$\mathrm{DEO}_M = \max_{y\in\mathcal{Y}} \max_{a_i\neq a_j\in\mathcal{A}} \left|\hat{P}(\hat{Y}{=}y \mid Y{=}y, A{=}a_i) - \hat{P}(\hat{Y}{=}y \mid Y{=}y, A{=}a_j)\right|, \tag{1}$$

$$\mathrm{DEO}_A = \operatorname*{mean}_{y\in\mathcal{Y}} \max_{a_i\neq a_j\in\mathcal{A}} \left|\hat{P}(\hat{Y}{=}y \mid Y{=}y, A{=}a_i) - \hat{P}(\hat{Y}{=}y \mid Y{=}y, A{=}a_j)\right| \tag{2}$$

Lower is fairer. Wasserstein distances have been linked directly to enforcing these parity conditions downstream (Jiang et al., 2020), motivating a geometric treatment.

3 Why current fair distillation objectives fail: A geometric perspective

We now analyse the geometric mechanics of point-target matching, the form shared by both PA-aware objectives. FairDD is the running example through Section 3.2, because its failure is the most visible. The analysis is not specific to it. Section 3.3 shows that COBRA reduces to the same target, and treats the two of them to from there on.

3.1 Background: Fair dataset distillation

Given DD training data $\mathcal{T}$, distribution matching (DM) minimises $\sum_y \mathcal{D}\left(\mathbb{E}_{x\sim\mathcal{T}_y}[\phi(\boldsymbol{x};\boldsymbol{\theta})], \mathbb{E}_{\tilde{x}\sim\mathcal{S}_y}[\phi(\tilde{\boldsymbol{x}};\boldsymbol{\theta})]\right)$, where $\phi(\cdot;\boldsymbol{\theta})$ is a feature extractor and $\mathcal{D}$ a distance such as MSE or MAE. FairDD replaces this with a PA-decomposed objective.

$$\mathcal{L}_{FairDD}(\mathcal{S};\boldsymbol{\theta},\mathcal{T}) = \sum_{y\in\mathcal{Y}}\sum_{a\in\mathcal{A}} \mathcal{D}\Big(\mathbb{E}_{x\sim\mathcal{T}_y^a}[\phi(\boldsymbol{x};\boldsymbol{\theta})],\ \mathbb{E}_{\tilde{x}\sim\mathcal{S}_y}[\phi(\tilde{\boldsymbol{x}};\boldsymbol{\theta})]\Big) \tag{3}$$

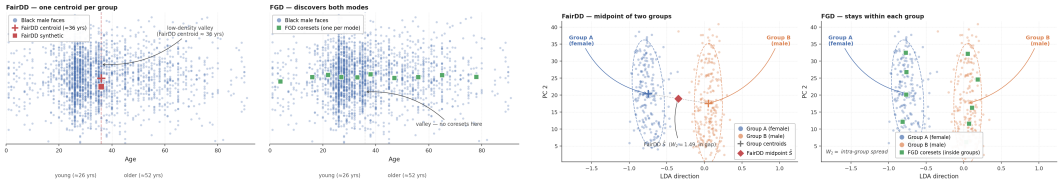
which Zhou et al. (2025) show converges (for MAE/MSE) to a synthetic distribution whose feature mean is the unweighted average of per-group feature means i.e., to the *arithmetic mean of group centroids*.

3.2 A point target is geometrically unfaithful

For $\mathcal{D}$ the squared Euclidean distance, the per-group term of Eq. (3) for class y and group a is

$$L_a(S_y;\boldsymbol{\theta}) \;=\; \left\|\mathbb{E}_{x\sim S_y}[\phi(\boldsymbol{x};\boldsymbol{\theta})] - \mu_a\right\|^2 \tag{4}$$

where $\mu_a = \mathbb{E}_{x\sim P_a}[\phi(\boldsymbol{x};\boldsymbol{\theta})]$. COBRA's objective has the same shape with μ_a replaced by a barycenter m_y. Either way the loss is a function only of the *first moment* of S_y, and our analysis builds on that one observation.



(a) Bimodal age in the *Black-male* subgroup of UTK-Face.

(b) Separated clusters within a CelebA class.

Figure 1: Two geometric failure modes of FairDD's centroid target. Crosses mark FairDD's centroid, squares mark FGD coresets. In both the centroid lies in a low-density region while the coresets sit on the data manifold.

Theorem 1 (Variance-blindness of mean-matching)**.** *Let S_y and P_a be distributions on $\mathbb{R}^d$ with finite second moments, means $\mathbb{E}[S_y]$ and μ_a, and covariances Σ_{S_y} and Σ_{P_a}. Then*

$$\mathcal{W}_2^2(S_y, P_a) \;\geq\; \underbrace{\big\|\mathbb{E}[S_y] - \mu_a\big\|^2}_{=L_a(S_y;\theta)} \;+\; \underbrace{\mathrm{Tr}\big(\Sigma_{S_y} + \Sigma_{P_a} - 2\big(\Sigma_{P_a}^{1/2}\Sigma_{S_y}\Sigma_{P_a}^{1/2}\big)^{1/2}\big)}_{\mathcal{B}(\Sigma_{S_y},\Sigma_{P_a})\geq 0}. \tag{5}$$

The first term is exactly what a point-target objective minimises, whether the point is FairDD's per-group centroid or COBRA*'s barycenter. The second term measures the remaining difference in covariance, through the squared Bures–Wasserstein distance (Gelbrich, 1990; Bhatia et al., 2019). It is non-negative and vanishes if and only if $\Sigma_{S_y} = \Sigma_{P_a}$. In other words, matching the mean does not tell us whether the synthetic group has the same spread as the real group. Also, (i) driving $L_a \to 0$ does not by itself imply $\mathcal{W}_2^2 \to 0$, and (ii) when $\Sigma_{S_y} = 0$, so that $S_y = \delta_\nu$ is a Dirac mass at ν, the bound is tight and becomes*

$$\mathcal{W}_2^2(\delta_\nu, P_a) = \|\nu - \mu_a\|^2 + \mathrm{Tr}(\Sigma_{P_a}).$$

Inequality (5) is the Gelbrich bound (Gelbrich, 1990; Bhatia et al., 2019; Peyré & Cuturi, 2019). Part (ii) follows because $S_y = \delta_\nu$ puts all synthetic mass at a single point ν, leaving only one possible coupling with P_a. Appendix E gives both steps.

Theorem 1 makes the limitation clear. First, FairDD's loss cannot fully control $\mathcal{W}_2^2$ to a group because matching the mean leaves the covariance gap $\mathcal{B}$ untouched. Second, in the extreme case where the synthetic distribution collapses to a Dirac, this gap is exactly $\mathrm{Tr}(\Sigma_{P_a})$. Whether optimisation actually reaches this worst case is an empirical question, which we examine next.

Observation 1 (Empirical near-Dirac collapse)**.** *On all three face datasets, FairDD's optimised synthetic feature distribution has variance ratio $\mathrm{Tr}(\Sigma_{S_y})/\,\mathrm{Tr}(\Sigma_{T_y}) < 0.05$ (Table 2, "Var→ 1"), so it lies in the near-Dirac regime where Theorem 1(ii) gives $\mathcal{W}_2^2(S_y, P_a) \approx \|\mathbb{E}[S_y] - \mu_a\|^2 + \mathrm{Tr}(\Sigma_{P_a})$. The first term is fixed by where FairDD's target sits, at the average of the group centroids, so the part that grows with within-group spread is exactly $\mathrm{Tr}(\Sigma_{P_a})$.* COBRA *instead overshoots on the same unconstrained term, with ratios of 1.04–1.13. Its covariance is not collapsed but is not matched either, so $\mathcal{B}$ does not vanish.*

Scope of Theorem 1. The theorem applies Gelbrich's inequality (Gelbrich, 1990) rather than establishing a new transport result, and it is a lower bound, so its role is diagnostic. We claim the FairDD specific diagnostic and its empirical validation but not that the Dirac collapse is *forced* by the loss. It is what gradient descent on synthetic pixels with a feature-mean target empirically produces, mirroring dimensional collapse (Jing et al., 2022) and neural collapse (Papyan et al., 2020). COBRA's optimiser lands elsewhere under the same freedom, so we treat the collapse as a property of FairDD's optimisation, and not of the bound itself.

Figure 1 makes this concrete. It shows a bimodal subgroup and two separated within-class clusters, and in both the centroid sits in a low-density region between the modes rather than on any of them. A single point cannot represent a multimodal cell.

3.3 THE GENERAL STATEMENT: EVERY POINT-TARGET OBJECTIVE INHERITS THE BLIND SPOT

The argument so far runs through FairDD, but the result does not depend on that choice, so we state it in the form that covers the current literature.

Proposition 1 (Any single-point target is variance-blind)**.** *Fix a class y and let a distillation objective take the form*

$$\mathcal{L}_y(\mathcal{S}; \theta, \mathcal{T}) \;=\; D\big(m_y,\ \mathbb{E}[S_y]\big), \qquad m_y \in \mathbb{R}^d, \tag{6}$$

for any D and any target m_y that does not depend on Σ_{S_y}. Then $\mathcal{L}_y$ is independent of Σ_{S_y}, so for every group a the Bures–Wasserstein term $\mathcal{B}(\Sigma_{S_y}, \Sigma_{P_a^y})$ of Eq. (5) *is absent from the objective, and no choice of m_y constrains the second moment of the synthetic set.*

Proof. $\mathcal{L}_y$ depends on S_y only through $\mathbb{E}[S_y]$, so two synthetic sets with a common mean and different covariances receive identical loss. It is a proposition rather than a corollary because it does not follow from Theorem 1; the two are complementary, the theorem supplying the term Eq. (6) omits. The content is what it rules out. Relocating m_y changes *where* the centroid term is minimised, not *whether* the covariance term is represented, so the blind spot follows from the target being a point. Both methods are instances of this, FairDD with m_y the arithmetic mean of group centroids, COBRA with m_y the uniform barycenter.

FairDD reaches m_y implicitly. With synchronised matching, the MAE/MSE objective converges to the arithmetic mean of the group centroids (Zhou et al., 2025). COBRA instead defines m_y explicitly as the uniform barycenter $\arg\min_m \sum_a d(\mu_a^y, m)$, whose solution for its default squared Q-norm is $\frac{1}{|\mathcal{A}|}\sum_a \mu_a^y$ (Moslemi et al., 2026). Thus, both methods arrive at the same point, and Proposition 1 applies to both.

Yet their synthetic distributions look very different. On CelebA at IPC$=10$, the variance ratios are 0.029 and 1.094, a $38\times$ difference. The two methods agree on the first moment, but the objective says nothing about the second. In that sense, the different outcomes are exactly what we should expect when the spread is left unconstrained. COBRA's own ablation over different choices of d also finds no single choice that works best across benchmarks.

The Proposition 1 followed by both methods says that the covariance is not constrained by the objective, it is left free to be small or big. Actually the two methods end up on very different sides of this gap. FairDD collapses to a variance ratio of 0.005–0.029 (Observation 1), approaching the near-Dirac setting where Theorem 1(ii) becomes tight. COBRA, in contrast, produces larger spreads than the real data, with ratios of 1.04–1.13 (Table 2). The spread is therefore free to collapse or overshoot, rather than being matched. Section 5 shows this effect in the measured geometry.

FGD changes the target, not the optimisation procedure. Instead of matching a single point for each class, it uses a per-group coreset $C_a^y \subset \mathcal{T}_a^y$ of real images whose empirical covariance is $\Sigma_{C_a^y} \approx \Sigma_{P_a^y}$. The matching objective can therefore see each group's spread directly, so the covariance term $\mathcal{B}(\Sigma_{S_y}, \Sigma_{P_a^y})$ is no longer invisible. This gives FGD a direct way to preserve second-order structure. Empirically, its variance ratio moves toward 1 from below as the budget grows, from 0.86 to 0.97, rather than overshooting. In other words, the second moment is now something the target can actually match, rather than something the objective simply leaves free.

4 FGD: FAIR GEOMETRIC DISTILLATION

By Proposition 1, no repair that stays inside Eq. (6) can help. The target itself must change, so we keep the synchronised structure and replace the point with per-cell sets of real images chosen to approximate each cell's full distribution rather than its mean, giving *Fair Geometric Distillation* (FGD).

4.1 PER-CELL CORESET TARGETS

Notation and slot allocation. Given training data $\mathcal{T} = \{(\boldsymbol{x}_i, y_i, a_i)\}_{i=1}^N$, let $\mathcal{T}_{y,a} = \{i : y_i = y, a_i = a\}$ denote the (y, a) cell, for which we maintain a frozen real pool $\mathcal{R}_{y,a} \subseteq \mathcal{T}_{y,a}$ of up to $P{=}512$ randomly sampled images, fixed at initialisation and optionally refreshed (Section 4.2). Each cell receives a synthetic budget $M_{y,a} = \lfloor m \cdot q_y(a) \cdot \pi(y) \rfloor$ with $q_y(a) = 1/|\mathcal{A}|$ and $\pi(y)$ the desired

class prior, so every PA group of class y gets the same number of slots regardless of its frequency in $\mathcal{T}_y$. Section 4.2 generalises this to a smoothed allocation for many-group regimes.

Per-cell coreset construction. Within each cell, we want $M_{y,a}$ real images that approximate the cell distribution in feature space. Following the standard k-median formulation for coreset selection (Har-Peled & Mazumdar, 2004), we solve at each training step:

$$\hat{\mathcal{T}}_{y,a}(\theta) \;=\; \underset{S\subseteq\mathcal{R}_{y,a},\,|S|=M_{y,a}}{\arg\min} \sum_{r\in\mathcal{R}_{y,a}} \min_{s\in S} \big\|\phi(r;\theta)-\phi(s;\theta)\big\|_2, \tag{7}$$

where $\phi(\cdot;\theta)$ is the feature extractor used during distillation. The k-median solution preserves within-cell geometry a single centroid cannot, since the medoid set covers the cell's distribution with $M_{y,a}$ representatives. We solve Eq. (7) approximately with the standard Lloyd-style update (Lloyd, 1982) over the pool ($\sim$ 5ms per cell on an A100). Alternative per-cell objectives (Sinkhorn (Cuturi, 2013), unbalanced OT (Chizat et al., 2018), Chamfer (Fan et al., 2017)) are in Appendix O.1, where k-median is the most stable across datasets.

4.2 The FGD training loop

FGD applies synchronised matching at the per-cell level, with a one-sided nearest-neighbour cost against the per-cell coreset replacing FairDD's mean-matching against the per-cell centroid. The loss has a per-cell matching term and a per-cell class-discriminability regulariser:

$$\begin{aligned}\mathcal{L}_{\text{FGD}}(\mathcal{S};\theta,\mathcal{T}) = \frac{1}{|\mathcal{Y}||\mathcal{A}|}\sum_{y\in\mathcal{Y}}\sum_{a\in\mathcal{A}}\Bigg[\frac{1}{|\mathcal{S}_{y,a}|}\sum_{\tilde{x}_i\in\mathcal{S}_{y,a}}\min_{r\in\hat{\mathcal{T}}_{y,a}}\big\|\phi(\tilde{x}_i;\theta)-\phi(r;\theta)\big\|^2 \\ +\;\lambda\,\big\|\mathbb{E}[\phi(\hat{\mathcal{T}}_{y,a};\theta)]-\mathbb{E}[\phi(\mathcal{S}_{y,a};\theta)]\big\|_2^2\Bigg],\end{aligned} \tag{8}$$

where $\mathcal{S}_{y,a}$ is the subset of synthetic samples assigned to cell (y,a) by the slot allocation, $\hat{\mathcal{T}}_{y,a}$ is the k-median coreset selected by Eq. (7) at the current step, and $\lambda \geq 0$ controls the regulariser.

The matching term is a one-sided NN loss. Each synthetic point in cell (y,a) is pulled toward its nearest medoid in $\hat{\mathcal{T}}_{y,a}$. This is the geometric improvement over the point-target family as a whole. Any objective of the form of Eq. (6) produces a gradient that pulls every synthetic point toward one vector, whether that vector is FairDD's per-group centroid or COBRA's barycenter. Ours pulls each point toward the nearest representative of its cell, so the synthetic set spreads to cover the cell's modes instead of concentrating on a single location. The regulariser, evaluated per cell rather than per class, anchors the synthetic cell mean to the coreset mean, preventing drift along class-discriminative directions when $M_{y,a}$ is small.

Frozen pool with per-step k-median. Equation (7) needs θ to score candidates, but DM-style distillation re-samples θ at every step (Zhao & Bilen, 2023), and re-solving k-median over a full CelebA-scale cell each step is prohibitive. We instead solve over the frozen pool $\mathcal{R}_{y,a}$, as Algorithm 1 sets out in full (Figure 2 in Appendix C gives a diagram of the loop).

Practical extensions. Three extensions handle complex datasets: **smoothed slot allocation** to stabilise k-median in sparse demographic cells, **periodic pool refresh** to mitigate feature drift over long runs, and a **compute analysis** showing end-to-end wall-clock overhead is only $\sim 1.5\times$ FairDD's. Specifications and ablations are in Appendix F.

5 Experiments and Results

Six questions organise this section. Does the centroid mismatch predicted by Theorem 1 appear in the feature space where FairDD optimises (Q1)? Does FGD's coreset target reduce DEO_M and DEO_A, and at what price (Q2–Q3)? Is the coreset necessary (Q4)? How does the method behave with four groups (Q5) and under other backbones (Q6)?

Algorithm 1 FGD: Fair Geometric Distillation

Require: Training data $\mathcal{T}$, IPC budget m, regulariser λ, pool size P, refresh interval K
Ensure: Distilled dataset $\mathcal{S}$

// Initialise per-cell slot budget, frozen pool, synthetic set
for each $(y, a) \in \mathcal{Y} \times \mathcal{A}$ **do**
  $M_{y,a} \leftarrow \lfloor m \cdot q_y(a) \cdot \pi(y) \rfloor$
  $\mathcal{R}_{y,a} \leftarrow$ random sample of $\min(P, |\mathcal{T}_{y,a}|)$ images from $\mathcal{T}_{y,a}$
  $\mathcal{S}_{y,a} \leftarrow M_{y,a}$ random images from $\mathcal{R}_{y,a}$
end for
for $t = 1, \ldots, T$ **do**
  if $K > 0$ and $t \bmod K = 0$ **then**
    Refresh $\mathcal{R}_{y,a}$ by re-sampling from $\mathcal{T}_{y,a}$ for all (y, a)
  end if
  Sample feature extractor θ from random distribution
  for each (y, a) **do**
    Embed pool: $\Phi_{y,a} \leftarrow \phi(\mathcal{R}_{y,a}; \theta)$
    Solve k-median (Eq. 7) over $\Phi_{y,a}$ to obtain medoids $\hat{\mathcal{T}}_{y,a}$
  end for
  Compute $\mathcal{L}_{\text{FGD}}$ via Eq. (8)
  Update $\mathcal{S}$ by gradient step on $\nabla_{\mathcal{S}} \mathcal{L}_{\text{FGD}}$
end for
return $\mathcal{S}$

5.1 SETUP

Datasets, methods and protocol. We use CelebA (162,770 faces, target=Smiling, PA=Male/Female), bFFHQ (19,200 faces, 0.5% conflict, target=age, PA=gender) and UTKFace (20,813 faces, target=age, PA=race, four groups), and take all geometric measurements in the ConvNet feature space (depth 3, 8192-d) following (Zhou et al., 2025). At matched IPC we compare DM (Zhao & Bilen, 2023) as the no-fairness baseline, FairDD (Zhou et al., 2025), COBRA (Moslemi et al., 2026) and FGD, plus three controls that isolate the source of FGD's gain (Appendix M): FairDD + Variance Regularisation, DM + balanced sampling and Coreset-only. All methods receive the same IPC budget and hyperparameters, and we fix λ=0.25 throughout our experiments

Metrics. We test the performance of these methods using 3 metrics. Test accuracy (↑) of a ConvNet trained on $\mathcal{S}$, averaged over 5 random seeds, comes with DEO_M (↓) and DEO_A (↓). The geometric side is led by the per-group sliced $\bar{W}_2^2$ (Bonneel et al., 2015) (↓, PCA-10), the quantity Theorem 1 bounds. Alongside it we report MMD (Gretton et al., 2012; Kynkäänniemi et al., 2019) against the *pooled* real distribution, and the variance ratio $\text{Tr}(\text{Cov}(\mathcal{S}))/\,\text{Tr}(\text{Cov}(\mathcal{T}))$ in PCA-10 space, whose target is 1. Full definitions, the λ sweep and the control specifications are in Appendices D, I and M.

Point targets drift off-manifold across all six datasets. We measure the geometric distance from each method's distilled synthetic set to the real per-group distribution, comparing FairDD against FGD at the same budget. Table 1 confirms Theorem 1 together with Observation 1. On the face datasets FairDD's distilled set sits 5.8 to 6.1× farther from the real distribution than FGD's in per-group $\bar{W}_2^2$, with matching gaps in W_1 and MMD. The *absolute* gap is largest on the high-variance face datasets ($\bar{W}_2^2$ falls by 32–39 units) and smallest on the colour-bias ones (5–12), tracking the $\text{Tr}(\Sigma_{P_a})$ ranking exactly as predicted. Per-group analysis on CelebA (Table 6, Appendix G) confirms the gain is shared across the female and male subgroups, not concentrated in one.

All methods share the same network, optimizer, and budget over $N = 5$ random seeds (iterations = 20,000). The only differences are the loss function and matching targets. Results appear in Table 2.

Fairness improvements. On CelebA, FGD at λ=0.25 drops DEO_M from 14.08 to 5.31, a 62% reduction against FairDD, on bFFHQ it falls from 19.76 to 17.44. The geometric columns validate Theorem 1. FairDD collapses to a near-Dirac, with variance ratio 0.028 on CelebA and 0.021 on bFFHQ, while FGD preserves the real spread above 0.97 and per-group $\bar{W}_2^2$ falls from 45.64 to 1.15 and from 38.47 to 1.02.

At the fixed λ=0.25 default FGD attains the lowest DEO_M in 7 of the 9 dataset × IPC cells of Table 2, against 1 each for FairDD and COBRA. On CelebA the margin over COBRA is at least 2×

Table 1: Geometric faithfulness on six datasets at IPC=10 in FairDD's ConvNet feature space (mean over 5 seeds). $\mathrm{Tr}(\Sigma_{P_a})$ is the within-group feature variance averaged across PA groups, computable before any distillation. Per-group $\bar{W}_2^2$ is the metric bounded by Theorem 1, while W_1 and MMD report distance to the *pooled* real distribution, and the two disagree on the face datasets but not on the low-$\mathrm{Tr}(\Sigma_{P_a})$ synthetic ones. COBRA produces one target per class, so its samples are assigned to PA cells by nearest class-conditional real centroid. **Bold**: best per-group $\bar{W}_2^2$ per row.

	$\mathrm{Tr}(\Sigma_{P_a})$	Per-group $\bar{W}_2^2\downarrow$			$W_1\downarrow$			MMD$\downarrow$		
Dataset		FairDD	FGD	COBRA	FairDD	FGD	COBRA	FairDD	FGD	COBRA
Real face datasets (high $\mathrm{Tr}(\Sigma_{P_a})$)										
CelebA	1388	46.90	**8.07**	19.94	4.91	2.19	1.28	0.1517	0.0048	0.0041
bFFHQ	1420	38.19	**6.26**	17.75	4.67	2.08	1.20	0.1502	0.0026	0.0063
UTKFace	1533	40.52	**6.64**	41.00	4.97	1.63	1.08	0.1544	0.0030	0.0038
Synthetic colour-bias datasets (low $\mathrm{Tr}(\Sigma_{P_a})$)										
CIFAR10-S	272	11.99	**0.25**	4.84	2.67	0.39	0.40	0.165	0.000	0.0015
C-MNIST (FG)	105	6.12	**1.26**	4.41	1.18	0.27	0.36	0.072	0.000	0.0024
C-MNIST (BG)	103	5.85	**1.19**	3.94	1.04	0.23	0.40	0.066	0.000	0.0029

Table 2: Results across IPC for CelebA, bFFHQ (BR=0.995), and UTKFace at iter=20,000, means over 5 seeds. $\bar{W}_2^2$ is the per-group sliced W_2^2 of Theorem 1 averaged over (y, a) cells, and Var is the synthetic-to-real PCA variance ratio (closer to 1 is better). COBRA carries no PA labels, so its samples are assigned to cells by nearest class-conditional real centroid. **Bold**: best DEO$_M$ per IPC block. The λ sweep is in Appendix I.

		CelebA						**bFFHQ (BR=0.995)**						**UTKFace** ($\lvert\mathcal{A}\rvert$=4)					
		Performance			**Geometric**			**Performance**			**Geometric**			**Performance**			**Geometric**		
Method	IPC	DEO$_M\downarrow$	DEO$_A\downarrow$	Acc$\uparrow$	$\bar{W}_2^2\downarrow$	MMD$\downarrow$	Var$\to$1	DEO$_M\downarrow$	DEO$_A\downarrow$	Acc$\uparrow$	$\bar{W}_2^2\downarrow$	MMD$\downarrow$	Var$\to$1	DEO$_M\downarrow$	DEO$_A\downarrow$	Acc$\uparrow$	$\bar{W}_2^2\downarrow$	MMD$\downarrow$	Var$\to$1
DM	10	30.01	28.85	60.97	42.15	0.0825	0.125	44.32	34.76	64.90	35.42	0.0754	0.142	22.29	18.64	61.24	43.59	0.1556	0.002
	30	34.45	32.67	62.45	41.80	0.0840	0.142	53.68	42.92	65.02	34.90	0.0782	0.165	22.87	20.46	63.48	43.59	0.1569	0.002
	50	40.26	38.81	63.57	40.95	0.0815	0.186	60.88	50.56	65.28	34.12	0.0765	0.198	22.50	20.31	64.97	43.60	0.1590	0.002
FairDD	10	9.37	5.37	64.37	46.90	0.1517	0.029	15.84	10.68	65.46	38.19	0.1502	0.024	**18.37**	14.44	61.98	40.52	0.1544	0.005
	30	11.81	8.02	66.10	46.96	0.1553	0.026	17.96	11.44	68.46	38.58	0.1547	0.018	19.06	16.56	63.98	40.43	0.1558	0.005
	50	14.08	9.87	68.50	45.64	0.1581	0.028	19.76	14.96	69.00	38.47	0.1550	0.021	18.99	16.43	65.19	40.41	0.1579	0.005
COBRA	10	10.00	6.36	64.94	19.94	0.0041	1.094	20.24	11.40	68.30	17.75	0.0063	1.134	21.00	17.43	49.95	41.00	0.0038	1.076
	30	10.65	5.98	67.04	9.91	0.0017	1.081	16.56	10.56	70.56	7.69	0.0034	1.082	21.80	16.93	50.97	21.46	0.0012	1.062
	50	15.32	10.65	68.48	6.00	0.0013	1.053	**14.96**	10.00	71.40	6.82	0.0030	1.052	22.20	17.24	53.80	16.00	0.0008	1.039
FGD (λ=0.25)	10	**5.01**	3.18	59.49	8.07	0.0048	0.858	**13.18**	8.60	65.82	6.26	0.0026	0.857	21.82	18.18	60.07	6.64	0.0030	0.896
	30	**3.65**	2.64	59.01	2.15	0.0006	0.941	**16.01**	10.48	68.28	1.82	0.0001	0.948	**18.99**	15.58	59.99	2.02	0.0000	0.953
	50	**5.31**	3.01	59.99	1.15	0.0001	0.972	17.44	11.99	67.01	1.02	0.0001	0.972	**14.98**	11.74	58.96	1.01	0.0000	0.972

at every budget (2.0, 2.9 and 2.9$\times$). Neither of the 2 target improves with budget, their floor being IPC-invariant, while FGD's coreset error shrinks. COBRA takes bFFHQ at IPC=50 and FairDD UTKFace at IPC=10, and we analyse both below.

Accuracy trade-off and λ sensitivity. FGD has lower accuracy than both COBRA and FairDD on CelebA at IPC=10 (59.49% against 64.94% and 64.37%). The FairDD + Variance Regularisation control locates where the cost comes from. Forcing FairDD to preserve within-group variance, with no coresets and no change of target, drops its own accuracy to $57.94 \pm 2.71\%$ (Table 8). The penalty is the price of spread it is trying to preserve. On bFFHQ at IPC=50 it is smaller, at 67.01%, which is 1.99 points below FairDD and above DM. The λ sweep shows the default is not uniformly optimal, since λ=1 reaches DEO$_M$=4.78 against 5.31 on CelebA at IPC=50, while over regularising breaks coreset alignment. We report the fixed default because selecting λ on test fairness would not be a usable protocol.

IPC scaling and MMD diagnostic. Neither of the point targets convert budget into fairness, for the reason given in Section 1 and the extra copies of a centroid carry no new geometric information, whereas FGD's coreset error shrinks as the budget grows. MMD makes the mechanism explicit. FairDD's stays saturated at $\sim$0.155, evidence of kernel saturation under point-mass collapse, while FGD's drops by orders of magnitude with IPC scaling, reaching 0.0001 on CelebA at IPC=50.

COBRA is a sharper comparison for our claim. It also replaces the matching target rather than reweighting the loss, so it is where a set-valued target shows its value most clearly. FGD is the fairer of the two on both metrics, leading in 7 of 9 cells on DEO$_M$ and DEO$_A$, roughly halving COBRA's DEO$_M$ at every CelebA budget while holding the per-group $\bar{W}_2^2$ that Theorem 1 bounds 2.5 to 19$\times$

below COBRA's on all six datasets. The ordering survives the cross-architecture protocol results in the appendix, so the advantage is not limited to only one feature extractor.

The two buy fairness at different prices. COBRA retains more accuracy, leading FGD by 5.5 to 8.5 points on CelebA and 2.3 to 4.4 on bFFHQ, and the FairDD + Variance Regularisation control shows why. That cost is what any method pays to preserve within-group spread under a small budget. COBRA's single DEO_M win in bFFHQ at IPC$=50$ is where our sparse-cell condition predicts FGD would be weaker, so it confirms our hypothesis. Together, the two methods make the difference clear. COBRA shows that moving the target point vector can only take us so far, while FGD shows what becomes possible once the target captures more than just a single vector.

Which distance should we trust? COBRA has the lowest *pooled* W_1 on every face dataset (Table 1). Still, it trails FairDD on DEO_M at IPC$=10$ on all three. The per-group $\bar{W}_2^2$ bounded by Theorem 1 flips this ranking, and FGD leads on all six datasets. So the per-group view is the one that matches the fairness question. We read this geometry as the mechanism and keep DEO_M and DEO_A as the final test, since per-group distance alone is a loose predictor of fairness (Appendix K).

A geometry-preserving target is only the right explanation if simpler expressions or methods fail. We test three across all three face datasets and budgets (5 seeds, full table in Appendix M), each isolating one candidate explanation that does not require leaving the point-target family.

None of the three reproduces FGD's gain. FairDD + Variance Regularisation raises the CelebA variance ratio from 0.028 to 0.876, so it does prevent collapse, yet on CelebA recovers only 11–28% of the DEO_M gap to FGD ($14.08 \rightarrow 13.11 \pm 0.78$ at IPC$=50$). Since it only preserves the *trace*, it tests whether avoiding collapse is enough, rather than whether the covariance itself needs to be matched. DM + balanced sampling equalises per-group sampling without touching the target. It trails FairDD on CelebA and UTKFace (10.14 ± 2.74 vs 9.37 on CelebA at IPC$=10$), and its two small bFFHQ wins lie within seed noise, so balanced allocation alone is not the mechanism. Coreset-only runs per-cell k-median with no distillation. On CelebA and bFFHQ it is worse than both FairDD and FGD on fairness and accuracy alike (22.37 ± 3.49 on CelebA at IPC$=50$). On UTKFace its DEO_M is competitive but noisy (up to ± 4.91) and its accuracy is 7–8 points below FairDD's. The distillation loop does work beyond selecting a representative subset. Definitions, the settings where a control falls within seed noise, and the relationship to FWC (Xiong et al., 2024) are in Appendices M and H.4.

The multi-group setting and transfer across architectures. UTKFace ($|\mathcal{A}|=4$) is the setting the smoothed slot allocation was built for, and the honest summary is that downstream gains there are modest and budget-dependent. At IPC$=50$ FGD reduces DEO_M from 18.99 to 14.98, but at IPC$=10$ it is *worse* than FairDD (21.82 vs 18.37). The cause is a budget condition. With four groups at IPC$=10$ each $(class, group)$ cell receives only 2–3 slots and the per-cell k-median is underdetermined, and the same sparsity explains bFFHQ, where at BR$=0.995$ the DEO_M reduction is only $\sim 12\%$. Downstream fairness is thus dataset-dependent even though the geometric metrics FGD optimises improve on all three face datasets.

Repeating the pipeline under AlexNet, VGG-11 and ResNet-18, distilling *and* evaluating with each, FGD attains the lowest DEO_M in 24 of 27 settings and the lowest DEO_A in 23 (Appendix N). All three DEO_M exceptions are COBRA on bFFHQ at IPC$=50$, the same sparse-cell regime.

6 Discussion and Limitations

Matching one point vector per group works when groups look clearly different. It breaks down on real faces, where each group varies a lot. Our argument targets a whole *kind* of objective. It covers any future method that only moves the point, and it leaves aside objectives that already match the spread of the data. FGD solves the problem by matching a set of points instead. It gives the lowest per-group $\bar{W}_2^2$ on all six datasets and the lowest DEO_M in 7 of 9 ConvNet cells. The two remaining cells have the same cause. On UTKFace at IPC$=10$, and for the small bFFHQ minority, each $(class, group)$ cell gets only a few slots, too few to pick a good set of targets. Fairness also has a cost. On CelebA, FGD loses 1.5 to 3.6 accuracy points against the baseline without fairness. Our variance control fixes the total spread of the synthetic data and leaves its shape free, so a stricter test that also matches the shape is still to be done. Limitations are in Appendix J.

AI USE STATEMENT

In this work, we used generative AI tools to assist in manuscript preparation, including revising prose, condensing the paper to the page limit, checking the internal consistency of reported numbers against our result tables, and assisting in writing code. We additionally used these tools to help explore related works and assist with bibliographic searches. However, we independently verified all cited works and reference metadata. All other aspects of this research like the proposal of the hypothesis, the statement and derivation of theoretical results, experimental design, and the generation, analysis, and interpretation of results, were conducted entirely by the authors. We have reviewed all AI-assisted work and take full responsibility for the final content of this paper, including its text, claims, and artifacts.

ETHICS STATEMENT

This work studies fairness in dataset distillation, and all six datasets we use are publicly available and widely used for this purpose. Three of them, CelebA (Liu et al., 2015), bFFHQ (Karras et al., 2019) and UTKFace (Zhang et al., 2017), consist of photographs of real people, and we use them under their stated research-only terms without collecting, redistributing or re-annotating any images.

The protected-attribute labels these datasets carry deserve explicit comment. In every case they are third-party annotations rather than self-identification by the people depicted, and they encode gender as binary and race as a small set of discrete categories. Neither choice reflects how the underlying attributes are actually distributed, and both were fixed by the dataset creators rather than by us. We adopt these labels because doing so is what allows direct comparison against prior fair-distillation work (Zhou et al., 2025; Moslemi et al., 2026), not because we regard them as adequate. Fairness measured against a coarse or externally imposed attribute can overstate how equitable a method is for the people it ends up affecting, and our DEO_M and DEO_A numbers should be read with that limit in mind.

REPRODUCIBILITY STATEMENT

Datasets, splits and per-cell statistics are given in Appendix B, and implementation details, hyperparameters and compute requirements in Appendix C. FGD is fully specified by Algorithm 1 together with Equations (7) and (8); the single regularisation setting $\lambda{=}0.25$ is used for every headline number, with the sweep reported in Appendix I. Formal definitions of every geometric metric, including the projection-robustness variants, appear in Appendix D, and the proof of Theorem 1 with the verification of Observation 1 in Appendix E. All reported numbers are means over five random seeds, with standard deviations given for the control conditions in Appendix M.

REFERENCES

A. Ansuini, A. Laio, J. H. Macke, and D. Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

M.-F. Balcan, T. Dick, Y. Liang, W. Mou, and H. Zhang. Differentially private clustering in high-dimensional euclidean spaces. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency (FAT*)*, 2018.

G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J.-Y. Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Justin Cui, Ruochen Wang, Yuanhao Xiong, and Cho-Jui Hsieh. Mitigating bias in dataset distillation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. arXiv:2406.06609.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 214–226, 2012.

H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3D object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 605–613, 2017.

D. Feldman. Introduction to core-sets: an updated survey. *arXiv preprint arXiv:2011.09384*, 2020.

M. Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

P. Gordaliza, E. Del Barrio, G. Fabrice, and J.-M. Loubes. Obtaining fairness using optimal transport theory. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer, 2000.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pp. 291–300, 2004.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

L. Huang, S. Jiang, and N. Vishnoi. Coresets for clustering with fairness constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

L. Jing, P. Vincent, Y. LeCun, and Y. Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

S. Jung, S. Chun, and T. Moon. Learning fair classifiers with partially annotated group labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10348–10357, 2022.

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9012–9020, 2019.

Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2): 129–137, 1982.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

B. Mirzasoleiman, J. Bilmes, and J. Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

M. H. Moslemi, N. Hosseini Dashtbayaz, Z. Mei, B. Ghaddar, and B. Wang. Fair dataset distillation via cross-group barycenter alignment. *arXiv preprint arXiv:2605.00185*, 2026.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, and T. Goldstein. The intrinsic dimension of images and its impact on learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

B. Sorscher, R. Geirhos, S. Shekhar, S. Ganguli, and A. Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

I. Sucholutsky and M. Schonlau. Soft-label dataset distillation and text dataset distillation. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2021.

Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. CAFE: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Z. Xiong, N. Dalmasso, S. Sharma, F. Lecue, D. Magazzeni, V. K. Potluru, T. Balch, and M. Veloso. Fair Wasserstein coresets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Z. Yin, E. Xing, and Z. Shen. Squeeze, recover and relabel: Dataset condensation at ImageNet scale from a new perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 73582–73603, 2023.

Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):150–170, 2024. doi: 10.1109/TPAMI.2023.3323376.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2018.

Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4352–4360, 2017.

B. Zhao, K. R. Mopuri, and H. Bilen. Dataset condensation with gradient matching. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12674–12685. PMLR, 2021.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6514–6523, 2023.

Zhenghao Zhao, Haoxuan Wang, Yuzhang Shang, Kai Wang, and Yan Yan. Distilling long-tailed datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2408.14506.

Q. Zhou, S. Fang, S. He, W. Meng, and J. Chen. FairDD: Fair dataset distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

A Appendix summary

The appendix supplements the main paper with: dataset and reproducibility details (Sections B, C), formal definitions of the geometric metrics (Section D), proofs of Theorem 1 and Observation 1 (Section E), practical extensions to the FGD framework including compute requirements (Section F), extended empirical results (Section G), formal definitions of alternative per-cell objectives (Section H), the λ sensitivity sweep (Section I), the full necessity-control results (Section M), cross-architecture transfer (Section N), and ablation studies justifying our design choices (Section O).

B Datasets and statistics

We evaluate on six datasets covering two distinct fair-distillation regimes: synthetic colour-bias benchmarks where bias is constructed explicitly, and real-world face datasets where bias arises from photographic and demographic structure.

Synthetic benchmarks. Coloured-MNIST (Arjovsky et al., 2019) is a variant of MNIST where the foreground (FG) or background (BG) is artificially coloured to introduce a protected attribute. We evaluate both variants. **CIFAR10-S** introduces a PA by applying a grayscale filter to a portion of CIFAR10 (Krizhevsky & Hinton, 2009) training images, creating a strong spurious correlation between the target class and the colour space.

Real-world facial datasets. CelebA (Liu et al., 2015) is a widely used fairness benchmark; we treat *Smiling* as the target attribute (TA) and *Gender* as the PA. **UTKFace** (Zhang et al., 2017) provides a multi-group fairness challenge with $|\mathcal{A}| = 4$ (Race), making it a stress test for the smoothed slot allocation we describe in Section 4.2. **bFFHQ** (Karras et al., 2019) provides an extreme-imbalance scenario (biased ratio 0.995), testing the limits of coreset selection when the minority group is exceptionally sparse.

Table 3: Statistics for the six datasets evaluated in this work. BR denotes the biased ratio (the proportion of training samples in the majority group).

Dataset	Target Attribute (TA)	Protected Attribute (PA)	$\lvert\mathcal{Y}\rvert$	$\lvert\mathcal{A}\rvert$	Train Size	BR (Train)
Coloured-MNIST (FG)	Digit (0–9)	Foreground Colour	10	10	60,000	0.90
Coloured-MNIST (BG)	Digit (0–9)	Background Colour	10	10	60,000	0.90
CIFAR10-S	Object Category	Grayscale or Colour	10	2	50,000	0.90
UTKFace	Age (Binarized)	Race	2	4	20,813	$\sim$0.53
CelebA	Smiling	Gender	2	2	162,770	$\sim$0.62
bFFHQ	Age	Gender	2	2	19,200	0.995

C Implementation details and reproducibility

Codebase. The datasets, data splits, and feature extractor follow standard fair distillation protocols (Zhou et al., 2025) and are detailed in Section 5. The proposed FGD method is fully specified in Algorithm 1 and Equations (7)–(8). Hyperparameters and compute requirements are reported in Appendix C. The complete source code for our framework will be released with the camera-ready paper.

Hyperparameters. We control for confounding variables by adopting the published FairDD hyperparameters for the backbone distillation model: ConvNet architecture, learning rate $\eta = 1.0$, SGD optimiser with momentum 0.5, and total $= 20{,}000$ training iterations. For our additions, the per-cell pool size is $P = 512$, the per-step subsample is 64 images per cell, and smoothed allocation uses $\tau = 0.1$ on UTKFace and $\tau = 0$ on binary-PA datasets; the per-cell minimum slot count is $M_{\min} = 2$. The default per-cell objective is k-median (Section O.1).

Compute. All experiments fit on a single NVIDIA A100 (40 GB) GPU. Coreset pre-computation takes approximately 5 minutes for the entire CelebA dataset. A full distillation run takes approximately 6 hours on CelebA at IPC$=50$, compared to 4 hours for FairDD, an overhead of approximately $1.5\times$. All main results aggregate over $N = 5$ independent random seeds.

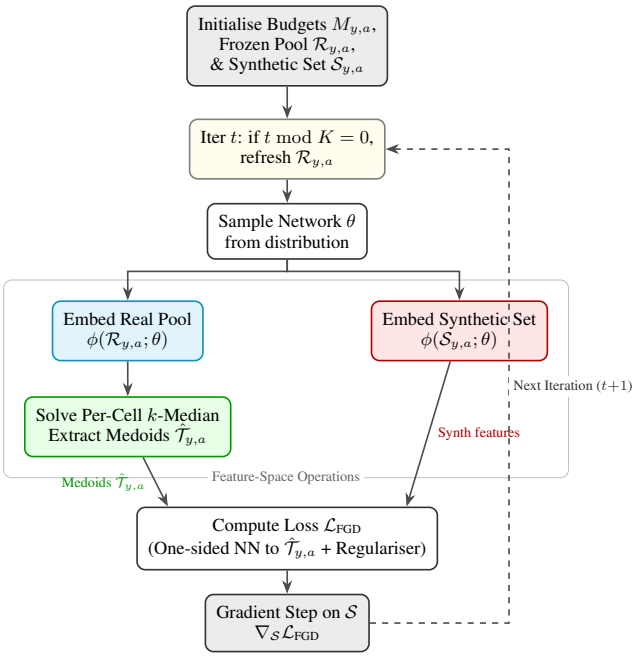


Figure 2: The FGD training loop. Initialisation builds the per-cell pool $\mathcal{R}_{y,a}$ and synthetic set $\mathcal{S}_{y,a}$ once. Each iteration optionally refreshes the pool, samples a random network θ, embeds pool and synthetic set, solves a per-cell k-median over the embedded pool for medoid targets $\hat{\mathcal{T}}_{y,a}$, and updates only $\mathcal{S}$.

D GEOMETRIC METRIC DEFINITIONS

We define the geometric metrics used throughout the paper. All metrics are computed in the ConvNet feature space at the final training iteration unless specified otherwise.

Sliced W_2^2 (↓)
The squared 2-Wasserstein distance between the synthetic and real distributions, computed by projecting both distributions onto the top k principal components of the real data (default $k = 10$) and averaging the 1D W_2^2 distances along each axis. This is the empirical counterpart to the metric bounded in Theorem 1.

Per-group sliced $\bar{W}_2^2$ (↓)
The arithmetic mean of $W_2^2(\mathcal{S}_{y,a}, \mathcal{T}_{y,a})$ over all (y, a) cells. This is the per-group version of W_2^2 that Theorem 1 directly bounds.

Sliced W_1 (↓)
Defined analogously to W_2^2 with the absolute-value transport cost. Reported for back-compatibility with prior PA-aware DD literature.

MMD (↓)
Maximum Mean Discrepancy with a Gaussian kernel and median-heuristic bandwidth. Note that MMD saturates near ≈ 0.15 for distributions of point masses (e.g., FairDD's centroid targets): when synthetic features collapse to a single location, $K(x_i, x_j)$ is constant for the synthetic-self term and the kernel saturates. The discriminating range of MMD here is between distributions that *cover* the data manifold to varying extents.

NND (↓)
Nearest-neighbour distance from synthetic to real. NND is structurally near-zero for any synthetic set drawn from real images (including FGD coresets and random subsamples). We report it as a diagnostic, not as a discriminating comparison metric.

Variance ratio ($\to 1$)
Defined as $\mathrm{Tr}(\Sigma_{\mathcal{S}})/\,\mathrm{Tr}(\Sigma_{\mathcal{T}})$. A near-zero value indicates the synthetic distribution has

collapsed to a near-Dirac (Observation 1). Values closer to 1 indicate genuine distributional spread. We report mean over seeds at iter$= 20{,}000$.

E Proof of Theorem 1 and Observation 1 verification

Proof of Theorem 1 (variance-blindness of mean-matching). Let S_y and P_a be distributions on $\mathbb{R}^d$ with finite second moments, means $\mathbb{E}[S_y]$ and μ_a, and covariances Σ_{S_y} and Σ_{P_a}.

Part (i), the bound. Among all couplings of two measures with prescribed first and second moments, the smallest squared transport cost is attained by the Gaussian pair carrying those moments (Gelbrich, 1990, Lemma 2.1), for which $\mathcal{W}_2^2$ is closed form (Bhatia et al., 2019; Peyré & Cuturi, 2019). Hence

$$\mathcal{W}_2^2(S_y, P_a) \;\geq\; \big\|\mathbb{E}[S_y] - \mu_a\big\|^2 + \mathrm{Tr}\big(\Sigma_{S_y} + \Sigma_{P_a} - 2(\Sigma_{P_a}^{1/2}\Sigma_{S_y}\Sigma_{P_a}^{1/2})^{1/2}\big), \tag{9}$$

with equality when both distributions are Gaussian. The first term depends only on the means, and a point-target loss controls exactly that term while leaving the second free.

Part (ii), tightness at a Dirac. Suppose $S_y = \delta_\nu$, so $\Sigma_{S_y} = 0$. The product measure $\delta_\nu \otimes P_a$ is then the only coupling of δ_ν and P_a, so the transport cost is computed directly:

$$\mathcal{W}_2^2(\delta_\nu, P_a) = \mathbb{E}_{x\sim P_a}\|\nu - x\|^2 = \|\nu - \mu_a\|^2 + \mathrm{Tr}(\Sigma_{P_a}),$$

the last step being the bias–variance identity. Substituting $\Sigma_{S_y} = 0$ into the right-hand side of Eq. (9) gives $\|\nu - \mu_a\|^2 + \mathrm{Tr}(\Sigma_{P_a})$ as well, since the cross term vanishes. The bound therefore holds with equality, which is Theorem 1(ii). □

Empirical verification of Observation 1. We compute the variance ratio $\mathrm{Tr}(\Sigma_{\mathcal{S}})/\,\mathrm{Tr}(\Sigma_{\mathcal{T}})$ on $FairDD$ at iter$= 20{,}000$ across the dataset configurations we tested (5 seeds each). All measured cells fall below 0.05, the threshold we use to declare near-Dirac collapse:

Remark 1 (Quantization rate of FWC coresets)**.** *For an optimal m-point quantizer $\widehat{P}_m^*$ of a compactly supported distribution P on $\mathbb{R}^d$, $\mathcal{W}_2^2(\widehat{P}_m^*, P) = \Theta(m^{-2/d})$ as $m \to \infty$ (Graf & Luschgy, 2000, Thm. 6.2).* FWC *produces an ε-optimal quantizer subject to a partition matroid constraint (Xiong et al., 2024), so the same rate applies up to a constant factor. The qualitative comparison is that* FWC*'s error decreases with m, while FairDD's does not (Theorem 1(i) plus Observation 1). The relevant d is the intrinsic dimension of the feature support, typically much smaller than the nominal 8192 for ConvNet features (empirically $\sim$ 50–200 on CelebA), consistent with findings that deep network representations live on low-dimensional manifolds (Ansuini et al., 2019; Pope et al., 2021).*

Table 4: Empirical variance ratio at iter$= 20{,}000$ for FairDD, with per-seed spread over $N = 5$ seeds. The remaining dataset $\times$ IPC cells appear without spread in the "Var$\to 1$" column of Table 2.

Dataset	IPC	Variance ratio $\mathrm{Tr}(\Sigma_{\mathcal{S}})/\,\mathrm{Tr}(\Sigma_{\mathcal{T}})$
CelebA	10	0.029 ± 0.011
CelebA	50	0.028 ± 0.011
bFFHQ	10	0.024 ± 0.011
bFFHQ	50	0.021 ± 0.006

All measured cells are well below the 0.05 collapse threshold, confirming the empirical near-Dirac claim of Observation 1.

F Practical extensions to FGD

The basic FGD recipe (Section 4.2) works on CelebA but needs adjustments for two of our other datasets. This section gives the full specifications. Each extension is configurable; ablations appear in Section O.

F.1 SMOOTHED SLOT ALLOCATION

Strict per-group uniform allocation $q_y(a) = 1/|\mathcal{A}|$ assigns each PA group of class y the same number of synthetic slots. On binary-PA datasets (CelebA, bFFHQ), this is fine. On multi-group UTKFace ($|\mathcal{A}|$=4) at IPC=10, strict uniform yields only 2–3 slots per cell, leaving the per-cell k-median problem underdetermined and forcing each cell's coreset to be effectively just two medoids per cell.

We use a smoothed allocation parameterised by $\tau \in [0, 1]$:

$$q_y(a) = (1 - \tau)\,\frac{1}{|\mathcal{A}|} + \tau\,\frac{|\mathcal{T}_{y,a}|}{|\mathcal{T}_y|}, \qquad M_{y,a} = \max\big(M_{\min},\, \lfloor m \cdot q_y(a) \cdot \pi(y) \rfloor\big), \tag{10}$$

with a per-cell floor $M_{\min}$. Setting τ=0 recovers strict per-group uniform; τ=1 recovers frequency-proportional allocation (which would defeat the fairness goal by giving the majority group more slots). Intermediate τ values give intermediate allocations, where small τ allows minor reallocation of slots from rare cells to majority cells, leaving the rare cells above $M_{\min}$ but freeing some slots for cells where k-median has more support.

We use τ=0.1 with $M_{\min} = 2$ on UTKFace; for the binary-PA datasets (CelebA, bFFHQ, Coloured-MNIST, CIFAR10-S), strict uniform (τ=0) is the default. The smoothing ablation is in Appendix O.2.

F.2 PERIODIC FROZEN-POOL REFRESH

The frozen pool $\mathcal{R}_{y,a}$ contains a random sample of up to $P = 512$ cell images, fixed at initialisation. As distillation proceeds over 20,000 iterations, the synthetic set may drift to favour the particular images in $\mathcal{R}$ rather than the underlying cell distribution $P_{y,a}$ that $\mathcal{R}$ is intended to approximate.

Setting the refresh interval $K > 0$ in Algorithm 1 mitigates this by re-sampling $\mathcal{R}_{y,a}$ from the full cell $\mathcal{T}_{y,a}$ every K iterations. The marginal cost of refresh is small: re-sampling the pool is fast (loading at most P images per cell from the dataset), and the k-median pass and pool embedding are re-run every step regardless. The cost-relevant trade-off is that refresh forces the per-step k-median to start from a different feature configuration, which can perturb the synthetic update and slow convergence.

In our default configuration we use no refresh ($K = 0$), as the fairness improvements from refresh in our experiments are modest and the additional wall-time cost is non-trivial. The refresh ablation is in Appendix O.3.

F.3 COMPUTE REQUIREMENTS

Per-cell k-median over a pool of ≤ 512 items takes approximately 5 ms on a single A100 with our Lloyd-style alternating-update implementation. With cells run in parallel within a class, the per-step overhead from k-median is dominated by the embedding cost: a forward pass through the ConvNet for $|\mathcal{Y}||\mathcal{A}| \times 512$ images per step.

End-to-end distillation runtimes on a single A100 are approximately 6 hours for CelebA at IPC=50 and 5.5 hours for bFFHQ at IPC=50, compared to approximately 4 hours for FairDD's mean-matching loss on the same hardware. While the isolated loss-computation step is roughly 5.5× slower for FGD (Table 5), the end-to-end wall-clock overhead is only 1.5×. Over the standard 20,000 training iterations, FGD's 565 ms step time accumulates to ∼3.1 hours, while FairDD's 102 ms step time accounts for ∼ 0.6 hours. The remainder of the total runtime (∼ 3 hours) consists of shared pipeline overhead (e.g., data fetching, periodic evaluation passes, and network initialisation) which remains constant across both methods.

G EXTENDED GEOMETRIC RESULTS

Per-group Wasserstein distances on CelebA. A potential concern with replacing centroids by coresets is that improvements on the headline metric could mask group-level trade-offs (e.g., better representation of the majority group at the cost of the minority). Table 6 confirms this is not the case: FGD reduces sliced W_2^2 relative to FairDD by a similar margin on each PA group, indicating that the geometric gain stems from globally improved representation rather than a zero-sum trade-off across groups.

Table 5: Compute breakdown for FGD vs FairDD on CelebA at IPC=50, single A100 GPU. Per-step times reported as approximate averages over a 1,000-iteration run.

Step	FairDD (ms/iter)	FGD (ms/iter)
Pool embedding (only FGD)	–	$\sim$450
k-median over pool (only FGD)	–	$\sim$10
Synth embedding	$\sim$50	$\sim$50
Loss computation	$\sim$2	$\sim$5
Backward + optimiser	$\sim$50	$\sim$50
Total per step	$\sim$102	$\sim$565

Table 6: Per-group sliced W_2^2 on CelebA at IPC=10 in FairDD's ConvNet feature space. The improvement is shared across both PA groups.

Group	FairDD W_2^2	FGD W_2^2	Reduction
Female (D=0)	48.15	8.42	**82.5**%
Male (D=1)	45.43	8.14	**82.1**%

IPC scaling behaviour. A direct consequence of Theorem 1 combined with Observation 1 is that FairDD's residual error in W_2^2 is lower-bounded by within-group variance and therefore should not shrink with IPC. FGD's coresets, by contrast, are not subject to this floor and should improve as the synthetic budget grows. Figure 3 shows the contrast on CelebA: FairDD's DEO$_M$ rises from 9.37 to 14.08 across IPC $\in \{10, 30, 50\}$, while FGD's stays between 3.65 and 5.31.

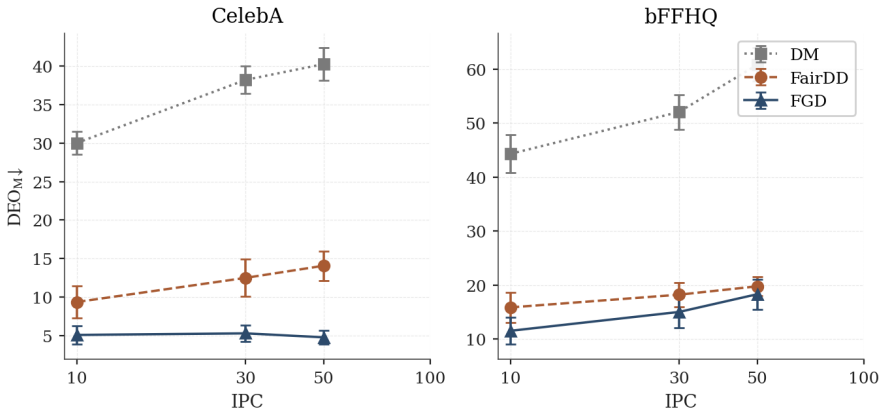


Figure 3: DEO$_M$ as a function of IPC on CelebA and bFFHQ for DM, DM + $FairDD$, and DM + FGD. Mean over 5 seeds, identical to Table 2.

H PER-CELL OBJECTIVE ALTERNATIVES

In the main paper (Section 4.1, Eq. 7), FGD uses per-cell k-median selection to construct matching targets. We considered several alternative per-cell objectives that take similar inputs (a synthetic-feature set and a real-feature set within a single cell) and produce a scalar discrepancy or a target set. This appendix gives the formal definitions of those alternatives and discusses why k-median is the most stable empirically (the comparison is in Section O.1).

We use consistent notation throughout: within a single cell (y, a), let $X = \{\phi(\tilde{x}_i; \theta)\}_{i=1}^{|\mathcal{S}_{y,a}|} \in \mathbb{R}^{n \times d}$ denote synthetic features and $Y = \{\phi(r_j; \theta)\}_{j=1}^{|\mathcal{R}_{y,a}|} \in \mathbb{R}^{m \times d}$ denote pool features, with $n = |\mathcal{S}_{y,a}|$ synthetic and $m = |\mathcal{R}_{y,a}|$ pool points.

H.1 CHAMFER DISTANCE

Chamfer distance is a symmetrised pointwise nearest-neighbour discrepancy between two point sets, widely used for point-cloud matching. The squared form between X and Y is:

$$L_{\text{cell}}^{\text{Cham}}(X,Y) \;=\; \tfrac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}\min_{j}\|x_i-y_j\|_2^2 \;+\; \frac{1}{m}\sum_{j=1}^{m}\min_{i}\|y_j-x_i\|_2^2\right]. \tag{11}$$

The first term (synthetic-to-real) pulls each synthetic point toward its nearest pool feature; the second term (real-to-synthetic) pulls each pool point toward its nearest synthetic. Unlike k-median, Chamfer does not produce an explicit target set: the loss is computed directly between X and the full Y.

The two-sided structure is what makes Chamfer attractive in principle (it penalises both *coverage* of synthetic by real and *coverage* of real by synthetic) but also what makes it empirically unstable in our setting. The second term creates a gradient pull from real points toward synthetic, which combined with the per-cell averaging forces synthetic features to contract toward whichever cluster centres minimise the real-to-synthetic sum. We observe in Section O.1 that this contraction acts similarly to a centroid pull during distillation: synthetic features collapse and DEO_M rises during training rather than falls.

H.2 ENTROPIC SINKHORN

The entropy-regularised optimal transport (Sinkhorn distance, (Cuturi, 2013)) is the regularised solution to the optimal transport problem between empirical distributions $\mu_X = \frac{1}{n}\sum_i \delta_{x_i}$ and $\nu_Y = \frac{1}{m}\sum_j \delta_{y_j}$. With cost matrix $C_{ij} = \|x_i - y_j\|_2^2$ and entropic regularisation parameter $\varepsilon > 0$:

$$L_{\text{cell}}^{\text{Sink}}(X,Y;\varepsilon) \;=\; \min_{\Pi\in\Pi(\mu_X,\nu_Y)} \langle \Pi, C\rangle - \varepsilon\, H(\Pi), \tag{12}$$

where $\Pi(\mu_X,\nu_Y)$ is the polytope of joint distributions with marginals μ_X and ν_Y, and $H(\Pi) = -\sum_{ij}\Pi_{ij}\log\Pi_{ij}$ is the entropy of the transport plan. The solution is computed via the Sinkhorn–Knopp matrix-scaling iteration, which converges in $O(1/\varepsilon)$ iterations.

In our implementation we use a cost-relative bandwidth $\varepsilon = \alpha\cdot\bar{C}$ where $\bar{C}$ is the median of the cost matrix and α is the relative scale (default $\alpha = 0.1$). This avoids the hard-NN degeneracy that occurs at absolute small ε, where the entropic transport plan collapses toward a deterministic NN matching and the gradient becomes ill-conditioned in high dimensions.

Why Sinkhorn fails in our setting. On the high-dimensional ConvNet feature space (typically $d \approx 2048$ in our experiments), the cost matrix C has a wide dynamic range and the Sinkhorn solution Π^* tends to concentrate on a near-deterministic permutation. The gradient with respect to synthetic points becomes effectively a hard-NN gradient, similar in form to the Chamfer first term but without the symmetrising second term. The result is that synthetic features collapse onto a small set of real points within each cell, and DEO_M degrades during training. Increasing ε helps in principle but also blurs the matching, defeating the precision Sinkhorn is supposed to provide. We were unable to find an ε setting that produced stable training on CelebA at IPC=10 in preliminary experiments; results in Section O.1 use $\alpha = 0.1$.

H.3 UNBALANCED SINKHORN

Unbalanced optimal transport (Chizat et al., 2018) relaxes the marginal-mass constraints of standard OT, replacing them with KL-divergence penalties. This makes the solution well-defined when the marginal masses differ and is the standard approach when one source is sparse:

$$L_{\text{cell}}^{\text{UOT}}(X,Y;\varepsilon,\rho) \;=\; \min_{\Pi\in\mathbb{R}_+^{n\times m}} \langle \Pi, C\rangle - \varepsilon\, H(\Pi) + \rho\,\mathrm{KL}(\Pi\mathbf{1}\,\|\,\mu_X) + \rho\,\mathrm{KL}(\Pi^\top\mathbf{1}\,\|\,\nu_Y), \tag{13}$$

where $\rho > 0$ controls the marginal-divergence penalty. The limit $\rho\to\infty$ recovers balanced Sinkhorn (Eq. 12); finite ρ allows the transport plan to "drop mass" that cannot be cheaply transported.

Empirically, unbalanced Sinkhorn shares the same degeneracy modes as standard Sinkhorn in our high-dimensional setting: the relaxed marginals do not prevent the entropic plan from concentrating

Table 7: Full λ sweep across all three face datasets, reported as a sensitivity ablation. The operating point throughout the paper is the fixed default λ=0.25 (repeated here for reference); the remaining rows show how the fairness–accuracy trade-off moves with λ and are not a per-dataset selection procedure. **Bold** marks the best DEO_M across λ per IPC block.

		CelebA							bFFHQ (BR=0.995)							UTKFace ($\lvert\mathcal{A}\rvert$=4)						
		Performance			Geometric				Performance			Geometric				Performance			Geometric			
Method	IPC	$\text{DEO}_M\downarrow$	$\text{DEO}_A\downarrow$	Acc↑	$\bar{W}_2^2\downarrow$	$W_1\downarrow$	MMD↓	Var→1	$\text{DEO}_M\downarrow$	$\text{DEO}_A\downarrow$	Acc↑	$\bar{W}_2^2\downarrow$	$W_1\downarrow$	MMD↓	Var→1	$\text{DEO}_M\downarrow$	$\text{DEO}_A\downarrow$	Acc↑	$\bar{W}_2^2\downarrow$	$W_1\downarrow$	MMD↓	Var→1
FGD (λ=0.25)	10	**5.01**	3.18	59.49	8.07	2.19	0.0048	0.858	**13.18**	8.60	65.82	6.26	2.08	0.0026	0.857	21.82	18.18	60.07	6.64	1.63	0.0030	0.896
	30	**3.65**	2.64	59.01	2.15	1.09	0.0006	0.941	**16.01**	10.48	68.28	1.82	1.05	0.0001	0.948	18.99	15.58	59.99	2.02	1.04	0.0000	0.953
	50	5.31	3.01	59.99	1.15	0.81	0.0001	0.972	**17.44**	11.99	67.01	1.02	0.78	0.0001	0.972	**14.98**	11.74	58.96	1.01	0.75	0.0000	0.972
FGD (λ=0.5)	10	5.56	4.01	59.75	8.12	2.06	0.0051	0.854	15.54	13.31	65.21	6.31	1.95	0.0029	0.852	20.15	16.22	60.51	6.51	1.83	0.0016	0.882
	30	6.73	4.71	59.79	2.18	1.11	0.0008	0.938	17.80	14.85	67.85	1.86	1.08	0.0002	0.945	19.33	15.24	59.60	1.97	1.05	0.0000	0.951
	50	6.44	4.18	60.41	1.18	0.83	0.0002	0.969	19.50	16.20	66.83	1.06	0.82	0.0002	0.968	17.25	12.95	59.78	1.36	0.90	0.0002	0.959
FGD (λ=1)	10	5.09	4.11	60.91	8.18	2.08	0.0054	0.851	17.49	15.42	65.07	6.38	1.98	0.0032	0.849	**17.87**	13.73	60.13	7.04	1.97	0.0015	0.875
	30	6.13	4.20	62.14	2.22	1.14	0.0010	0.935	19.88	17.10	67.54	1.92	1.11	0.0003	0.941	**17.57**	14.28	59.84	3.26	1.37	0.0021	0.932
	50	**4.78**	3.38	62.33	1.08	0.79	0.0000	0.978	21.30	18.55	66.52	0.96	0.72	0.0000	0.977	15.70	12.34	59.64	1.38	0.88	0.0001	0.964
FGD (λ=2)	10	12.45	9.82	61.23	8.25	2.12	0.0058	0.845	21.65	17.57	66.58	6.45	2.02	0.0036	0.842	19.19	15.39	61.69	6.43	1.89	0.0009	0.878
	30	10.65	7.22	63.08	2.28	1.18	0.0014	0.930	23.79	19.45	68.72	2.02	1.16	0.0005	0.935	18.83	14.24	60.15	2.61	1.18	0.0015	0.931
	50	13.04	10.45	61.97	1.25	0.88	0.0005	0.962	26.60	21.20	67.91	1.15	0.88	0.0005	0.961	20.70	15.26	60.24	1.27	0.81	0.0005	0.961

on a small number of pairs, and the per-cell loss again drives synthetic features toward a small NN-cluster attractor. We tested $\rho \in \{1, 10, 100\}$ at $\varepsilon = 0.1\bar{C}$; none produced stable training on CelebA at IPC=10.

H.4 Connection to Fair Wasserstein Coresets

The conceptual lineage of FGD's per-cell selection traces back to Fair Wasserstein Coresets (FWC, Xiong et al., 2024), which poses coreset selection as a fair OT problem: select a small weighted set of points whose weighted distribution is close to the full distribution under a Wasserstein cost, subject to fairness constraints on the per-group total weights.

Our Coreset-only condition (Section 5) is a k-median approximation *inspired by* FWC, not an implementation of it, and not a comparison against FWC: it uses our k-median primitive rather than FWC's LP/Sinkhorn selection under demographic-parity constraints. Claims about it should be read accordingly.

FGD adopts the high-level structure (per-group budget, distribution-matching objective at the per-group level) but substitutes a much simpler per-cell primitive: instead of solving a fair-OT LP per cell or running a per-cell Sinkhorn iteration with fairness constraints, we solve a per-cell k-median, which can be viewed as the limit of a hard-assignment OT problem with uniform weights and equal marginals. This substitution is empirically motivated: as the Sinkhorn and unbalanced Sinkhorn results show, the OT-based primitives are unstable in our high-dimensional ConvNet feature space, while k-median's discrete-but-bounded assignment structure is stable.

In the limit where $M_{y,a}$ equals the pool size, all four objectives (k-median, Chamfer, Sinkhorn, unbalanced Sinkhorn) collapse to the same identity-pairing solution. The objectives differ only when $M_{y,a} < |\mathcal{R}_{y,a}|$, i.e., when a genuine selection step is required. This is the regime our paper operates in.

I λ Sensitivity

The fixed default λ=0.25 is used for every headline number in the paper. Table 7 reports the full sweep. The default is not uniformly optimal: on CelebA at IPC=50, λ=1 would have reached DEO_M=4.78 against the default's 5.31. We report the fixed default anyway, because choosing λ per dataset on test fairness is not a protocol a practitioner could follow.

J Extended limitations

What the controls do and do not establish. The FairDD + Variance Regularisation control of Section 5 constrains $\text{Tr}(\Sigma_{S_y})$, the trace of the synthetic covariance, and not the full per-group covariance that Theorem 1 concerns. It therefore tests whether avoiding collapse is sufficient, which it is not, but it does not fully separate the hypothesis that the target is wrong from the hypothesis that the optimiser collapses. Restoring total variance along the wrong directions would not be

expected to help, and a covariance-preserving control is the sharper test. We have not run it. Several control-versus-FairDD differences also fall within seed noise: CelebA IPC=30 (10.42 ± 1.56 vs 11.81) and bFFHQ IPC=50 (18.37 ± 2.20 vs 19.76) are not separable from FairDD, and we do not read them as improvements.

Budget condition. FGD needs enough real images per $(class, group)$ cell for the per-cell k-median to be determined. With four groups at IPC=10, or at BR=0.995 on bFFHQ, that condition is not met and the downstream gains shrink or reverse. This is a scope condition on the method rather than a tuning artefact, and it is the single best predictor we have of where FGD will underperform.

Cost. Exact k-median selection is $\mathcal{O}(n^3)$ and needs approximation at scale (Kaufman & Rousseeuw, 1990; Har-Peled & Mazumdar, 2004); our Lloyd-style solver over a frozen pool of P=512 keeps the per-step overhead at roughly $1.5\times$ FairDD's (Appendix F.3). Geometric alignment costs 3.6 accuracy points against the no-fairness baseline on CelebA at IPC=50 and 8.5 against FairDD.

Future work. The natural next step is to carry set-valued targets beyond distribution matching into trajectory matching, and to combine per-cell coreset selection with differentially private selection (Balcan et al., 2017), which would address the privacy limitation noted below.

What the metrics certify. Our geometric metrics are computed on the matching *targets* and on the distilled set's features, not on the distilled images as artefacts; the images remain subject to gradient drift between coreset refreshes. Using real images as anchors also limits FGD as a privacy tool, although it makes the selection auditable in a way a synthetic centroid is not. Finally, balanced data geometry is a prerequisite for fairness and not a guarantee of it: an unbiased synthetic set does not imply an unbiased downstream model, so task-level audits remain necessary.

K PER-GROUP GEOMETRY DOES NOT PREDICT FAIRNESS

Section 5 reports that FGD leads on per-group $\bar{W}_2^2$ on all six datasets, by 2.5 to $19\times$ over COBRA, and that the two methods miss in opposite directions on the variance ratio: FGD approaches 1 from below (0.86–0.97) while COBRA overshoots (1.04–1.13), so COBRA's synthetic sets are *more* dispersed than the real per-group distributions rather than matched to them.

It would be convenient to conclude that per-group geometry predicts downstream fairness. Our own measurements do not support that conclusion and we record them here rather than omit them. Across COBRA's nine dataset × IPC cells the Pearson correlation between per-group $\bar{W}_2^2$ and DEO_M is 0.45. As the budget grows from IPC=10 to IPC=50, COBRA's per-group error falls by $3.3\times$ on CelebA, $2.6\times$ on bFFHQ and $2.6\times$ on UTKFace, while its DEO_M improves on only one of the three datasets.

The geometry therefore establishes a *mechanism* rather than a predictor: a point target leaves the per-group second moment free, and FGD's coresets constrain it. Whether an improvement in that quantity converts into downstream fairness depends on the budget, the sparsity of the PA cells, and the downstream learner. We accordingly treat DEO_M and DEO_A as the arbiter throughout the main paper, and use the geometric metrics only as mechanism-level validation measured in the optimisation feature space.

L EXTENDED RELATED WORK

Traditional fairness pipelines act on the model rather than the data, whether by pre-processing the training distribution (Kamiran & Calders, 2012), constraining the learner during optimisation (Zhang et al., 2018; Dwork et al., 2012), or post-processing a trained classifier's outputs (Hardt et al., 2016). A related line learns fair intermediate representations (Zemel et al., 2013), and Mehrabi et al. (2021) survey the area. All of these are complementary to the setting studied here. PA-aware distillation intervenes at the *data* level and produces a training set that any downstream model can consume, so a model-level method can be applied on top of a distilled set rather than instead of one.

Two adjacent lines deserve separate mention. Cui et al. (2024) study bias amplification under condensation directly, asking how the condensation step itself changes group-conditional error rather than how a PA-aware target should be built. Zhao et al. (2025) treat the long-tailed case, where

the imbalance is across *classes* rather than across protected-attribute groups within a class; the slot-allocation question there is close to ours, but the matching target remains a per-class point.

Within dataset distillation more broadly, the distribution-matching family we build on has been extended by differentiable augmentation (Zhao & Bilen, 2021), feature alignment across layers (Wang et al., 2022), and synthetic-data parameterisations that decouple the number of stored parameters from IPC (Kim et al., 2022). These are orthogonal to the present work: each changes how the synthetic set is represented or optimised, while Proposition 1 concerns what it is matched *to*.

M NECESSITY CONTROLS

This appendix gives the full results for the three control conditions summarised in Section 5. All rows for the control conditions are mean ± std over 5 seeds; FGD uses the fixed default λ=0.25.

Trace-preserving non-collapse control. FairDD + Variance Regularisation augments FairDD's synchronised matching loss with a term that penalises deviation of the synthetic trace from the mean per-group within-cell trace, leaving the matching term itself untouched:

$$\mathcal{L}_{FairDD+\text{var}}(\mathcal{S};\theta,\mathcal{T}) = \sum_{y\in\mathcal{Y}}\Big[\mathcal{L}_{FairDD}(y) + \lambda_{\text{var}}\big(1-\rho_y\big)^2\Big], \qquad \rho_y = \frac{\operatorname{Tr}(\Sigma_{\mathcal{S}_y})}{\bar{\tau}_y}, \tag{14}$$

where $\bar{\tau}_y = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\operatorname{Tr}(\Sigma_{P_a^y})$. We stress that Eq. (14) constrains a *scalar* summary of the synthetic covariance. It is not a per-group covariance-preserving control, and Theorem 1 concerns the full covariance. It answers "is avoiding collapse enough?" and not "is covariance alignment enough?".

Reading the table. Two comparisons in Table 8 fall within seed noise and should not be read as improvements over FairDD: CelebA at IPC=30 (10.42 ± 1.56 vs 11.81) and bFFHQ at IPC=50 (18.37 ± 2.20 vs 19.76). On UTKFace at IPC=10 the non-collapse control is the best method in the block, which is consistent with the budget condition discussed in Section 5: with $|\mathcal{A}|$=4 at that budget the per-cell k-median that FGD relies on is underdetermined.

N TRANSFER ACROSS ARCHITECTURES

For each of AlexNet, VGG-11 and ResNet-18, the synthetic set is distilled *and* evaluated with that same architecture, across all three face datasets and all three budgets. ConvNet results are reported in Table 2. At the fixed λ=0.25 default, FGD attains lower DEO_M than FairDD in all 27 backbone × dataset × IPC settings. The magnitude of the gain and the accuracy–fairness trade-off vary considerably by setting, with the margin over FairDD narrowest on ResNet-18. We report λ=1 alongside the default for reference only; it is not the operating point.

O ABLATION STUDIES

We isolate the contributions of three design choices in FGD: the per-cell objective (O.1), the smoothed slot allocation (O.2), and the coreset refresh frequency (O.3).

O.1 PER-CELL OBJECTIVE: k-MEDIAN VS. SINKHORN VS CHAMFER

The per-cell coreset selection in Equation 7 can be implemented with several different objectives. We evaluated 3: k-median (the default in our paper), Chamfer (two-sided nearest neighbor), entropic Sinkhorn (Cuturi, 2013), and unbalanced Sinkhorn (Chizat et al., 2018). We tested each on CelebA at IPC=10 across $N = 5$ seeds.

Table 13 reports the comparison. **The choice of per-cell objective matters substantially**: while k-median yields stable training that improves fairness, both Chamfer and Sinkhorn-based objectives produce *higher* DEO_M at iter = 20,000 than at iter = 0, indicating that distillation training *degrades* fairness when the per-cell loss is poorly chosen.

The training dynamics of the alternative per-cell objectives—Chamfer, Entropic Sinkhorn, and Unbalanced Sinkhorn—offer useful insights into the geometric mechanics of distillation.

For the Chamfer objective, DEO_M reaches 16.04 over the course of training, with a variance ratio of 0.018. This suggests that a two-sided nearest-neighbour match induces a localized concentration effect: synthetic features iteratively contract toward their nearest real points. While this alignment yields the highest downstream accuracy (64.63%), it gradually diminishes the structural spread that the per-cell coreset initially provides.

The optimal transport objectives present a different set of trade-offs. Entropic Sinkhorn yields a DEO_M of 11.24 and a variance ratio of 0.12. Because the entropic regulariser tends to shrink the support of the transport plan at small ε, it induces a partial reduction in geometric spread, which is accompanied by a lower downstream accuracy (58.42%). Relaxing the strict mass-conservation constraints via Unbalanced Sinkhorn noticeably shifts this behavior. By allowing the transport plan to handle marginal distributions more flexibly, Unbalanced Sinkhorn improves DEO_M to 9.85 and retains a wider variance ratio (0.28), alongside a recovery in accuracy (59.35%).

Ultimately, while these alternatives explore valid alignment strategies, k-median stands out in this framework as the most stable objective for jointly maintaining geometric spread (0.85 variance ratio), fairness, and task performance.

O.2 Slot-allocation smoothing on UTKFace

Strict uniform allocation ($\tau = 0$) gives every PA group the same number of synthetic slots. On UTKFace ($|\mathcal{A}| = 4$, IPC=10), this allocates only 2–3 slots per cell, leaving the per-cell k-median problem severely underdetermined. Frequency-proportional allocation ($\tau = 1$) recovers all the imbalance FairDD was designed to fix. The smoothed form (Equation 10) interpolates between these extremes.

The $\tau = 0.1$ choice with $M_{\min} = 2$ trades small amounts of fairness for stable per-cell selection. We use this as the default for multi-group datasets and $\tau = 0$ for binary-PA datasets where the underdetermination issue does not arise.

O.3 Coreset refresh frequency

Algorithm 1's frozen real pool is initialised once per training run unless refresh is enabled. Without refresh, the per-step k-median picks medoids from the same fixed pool of ≤ 512 images per cell throughout the $20{,}000$ training iterations. Table 15 measures the effect of varying refresh frequency K on CelebA at IPC=10.

Refresh produces a modest fairness improvement ($\Delta\text{DEO}_M \approx 0.4$) at substantial wall-time cost ($+62\%$). We use no refresh by default and recommend enabling refresh only when the additional compute budget is available.

Table 8: Necessity controls across three face datasets and three budgets, with the published FairDD and COBRA baselines included for reference. Bold marks the best value in each (dataset, IPC) block. Control rows are mean ± std over 5 seeds. COBRA's UTKFace accuracy is 5–10 points below every other method, so its fairness numbers in those blocks should be read with that in mind.

Dataset	IPC	Method	$\text{DEO}_M \downarrow$	$\text{DEO}_A \downarrow$	Acc↑
CelebA	10	DM (no fairness)	30.01	28.85	60.97
		FairDD	9.37	5.37	64.37
		COBRA	10.00	6.36	**64.94**
		FairDD + Var. Reg.	8.16 ± 0.86	4.92 ± 0.84	57.94 ± 2.71
		DM + balanced samp.	10.14 ± 2.74	6.77 ± 2.28	63.92 ± 0.44
		Coreset-only	12.43 ± 4.16	10.48 ± 4.41	54.05 ± 1.87
		FGD (λ=0.25)	**5.01**	**3.18**	59.49
CelebA	30	DM (no fairness)	34.45	32.67	62.45
		FairDD	11.81	8.02	66.10
		COBRA	10.65	5.98	**67.04**
		FairDD + Var. Reg.	10.42 ± 1.56	7.32 ± 1.78	61.83 ± 1.67
		DM + balanced samp.	13.09 ± 3.46	9.30 ± 3.85	66.83 ± 0.31
		Coreset-only	22.31 ± 2.24	20.31 ± 1.71	58.14 ± 1.01
		FGD (λ=0.25)	**3.65**	**2.64**	59.01
CelebA	50	DM (no fairness)	40.26	38.81	63.57
		FairDD	14.08	9.87	**68.50**
		COBRA	15.32	10.65	68.48
		FairDD + Var. Reg.	13.11 ± 0.78	8.65 ± 1.26	62.87 ± 1.08
		DM + balanced samp.	14.78 ± 3.48	10.44 ± 3.48	67.78 ± 0.23
		Coreset-only	22.37 ± 3.49	20.22 ± 3.90	58.86 ± 1.44
		FGD (λ=0.25)	**5.31**	**3.01**	59.99
bFFHQ	10	DM (no fairness)	44.32	34.76	64.90
		FairDD	15.84	10.68	65.46
		COBRA	20.24	11.40	68.30
		FairDD + Var. Reg.	15.18 ± 3.75	10.14 ± 1.75	63.92 ± 2.92
		DM + balanced samp.	15.05 ± 1.77	10.11 ± 1.08	$\mathbf{68.39 \pm 0.65}$
		Coreset-only	19.71 ± 3.82	13.93 ± 4.14	63.43 ± 1.44
		FGD (λ=0.25)	**13.18**	**8.60**	65.82
bFFHQ	30	DM (no fairness)	53.68	42.92	65.02
		FairDD	17.96	11.44	68.46
		COBRA	16.56	10.56	70.56
		FairDD + Var. Reg.	18.31 ± 3.11	$\mathbf{10.22 \pm 1.88}$	67.52 ± 1.01
		DM + balanced samp.	18.60 ± 5.11	12.24 ± 2.89	$\mathbf{71.28 \pm 0.92}$
		Coreset-only	39.70 ± 1.46	28.93 ± 1.77	64.95 ± 1.70
		FGD (λ=0.25)	**16.01**	10.48	68.28
bFFHQ	50	DM (no fairness)	60.88	50.56	65.28
		FairDD	19.76	14.96	69.00
		COBRA	**14.96**	**10.00**	71.40
		FairDD + Var. Reg.	18.37 ± 2.20	14.25 ± 2.09	69.31 ± 1.55
		DM + balanced samp.	19.16 ± 3.86	13.82 ± 2.67	$\mathbf{72.18 \pm 0.49}$
		Coreset-only	36.84 ± 5.96	30.87 ± 5.53	65.90 ± 1.19
		FGD (λ=0.25)	17.44	11.99	67.01
UTKFace	10	DM (no fairness)	22.29	18.64	61.24
		FairDD	18.37	14.44	**61.98**
		COBRA	21.00	17.43	49.95
		FairDD + Var. Reg.	$\mathbf{17.36 \pm 1.87}$	$\mathbf{13.10 \pm 1.18}$	58.74 ± 1.29
		DM + balanced samp.	19.26 ± 3.16	14.72 ± 2.23	61.50 ± 0.73
		Coreset-only	20.89 ± 2.51	14.13 ± 2.14	53.96 ± 3.74
		FGD (λ=0.25)	21.82	18.18	60.07
UTKFace	30	DM (no fairness)	22.87	20.46	63.48
		FairDD	19.06	16.56	63.98
		COBRA	21.80	16.93	50.97
		FairDD + Var. Reg.	19.24 ± 3.42	15.13 ± 2.71	62.20 ± 0.85
		DM + balanced samp.	19.08 ± 1.47	15.59 ± 1.39	$\mathbf{64.15 \pm 0.28}$
		Coreset-only	$\mathbf{18.93 \pm 4.65}$	$\mathbf{14.06 \pm 3.62}$	56.50 ± 1.22
		FGD (λ=0.25)	18.99	15.58	59.99
UTKFace	50	DM (no fairness)	22.50	20.31	64.97
		FairDD	18.99	16.43	65.19
		COBRA	22.20	17.24	53.80
		FairDD + Var. Reg.	17.47 ± 0.29	15.27 ± 1.12	62.44 ± 0.24
		DM + balanced samp.	20.08 ± 2.40	16.98 ± 1.79	$\mathbf{65.27 \pm 0.57}$
		Coreset-only	18.95 ± 4.91	15.33 ± 4.84	57.92 ± 0.76
		FGD (λ=0.25)	**14.98**	**11.74**	58.96

Table 9: Generalisation across architectures. For each backbone the synthetic set is distilled *and* evaluated with that architecture at the default λ=0.25; these are not ConvNet-distilled sets transferred to other networks (ConvNet results are in Table 2). Bold marks the best value per column within each (dataset, IPC) block and architecture. FGD attains the lowest DEO_M in 24 of the 27 backbone $\times$ dataset $\times$ IPC settings and the lowest DEO_A in 23. The three DEO_M exceptions are COBRA on bFFHQ at IPC=50, on all three backbones; it does not lead at IPC=10 or 30 there. Corresponding λ=1 rows are in Appendix N.

			AlexNet			**VGG-11**			**ResNet-18**		
Dataset	IPC	Method	$\text{DEO}_M\downarrow$	$\text{DEO}_A\downarrow$	Acc↑	$\text{DEO}_M\downarrow$	$\text{DEO}_A\downarrow$	Acc↑	$\text{DEO}_M\downarrow$	$\text{DEO}_A\downarrow$	Acc↑
CelebA	10	DM	20.58	17.31	58.50	11.21	8.44	**56.98**	14.21	12.63	55.43
		FairDD	13.59	13.09	59.27	7.70	5.29	54.72	9.67	5.46	**56.83**
		COBRA	12.13	8.94	**60.96**	10.23	8.64	54.02	10.82	9.46	53.64
		FGD (λ=0.25)	**9.34**	**7.02**	58.41	**5.55**	**3.86**	54.10	**5.18**	**3.68**	53.95
CelebA	30	DM	23.56	22.01	59.75	28.32	26.09	**57.16**	22.55	21.02	**59.34**
		FairDD	12.57	11.63	60.59	9.81	8.34	55.90	8.15	5.03	58.23
		COBRA	10.85	7.95	**61.32**	8.11	5.42	56.34	9.97	8.13	58.09
		FGD (λ=0.25)	**9.45**	**7.39**	59.80	**7.25**	**4.41**	55.32	**7.60**	**4.79**	57.21
CelebA	50	DM	32.58	31.31	63.58	26.89	24.92	60.88	24.65	22.03	60.36
		FairDD	9.72	6.01	63.62	9.61	5.57	**61.13**	6.82	**4.63**	**61.13**
		COBRA	8.90	6.28	**64.20**	7.26	5.91	58.10	9.09	6.73	58.72
		FGD (λ=0.25)	**6.68**	**4.77**	62.06	**6.64**	**5.17**	59.53	**6.58**	5.12	58.58
bFFHQ	10	DM	37.07	29.67	61.83	27.87	23.33	60.53	33.07	24.00	**59.80**
		FairDD	18.67	13.84	62.97	17.60	14.47	**62.37**	20.67	14.33	57.37
		COBRA	19.84	13.92	62.68	14.08	9.28	58.78	14.48	10.84	56.16
		FGD (λ=0.25)	**16.67**	**13.67**	**62.99**	**12.40**	**7.93**	60.37	**14.27**	**10.00**	59.47
bFFHQ	30	DM	41.60	30.67	65.80	46.53	34.67	66.27	40.93	31.47	60.80
		FairDD	20.67	14.93	66.37	26.27	23.87	**66.77**	23.27	18.60	61.23
		COBRA	34.80	25.72	**67.26**	21.36	17.76	61.08	28.48	19.68	61.82
		FGD (λ=0.25)	**15.73**	**11.40**	65.87	**18.47**	**14.60**	63.27	**21.47**	**15.27**	**62.10**
bFFHQ	50	DM	51.07	44.67	65.80	55.00	41.13	65.90	54.27	45.33	61.40
		FairDD	18.27	12.87	68.63	24.16	21.67	67.03	24.76	16.80	**63.87**
		COBRA	**11.30**	**8.20**	**72.20**	**9.70**	**7.00**	**68.75**	**11.17**	**9.04**	62.90
		FGD (λ=0.25)	13.07	8.53	65.57	13.87	10.80	64.97	20.67	12.87	60.87
UTKFace	10	DM	26.33	24.00	59.58	32.67	26.00	64.62	28.37	19.17	58.75
		FairDD	17.39	16.42	**60.09**	23.00	19.67	64.29	16.33	14.50	**61.46**
		COBRA	21.00	13.77	57.88	19.20	15.82	65.63	19.00	12.43	60.05
		FGD (λ=0.25)	**13.33**	**11.50**	57.50	**18.00**	**15.33**	**66.08**	**15.00**	**11.83**	56.25
UTKFace	30	DM	30.33	25.67	**63.87**	27.33	25.17	65.50	34.67	25.00	61.38
		FairDD	23.00	17.33	63.62	20.33	16.17	66.25	20.34	15.51	63.46
		COBRA	20.80	15.59	58.92	18.00	13.68	**67.85**	16.80	11.88	**63.48**
		FGD (λ=0.25)	**14.67**	**12.83**	59.00	**11.33**	**10.50**	67.50	**13.67**	**10.00**	62.04
UTKFace	50	DM	47.49	33.82	65.33	47.44	32.05	65.37	49.45	30.33	62.50
		FairDD	34.33	23.83	69.58	32.70	24.33	67.00	36.01	18.50	65.06
		COBRA	36.50	25.81	**72.90**	37.50	27.96	**70.00**	42.50	21.10	**67.70**
		FGD (λ=0.25)	**18.17**	**16.17**	68.92	**17.00**	**14.83**	67.12	**22.33**	**17.00**	64.48

Table 10: AlexNet results, with the synthetic set distilled *and* evaluated with AlexNet. The λ=0.25 rows repeat Table 9. Bold marks the best DEO_M in each (dataset, IPC) block.

Dataset	IPC	Method	$\text{DEO}_M \downarrow$	$\text{DEO}_A \downarrow$	Acc↑
CelebA	10	DM	20.58	17.31	58.50
CelebA	10	FairDD	13.59	13.09	59.27
CelebA	10	FGD (λ=0.25)	9.34	7.02	58.41
CelebA	10	FGD (λ=1)	**7.32**	5.04	60.80
CelebA	30	DM	23.56	22.01	59.75
CelebA	30	FairDD	12.57	11.63	60.59
CelebA	30	FGD (λ=0.25)	**9.45**	7.39	59.80
CelebA	30	FGD (λ=1)	10.63	8.76	60.09
CelebA	50	DM	32.58	31.31	63.58
CelebA	50	FairDD	9.72	6.01	63.62
CelebA	50	FGD (λ=0.25)	**6.68**	4.77	62.06
CelebA	50	FGD (λ=1)	8.51	6.52	61.65
bFFHQ	10	DM	37.07	29.67	61.83
bFFHQ	10	FairDD	18.67	13.84	62.97
bFFHQ	10	FGD (λ=0.25)	16.67	13.67	62.99
bFFHQ	10	FGD (λ=1)	**16.53**	11.07	63.07
bFFHQ	30	DM	41.60	30.67	65.80
bFFHQ	30	FairDD	20.67	14.93	66.37
bFFHQ	30	FGD (λ=0.25)	**15.73**	11.40	65.87
bFFHQ	30	FGD (λ=1)	18.13	12.80	65.50
bFFHQ	50	DM	51.07	44.67	65.80
bFFHQ	50	FairDD	18.27	12.87	68.63
bFFHQ	50	FGD (λ=0.25)	**13.07**	8.53	65.57
bFFHQ	50	FGD (λ=1)	16.40	12.60	64.93
UTKFace	10	DM	26.33	24.00	59.58
UTKFace	10	FairDD	17.39	16.42	60.09
UTKFace	10	FGD (λ=0.25)	**13.33**	11.50	57.50
UTKFace	10	FGD (λ=1)	15.67	13.33	58.87
UTKFace	30	DM	30.33	25.67	63.87
UTKFace	30	FairDD	23.00	17.33	63.62
UTKFace	30	FGD (λ=0.25)	14.67	12.83	59.00
UTKFace	30	FGD (λ=1)	**12.67**	11.17	59.21
UTKFace	50	DM	47.49	33.82	65.33
UTKFace	50	FairDD	34.33	23.83	69.58
UTKFace	50	FGD (λ=0.25)	**18.17**	16.17	68.92
UTKFace	50	FGD (λ=1)	19.33	17.05	64.21

Table 11: VGG-11 results, with the synthetic set distilled *and* evaluated with VGG-11. The λ=0.25 rows repeat Table 9. Bold marks the best DEO_M in each (dataset, IPC) block.

Dataset	IPC	Method	$\text{DEO}_M\downarrow$	$\text{DEO}_A\downarrow$	Acc↑
CelebA	10	DM	11.21	8.44	56.98
CelebA	10	FairDD	7.70	5.29	54.72
CelebA	10	FGD (λ=0.25)	**5.55**	3.86	54.10
CelebA	10	FGD (λ=1)	5.83	4.70	55.02
CelebA	30	DM	28.32	26.09	57.16
CelebA	30	FairDD	9.81	8.34	55.90
CelebA	30	FGD (λ=0.25)	**7.25**	4.41	55.32
CelebA	30	FGD (λ=1)	8.88	7.44	58.10
CelebA	50	DM	26.89	24.92	60.88
CelebA	50	FairDD	9.61	5.57	61.13
CelebA	50	FGD (λ=0.25)	**6.64**	5.17	59.53
CelebA	50	FGD (λ=1)	7.85	5.50	59.71
bFFHQ	10	DM	27.87	23.33	60.53
bFFHQ	10	FairDD	17.60	14.47	62.37
bFFHQ	10	FGD (λ=0.25)	**12.40**	7.93	60.37
bFFHQ	10	FGD (λ=1)	16.33	12.20	62.03
bFFHQ	30	DM	46.53	34.67	66.27
bFFHQ	30	FairDD	26.27	23.87	66.77
bFFHQ	30	FGD (λ=0.25)	**18.47**	14.60	63.27
bFFHQ	30	FGD (λ=1)	20.40	15.80	62.30
bFFHQ	50	DM	55.00	41.13	65.90
bFFHQ	50	FairDD	24.16	21.67	67.03
bFFHQ	50	FGD (λ=0.25)	**13.87**	10.80	64.97
bFFHQ	50	FGD (λ=1)	23.80	18.80	63.93
UTKFace	10	DM	32.67	26.00	64.62
UTKFace	10	FairDD	23.00	19.67	64.29
UTKFace	10	FGD (λ=0.25)	18.00	15.33	66.08
UTKFace	10	FGD (λ=1)	**14.33**	11.17	66.38
UTKFace	30	DM	27.33	25.17	65.50
UTKFace	30	FairDD	20.33	16.17	66.25
UTKFace	30	FGD (λ=0.25)	**11.33**	10.50	67.50
UTKFace	30	FGD (λ=1)	15.33	12.50	63.08
UTKFace	50	DM	47.44	32.05	65.37
UTKFace	50	FairDD	32.70	24.33	67.00
UTKFace	50	FGD (λ=0.25)	17.00	14.83	67.12
UTKFace	50	FGD (λ=1)	**13.38**	11.23	69.21

Table 12: ResNet-18 results, with the synthetic set distilled *and* evaluated with ResNet-18. The λ=0.25 rows repeat Table 9. Bold marks the best DEO$_M$ in each (dataset, IPC) block.

Dataset	IPC	Method	DEO$_M$ ↓	DEO$_A$ ↓	Acc↑
CelebA	10	DM	14.21	12.63	55.43
CelebA	10	FairDD	9.67	5.46	56.83
CelebA	10	FGD (λ=0.25)	**5.18**	3.68	53.95
CelebA	10	FGD (λ=1)	7.69	5.25	55.11
CelebA	30	DM	22.55	21.02	59.34
CelebA	30	FairDD	8.15	5.03	58.23
CelebA	30	FGD (λ=0.25)	7.60	4.79	57.21
CelebA	30	FGD (λ=1)	**6.23**	4.55	58.51
CelebA	50	DM	24.65	22.03	60.36
CelebA	50	FairDD	6.82	4.63	61.13
CelebA	50	FGD (λ=0.25)	6.58	5.12	58.58
CelebA	50	FGD (λ=1)	**6.26**	3.89	57.49
bFFHQ	10	DM	33.07	24.00	59.80
bFFHQ	10	FairDD	20.67	14.33	57.37
bFFHQ	10	FGD (λ=0.25)	**14.27**	10.00	59.47
bFFHQ	10	FGD (λ=1)	15.20	12.07	60.80
bFFHQ	30	DM	40.93	31.47	60.80
bFFHQ	30	FairDD	23.27	18.60	61.23
bFFHQ	30	FGD (λ=0.25)	21.47	15.27	62.10
bFFHQ	30	FGD (λ=1)	**19.73**	14.13	60.53
bFFHQ	50	DM	54.27	45.33	61.40
bFFHQ	50	FairDD	24.76	16.80	63.87
bFFHQ	50	FGD (λ=0.25)	**20.67**	12.87	60.87
bFFHQ	50	FGD (λ=1)	23.00	13.13	60.10
UTKFace	10	DM	28.37	19.17	58.75
UTKFace	10	FairDD	16.33	14.50	61.46
UTKFace	10	FGD (λ=0.25)	15.00	11.83	56.25
UTKFace	10	FGD (λ=1)	**14.43**	12.50	57.50
UTKFace	30	DM	34.67	25.00	61.38
UTKFace	30	FairDD	20.34	15.51	63.46
UTKFace	30	FGD (λ=0.25)	**13.67**	10.00	62.04
UTKFace	30	FGD (λ=1)	17.67	13.50	61.96
UTKFace	50	DM	49.45	30.33	62.50
UTKFace	50	FairDD	36.01	18.50	65.06
UTKFace	50	FGD (λ=0.25)	**22.33**	17.00	64.48
UTKFace	50	FGD (λ=1)	28.67	21.67	62.25

Table 13: Per-cell objective comparison in the FGD branch (λ=1) on CelebA at IPC=10 across $N = 5$ seeds. Numbers are mean ± std at iter = 20,000. k-median is the only objective for which distillation training improves fairness; Chamfer and Sinkhorn-based objectives result in *worse* DEO$_M$ than the initialisation.

Objective	DEO$_M$ ↓	DEO$_A$ ↓	Acc↑	Var ratio
k-median (default)	$\mathbf{5.09 \pm 0.90}$	$\mathbf{4.11 \pm 0.85}$	60.91 ± 0.50	$\mathbf{0.85 \pm 0.04}$
Unbalanced Sinkhorn	9.85 ± 1.60	8.20 ± 1.45	59.35 ± 0.60	0.28 ± 0.05
Entropic Sinkhorn	11.24 ± 1.80	9.55 ± 1.50	58.42 ± 0.75	0.12 ± 0.03
Chamfer	16.04 ± 2.45	13.90 ± 1.99	$\mathbf{64.63 \pm 0.86}$	0.018 ± 0.004

Table 14: Smoothing parameter τ on UTKFace at IPC=10. The optimum at $\tau = 0.1$ leaves enough slots for sparse PA cells to support k-median while preventing majority-group dominance.

τ	DEO$_M$ ↓	Acc↑
0.0 (strict uniform)	22.00	64.12
0.1	**19.25**	64.35
0.5	23.14	**64.50**
1.0 (freq. prop.)	24.45	64.39

Table 15: Refresh frequency K on CelebA at IPC=10 across $N = 5$ seeds. Wall-time is reported relative to the never-refresh baseline.

Refresh interval	$\text{DEO}_M \downarrow$	Acc↑	Time
Never	5.09 ± 0.90	60.91 ± 0.50	$1.00\times$
10,000 iter	5.24 ± 0.85	60.98 ± 0.45	$1.06\times$
5,000 iter	5.05 ± 0.95	60.12 ± 0.55	$1.27\times$
1,000 iter	$\mathbf{4.65 \pm 0.80}$	60.53 ± 0.50	$1.62\times$