Data Mixing for Group Preference Heterogeneity in Collaborative Filtering

Abstract

Machine learning models often fail to capture the unique preferences of social groups. While much work has focused on model development to capture heterogeneity, we examine how the composition of training data, mediated by group-level data mixing, affects group-preference alignment. The central question is: given a target group, how does adding training data from augmentation groups impact model alignment with the target group's preferences? We examine preference alignment in the context of collaborative filtering. While it is difficult to specify the optimal data mix for generic prediction models, we show that in a matrix completion setting, the recovery-bound optimal mix minimizes the prevalence disparity among item classes. Strikingly, experiments on benchmark recommendation datasets reveal that optimizing the data mix does not reliably increase group alignment, because at standard embedding dimensions such as $d = 64$, the differences among groups are insufficient to warrant data mixing. In very low-dimensional models, however, data mixing can leverage group differences to increase preference alignment. The experiments are consistent with our theoretical result showing that the augmentation groups most similar to the target are not necessarily the most beneficial for alignment.

1. Introduction

Despite the heterogeneity of human preferences, machine learning (ML) models continue to homogenize preference plurality, resulting in outputs that do not reflect group differences (Bommasani et al., 2022; Wang et al., 2025b;a). Large language models, for instance, have been shown to flatten the identities of minority groups, failing to capture the range of values and identities that exist within coarse demographic partitions (Wang et al., 2025a; Santurkar et al., 2023; Zhang et al., 2026). Recommender systems have simultaneously been shown to exhibit popularity bias, recommending already popular, mainstream items despite a user's history of preferring niche items (Klimashevskaia et al., 2024; Abdollahpouri et al., 2021; 2019).

Many existing approaches to developing more preference-aligned models focus largely on the machine learning model and post-processing techniques. In LLMs, studies emphasize fairness-aware objectives (Wang et al., 2025b), representation audits of identity flattening (Wang et al., 2025a; Santurkar et al., 2023), and prompt-time controls for simulated participants (Argyle et al., 2023); in recommender systems, mitigation efforts similarly focus on algorithmic re-ranking and exposure control (Klimashevskaia et al., 2024). While recent work has established that data composition can strongly affect model behavior (Chen et al., 2025), its role in preference alignment remains undercharacterized compared with model-centric interventions. Specifically, there is a need to understand better how interaction effects among data from diverse social groups impact homogenization. Thus, we ask: given a *target* group with distinct preferences, how does adding training data from *augmentation* (other) groups impact performance for target users?

Prior work suggests two possible outcomes from group-level data augmentation. First, augmentation-group preferences may exhibit a distributional shift relative to target-group preferences (Shen et al., 2024). In these scenarios, augmentation data may *harm* the target group. On the other hand, it is also possible that all groups share a single underlying data distribution, and additional data mitigates sparsity. These outcomes correspond to two data-composition extremes: training isolated per-group models (no cross-group augmentation) versus training a single pooled model that ignores group identity in data composition. If practical settings lie between these extremes, our question is how to optimally interpolate the data composition.

To better understand the conditions under which the two extremes occur, we focus on collaborative filtering (CF). The goal of CF is to learn a shared low-rank representation of users and items that best recovers unobserved values. We define group-preference alignment as the performance of the CF model trained on aggregate data for recovering unobserved values for target users. We note that by definition, the model is already collaborative: leveraging data across

. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

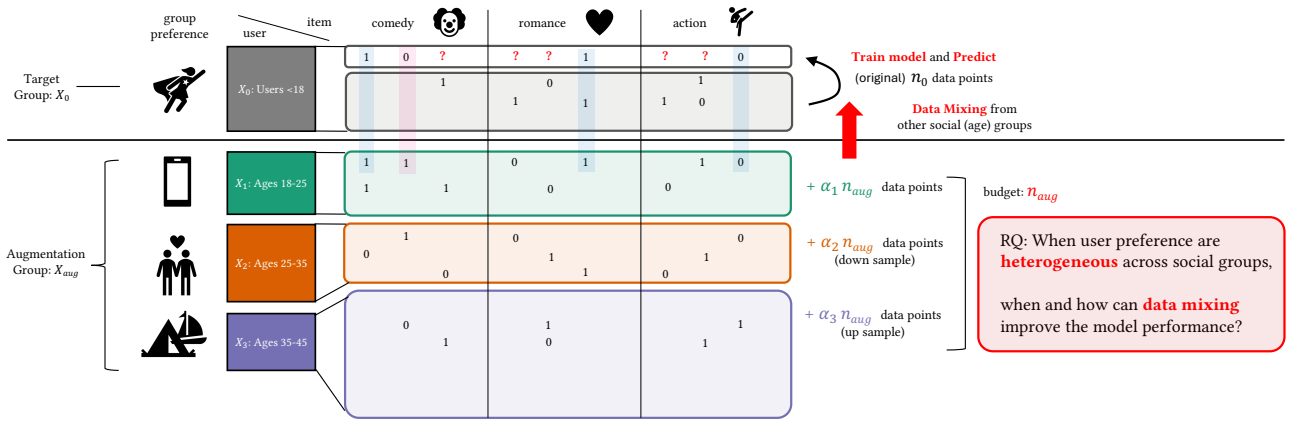


Figure 1. An example illustration of data mixing for a collaborative filtering (CF) task where different user groups have heterogeneous preferences. The input is sparse historical movie-viewership data from users bucketed by age group. The task is to train a CF model to best recommend movies to users under 18, leveraging data from other age groups as needed. Such data mixing can help provide preference signals that overlap with the target group (illustrated by the blue-shaded region); however, it may also introduce noisy signals (illustrated by the red-shaded region). The core decision is determining the ratio of data, $\boldsymbol{\alpha}_j$, to sample from each age group, given the total data-mixing budget of n_{aug}, to improve model performance (e.g., precision) for the target social group.

groups to uncover shared latent structure. However, it remains unclear how augmentation groups impact the model and whether the learned structure serves the target group.

To operationalize alternative data compositions, we adapt data mixing, a technique for curating LLM training data (Xie et al., 2023; Chen et al., 2025; 2026), for the goal of heterogeneous group-preference alignment. Past work demonstrates that curating the ratios of training data from input sources, such as the ratio of code files to chat logs in the training corpus, can significantly increase model performance (Xie et al., 2023; Chen et al., 2025). Instead, we apply data mixing to the socially motivated question of group-preference alignment. Figure 1 provides a concrete example involving data mixing for group alignment in CF.

Contributions We provide theoretical and empirical contributions to clarify the role of data mixing for CF. For instance, we prove that in a matrix completion setting, the high-probability recovery bound is minimized by data mixes that reduce item-class prevalence disparity. Our result implies that when groups share the same latent item classes, e.g., movie genre definitions, mixing data from augmentation groups *differing* from the target group improves alignment. The prevalence of the least-popular class (e.g., genre) in the target group is the bottleneck for learning. So, adding data from groups that do engage with the bottleneck genre complements the target-group data.

Empirically, we show that, strikingly, data mixing on two benchmark recommendation datasets does not reliably increase target-group performance. To explain this negative result, we find that at standard ($d = 64$) embedding dimensions, groups in benchmark datasets do not exhibit sufficiently differing preferences, corroborating prior work (Zhang et al., 2026). However, at very low embedding dimensions (e.g., $d = 4$), groups exhibit heterogeneity, and data mixing yields more aligned models. Consistent with our theory, the augmentation groups with the greatest representation in the training data are not necessarily those most similar to the target group, but rather those that complement.

2. Related Work

We tackle a problem motivated by the literature on homogenization in ML models. In the process, we use data mixing as the main intervention: a technique typically used to tune the ratios of data sources in language-model training, which we adapt for CF to align group preferences.

2.1. Homogenization

Prior work identifies homogenization as a risk of algorithmic monoculture, where shared models can induce repeated outcomes across deployments (Bommasani et al., 2022). In LLM settings, models can misportray and flatten identity groups when used as substitutes for human participants (Wang et al., 2025a). Relatedly, fairness work argues that some forms of group-differentiated behavior are desirable, and that parity alone can mask legitimate differences in group needs and preferences (Wang et al., 2025b). Complementary evidence on opinion alignment also finds demographic misalignment in current LLMs (Santurkar et al., 2023).

This body of work points to a data problem: learning pluralistic models requires datasets that exhibit heterogeneous preference signals. Recent dataset efforts explicitly pursue

this goal (Zhang et al., 2026). In collaborative filtering, an analogous concern appears as popularity bias: recommenders over-prioritize popular items and can under-serve niche tastes (Klimashevskaia et al., 2024; Abdollahpouri et al., 2021; 2019). Our work connects these threads by asking when cross-group augmentation improves target-group alignment and when it does not.

This question is also related to work on emergent specialization in interactive learning systems. These studies show that user-specific or subgroup-specialized behavior can arise naturally from participation dynamics and retraining, and can improve outcomes for diverse users (Dean et al., 2024; Su & Dean, 2024; Bose et al., 2024). Our setting is complementary: rather than changing the deployment architecture to user-specific services, we ask how much alignment can be recovered by re-composing the shared training data itself.

2.2. Data Mixing and Addition

In LLMs, data mixing usually involves selecting mixture weights across domains (for example, web, code, and books). Recent methods frame this as an optimization problem over domain weights and report gains over fixed heuristics (Xie et al., 2023; Chen et al., 2025; 2026). This line of work is related to reweighting for distributionally robust neural networks (Sagawa et al., 2020). In parallel, data-centric benchmark work continues to emphasize that dataset composition is a first-order driver of model quality (Li et al., 2024).

However, due to distribution shift, additional data is not always beneficial. In multi-source settings, adding data can reduce accuracy and worsen subgroup outcomes when distribution shift dominates scaling gains (Shen et al., 2024). Similar concerns arise in LLM-based human simulation settings, where representational fidelity can vary across tasks and groups (Argyle et al., 2023).

Against this backdrop, our contribution is to characterize conditions under which group-level data mixing helps in collaborative filtering. Rather than assuming that more augmentation data is always better, we analyze when mixing across groups improves target-group alignment and when gains are limited. In contrast to primarily empirical LLM data-mixing studies, the CF setting here admits explicit theoretical guarantees.

3. Data Mixing for Collaborative Filtering

In this section, we formally define the CF data mixing problem. While principled optimal data mixes are not evident for general prediction models, in the case of CF, we show that the optimal data mix can be characterized. Under a stylized matrix-completion setting, the optimal data mix minimizes the prevalence gap among latent item classes, which is our main theoretical contribution. A consolidated notation table is provided in Appendix A.

3.1. Problem Definition

Suppose there are n users partitioned into g mutually-exclusive groups. For each user, we have access to their historical interactions over m items. The goal is to predict unseen interactions for each user.

The data mixing problem involves optimizing the proportion of training data to use from each *augmentation* user group to maximize performance for a *target* group. Let $X_0 \in \{0,1\}^{n_0 \times m}$ denote training interaction data for n_0 users from the target group over m items, where entry $X_0(i,j) = 1$ if and only if user i interacted with item j in the training data. Similarly, let $X_1, \ldots, X_g$ denote the training data for g augmentation groups, where group j has n_j users.

The CF model f is trained on the target-group data X_0 concatenated with a probabilistic augmentation dataset X_{aug}. The augmentation dataset is parameterized by *mixing vector* $\boldsymbol{\alpha} \in \Delta_{g-1}$ and a budget ratio α_{aug}. The mixing vector specifies the amount of data sampled from each group, whereas the budget ratio specifies the total amount of augmentation data. For example, given α_{aug} and $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_j n_{\text{aug}}$ users are sampled with replacement from group j, where $n_{\text{aug}} = \alpha_{\text{aug}}(n - n_0)$. Let X_α be the aggregate training dataset.

The trained CF model is evaluated using a metric $\mathcal{M}$ that measures the performance of f only for *target-group* users. The data-mixing problem for CF is:

$$\arg \max_{\boldsymbol{\alpha} \in \Delta_{g-1}} \mathcal{M}\left(f\left(X_\alpha\right)\right) \tag{1}$$

Standard ranking metrics, such as Precision@k, are possible instantiations of $\mathcal{M}$. Henceforth, we use "target alignment" to refer to $\mathcal{M}$.

3.2. Optimal Data Mixing

We note several questions that emerge from the data-mixing problem formulation. First, we can ask whether adding augmentation data helps performance for the target group; perhaps the target group is sufficiently distinct that additional training data hurts target alignment. Second, if augmentation data improves performance, which augmentation groups have the highest mixing ratios? Are groups more similar to the target group sampled with greater probability? Third, does optimizing the mixing ratio and budget strictly increase target alignment relative to training on the data as is without sampling?

We characterize the optimal data mix for a matrix completion setting to answer these questions. We show that the target alignment benefits most from augmentations that re-

duce disparities in item-class prevalence. In the language of our movies-running example, alignment is highest when the popularity gap between the most and least popular genres is minimized. Thus, complementary augmentation groups that increase underrepresented item classes most improve performance.

3.2.1. Data Mixing for Matrix Completion

Building on the CF data-mixing objective in Eq. (1), we instantiate f and $\mathcal{M}$ for the narrower matrix completion setting, where an optimal data mix is analytically characterizable.

In matrix completion, entries of a true low-rank data matrix are observed uniformly at random with probability p. Let X_α^* denote the true aggregate binary interaction matrix, whereas X_α is partially observed, with unobserved entries set to zero.

We follow existing guarantees for matrix completion and define f as the singular-value decomposition of $p^{-1}X_\alpha$. That is, the rank-d CF model f is defined as

$$f(X_\alpha) = \arg\min_{U_\alpha,\Sigma_\alpha,V_\alpha} \left\|U_\alpha\Sigma_\alpha V_\alpha^T - p^{-1}X_\alpha\right\|_F^2 \quad \text{s.t. } U_\alpha^T U_\alpha = I_d,\ V_\alpha^T V_\alpha = I_d. \tag{2}$$

Chen et al. (2021) shows that f, as defined in Eq. (2), achieves near sample-size optimality for reliable matrix completion in that the lower-bound on p necessary for reliable completion matches the information-theoretic sampling requirement introduced in Candès & Tao (2010).

We evaluate model performance, as a consequence of a mixing vector $\boldsymbol{\alpha}$, based on the item singular vectors V_α. In the matrix completion setting, it is assumed that the complete interaction matrix for the target group is low rank and can be decomposed as $X_0^* = U_0^*\Sigma_0^* V_0^{*T}$. Thus, reducing the error between V_α and V_0^* is necessary for recovering the ground truth. However, to account for rotational symmetry in singular vectors, we define the item subspace error in terms of the projection matrix $V_\alpha V_\alpha^T$, which is invariant to rotations of V_α. Let the item-subspace discrepancy δ be

$$\delta = \left\|V_\alpha V_\alpha^T - V_0^* V_0^{*T}\right\|_F^2. \tag{3}$$

Matrix-completion data mixing entails identifying the augmentation group ratios that best help recover the target group's item subspace.

We define the evaluation metric $\mathcal{M}$ in terms of the high-probability radius of the item-subspace discrepancy. The discrepancy's stochasticity emerges from the randomness of missing values in X_α. Given a mixing vector $\boldsymbol{\alpha}$, let this radius be defined as:

$$R_\alpha = \inf\left\{r \geq 0 : P\left(\sqrt{\delta} > r\right) \lesssim m^{-1}\right\}. \tag{4}$$

Smaller values of radius R_α provide stronger guarantees on the discrepancy. Accordingly, we define the matrix-completion evaluation metric $\mathcal{M} = -R_\alpha$.

3.2.2. Stylized Matrix-Completion Setting

We detail a stylized setting under which the optimal matrix-completion data mix is well characterized.

Items are partitioned into k latent classes of equal size, and each user prefers exactly one item class. Using the movie example, under the stylized setting, item classes are genres, and a user watches movies only in their preferred genre. Under this construction, the item singular vectors encode item-class membership and are shared across groups. The matrix completion task reduces to: identifying which items belong to each item class, and which users prefer each item class.

Let $\Theta \in \mathbb{R}^{g\times k}$ denote group-level class proportions, where Θ_{jc} is the fraction of users in group j from class c, e.g., the percentage of users aged $25-35$ who prefer comedy movies. Let $\theta \in \Delta_{k-1}$ denote target-group class proportions. For any mixing vector $\boldsymbol{\alpha}$, define class prevalence as the number of users in the aggregate dataset X_α who prefer the item class. Informally, the prevalence is the popularity of the item class. Formally, define the prevalence of item class c as

$$\sigma_c(\boldsymbol{\alpha}) = n_0\theta_c + n_{\text{aug}}\sum_{j=1}^{g}\boldsymbol{\alpha}_j\Theta_{jc}, \tag{5}$$

where $\sigma_{\max}(\boldsymbol{\alpha}) = \max_c \sigma_c(\boldsymbol{\alpha})$ and $\sigma_{\min}(\boldsymbol{\alpha}) = \min_c \sigma_c(\boldsymbol{\alpha})$.

Our theorem and Eq. (5) rely on three assumptions: (i) without loss of generality, there are more items than users, (ii) the data sampling probability p is above a standard logarithmic sampling threshold, and (iii) augmentation sampling is class-proportional within each group. These three assumptions are formally defined in Appendix B.1.

3.2.3. Main Theoretical Result

Under the matrix-completion setting detailed in Section 3.2.2, the high-probability discrepancy bound is controlled by item-class prevalence disparity, as elaborated in Theorem 3.1. Thus, minimizing the bound requires minimizing $\sigma_{\max}(\boldsymbol{\alpha})/\sigma_{\min}^2(\boldsymbol{\alpha})$.

Theorem 3.1. *Under the stylized setting, for any mixing vector $\boldsymbol{\alpha}$:*

$$R_\alpha \lesssim \sqrt{\frac{k\,\sigma_{\max}(\boldsymbol{\alpha})\log m}{\sigma_{\min}(\boldsymbol{\alpha})^2\,p}}. \tag{6}$$

The proof for Theorem 3.1 is provided in Appendix B.1, and is closely connected to matrix-completion results linking the high-probability discrepancy radius to the ratio of the largest

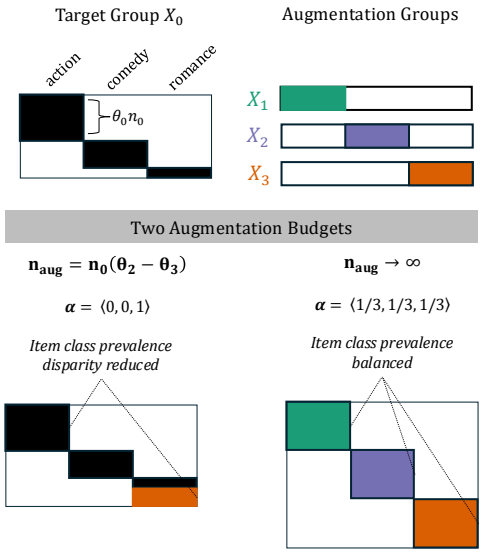


Figure 2. To provide intuition for our theoretical result, we show a stylized example where augmentation groups map to single genres. The target group has imbalanced genre prevalence, with the ratio between the most prevalent (action) and least prevalent (romance) genres determining the alignment bottleneck. At small budgets, the optimal mix concentrates augmentation on the least prevalent genre to reduce prevalence disparity; as $n_{\text{aug}} \to \infty$, the augmentation budget is sufficiently large to balance all genre prevalences, the optimal condition for target alignment in our matrix-completion setting.

and smallest singular values. Under the stylized setting, the item-class prevalences determine the singular values of X_α and, therefore, the condition number.

To illustrate the implication of Theorem 3.1 for the optimal mixing vector and budget ratios, we consider a further stylized setting. Figure 2 presents a simplified example in which augmentation groups map to item classes: for example, users in group 1 prefer only action movies. Per Theorem 3.1, the comparatively low prevalence of romance-preferring users ($\sigma_{\min}$) in the target group is the bottleneck for recovering the item singular vectors.

Figure 2 shows that the optimal mixing vector depends on the augmentation budget. At low augmentation budgets ($n_{\text{aug}} \in [0, n_0(\theta_2 - \theta_3)]$, the optimal augmentation includes only romance-preferring users from group 3 to minimize the prevalence disparity. However, as $n_{\text{aug}} \to \infty$, the augmentation budget is sufficient to eliminate item-class disparity ($\sigma_{\max} = \sigma_{\min}$), where the corresponding mixing vector is the constant vector. For clarity, the illustration of X^*_α in the $n_{\text{aug}} \to \infty$ setting does not depict the target group interactions, which vanish in the limit of the augmentation budget.

4. Evaluation Methodology

4.1. Data Mixing Approach

Optimizing Expected Loss Building on the data mixing problem definition specified in Section 3.1, suppose we have n users partitioned into a set $\mathcal{U}_0$ of n_0 target users and g augmentation sets, where group $j \in [g]$ has n_j users denoted by the set $\mathcal{U}_j$. Let $\ell(u)$ be the collaborative filtering loss for a user u. In our experiments, we define ℓ as the Bayesian Personalized Ranking (BPR) loss function (Rendle et al., 2009); however, the following approach is generic to the loss function.

Given a data mixing vector $\boldsymbol{\alpha} \in \Delta_{g-1}$ and a positive augmentation-budget ratio of α_{aug}, the aggregate data-mixed training dataset consists of target users $\mathcal{U}_0$ and $n_{\text{aug}} = \alpha_{\text{aug}} \sum_{j=1}^{g} n_j$ augmentation users. The α_{aug} parameter scales the relative size of the augmentation set relative to the target group, where $\alpha_{\text{aug}} = 1$ preserves the original ratio.

For an augmentation group j, let $\tilde{\mathcal{U}}_j$ be $\alpha_j n_{\text{aug}}$ users sampled with replacement from $\mathcal{U}_j$. The augmentation dataset is the union of all $\tilde{\mathcal{U}}_j$. Further, the probabilistic collaborative filtering loss $\mathcal{L}$ is

$$\mathcal{L}|\tilde{\mathcal{U}}_1, \ldots, \tilde{\mathcal{U}}_g = \sum_{u \in \mathcal{U}_0} \ell(u) + \left(\sum_{j \in [g]} \sum_{u' \in \tilde{\mathcal{U}}_j} \ell(u') \right). \quad (7)$$

To reduce the impact of sampling variance, in our experiments for a mixing vector $\boldsymbol{\alpha}$ and augmentation budget ratio of α_{aug}, we do not optimize the stochastic loss, but rather the *expected* loss

$$\mathbb{E}(\mathcal{L}) = \sum_{u \in \mathcal{U}_0} \ell(u) + n_{\text{aug}} \sum_{j \in [g]} \sum_{u' \in \mathcal{U}_j} \frac{\alpha_j}{n_j} \ell(u'). \quad (8)$$

The expected loss in Eq. 8 is equivalent to re-weighting a user in group j by $\frac{\alpha_j \alpha_{\text{aug}}(n - n_0)}{n_j}$.

Baseline Data Mixes We have three data-mix baselines: No Augmentation, Proportional, and Stratified. All are instances of Eq. (8). The No Augmentation baseline ($\alpha_{\text{aug}} = 0$) does not utilize any augmentation data; it is equivalent to the group-heterogeneity approach of training an isolated model for each target group. The Proportional baseline is equivalent to training on the data as is, where the mixing ratio for a group is proportional to n_j: $\alpha_{\text{aug}} = 1$, $\alpha_j = \frac{n_j}{\sum_{j'=1}^{g} n_{j'}}$ for all $j \in [g]$. The Stratified baseline weights all groups equally: $\alpha_{\text{aug}} = 1$, $\boldsymbol{\alpha} = \{g^{-1} | \forall j \in [g]\}$. We include the stratified baseline because of its demonstrated efficacy for LLM corpus data mixing (Chen et al., 2025) as well as its optimality in the limit in the stylized example from Figure 2.

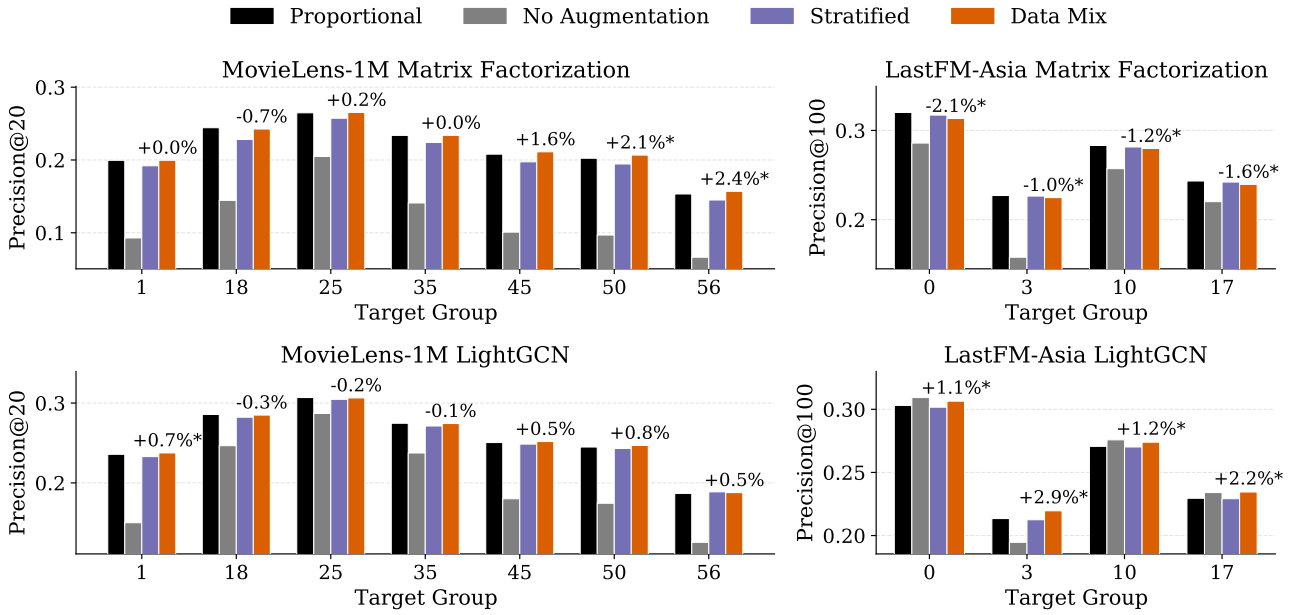


Figure 3. Data mixing does not reliably increase performance for MovieLens-1M and LastFM-Asia when trained with LightGCN or matrix factorization. Percentages refer to the change in Precision@k relative to the Proportional baseline. In most cases, except for LastFM-Asia trained with LightGCN, including augmentation data drastically increases performance over the No Augmentation baseline. However, the exact mixing ratio between Proportional, Stratified, and Data Mix does not substantially impact performance. Asterisks denote statistical significance from two-sided bootstrap CIs.

Sampling Data Mixing Vectors To evaluate data mixes, we sample α_{aug} and $\boldsymbol{\alpha}$, train a model using Eq. (8) on each mix, and select the mix that yields the highest cross-validation Recall@k. While α_{aug} can take on any positive value, we sample uniformly from $\alpha_{\text{aug}} \in [0.5, 2]$. For the mixing vector $\boldsymbol{\alpha}$, we design multiple sampling mechanisms, which are motivated by the baseline mixes. For example, we sample from the Dirichlet distribution $\text{Dir}(\gamma\mathbf{1})$ with $\gamma = 0.7$, which has an expected value equivalent to the Stratified baseline. We also sample from mixing vectors by adding a random noise vector ϵ to the Proportional mixing vector, omitting invalid mixing vectors containing negative ratios. Appendix C contains additional sampling details.

4.2. Evaluation Setup

Datasets We evaluate on two benchmark recommendation datasets: MovieLens-1M (Harper & Konstan, 2015) and LastFM-Asia (Rozemberczki & Sarkar, 2020). MovieLens-1M contains movie ratings by 6,040 users over 3,706 movies; target groups are defined by user age buckets (e.g., 1, 18, 25, 35, 45, 50, 56+). We encode the original explicit dataset as an implicit interaction dataset, where a user interacts with a movie by rating it. LastFM-Asia captures the listening history of 4,320 Asian users over a library of 7,832 songs on LastFM; target groups are defined by anonymized country codes (e.g., 0, 3, 10, 17). In both datasets, for each user, we split interactions into 80/20 train/test splits. We then construct cross-validation folds within the training interactions for mixing-vector selection: $k = 5$ folds for MovieLens-1M and $k = 4$ folds for LastFM-Asia.

Models We evaluate two CF models: LightGCN (He et al., 2020) and matrix factorization (MF). In our implementation, both models are trained with the same BPR objective and an ℓ_2 regularization on the user and item embeddings. Hyperparameter optimization for learning rate and regularization coefficient is described in Appendix C.

5. Results

Result 1: Data mixing does not reliably increase group alignment At a standard embedding dimension ($d = 64$), we observe a clear negative result: optimizing the mix does not consistently improve target-group alignment compared with training on the original group proportions. As shown in Figure 3, the Proportional baseline is often matched by Stratified and by the optimized data mix. The dominant effect is instead whether augmentation is used at all. Taken together, these results suggest that fine-grained reallocation across augmentation groups has limited additional impact. In Figure 3, an asterisk indicates that the difference relative to Proportional is statistically significant under a 95% bootstrap confidence interval (CI). Throughout, we report only Precision@k; the same qualitative patterns hold for Recall@k and NDCG@k.

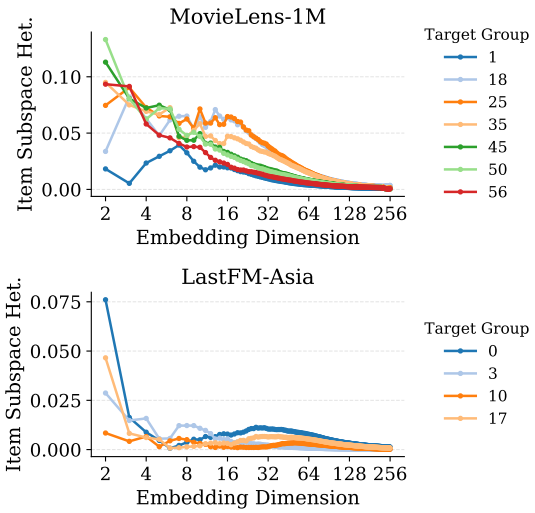


Figure 4. At larger embedding dimensions, the variation among the augmentation groups vanishes. For each target group, we calculate the discrepancy between the target and each augmentation item subspace. The y-axis represents the variance of these discrepancies for each target group. Smaller values indicate that augmentation groups are equally similar to the target group.

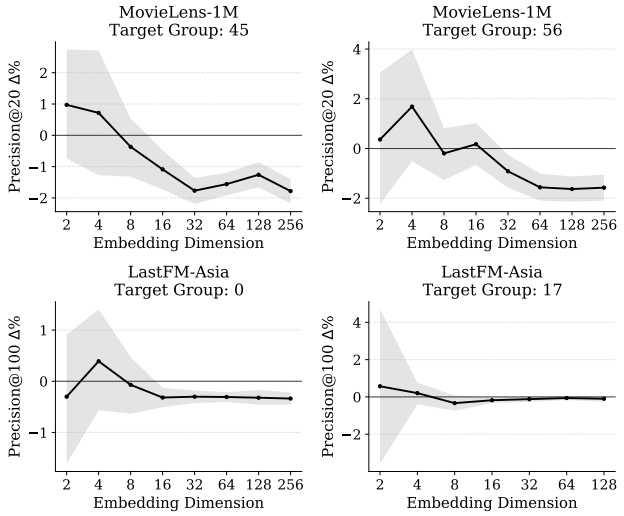


Figure 5. At lower embedding dimensions, Stratified outperforms Proportional for select target groups. We use the Stratified baseline as a heuristic lower bound on data-mixing performance, so the plots suggest that data mixing may also improve target alignment in low dimensions. The y-axes show the performance of Stratified relative to Proportional, with shaded bootstrapped CIs.

Result 2: Low group heterogeneity hinders data mixing. To explain why data mixing has a limited effect at $d = 64$, we analyze inter-group heterogeneity using a *dimension-specific* item-subspace discrepancy measure. For target groups i and j, let $V_i^*, V_j^* \in \mathbb{R}^{m \times d}$ contain the top-d right singular vectors of their fully observed interaction matrices. We define normalized discrepancy as

$$\Delta_d(i,j) = (2d)^{-1} \left\| V_i^* V_i^{*T} - V_j^* V_j^{*T} \right\|_F^2, \qquad (9)$$

where the denominator normalizes by the maximum possible squared Frobenius distance between two rank-d projection matrices. For each target group i, we compute $\{\Delta_d(i,j)\}_{j \neq i}$ to all augmentation groups and define heterogeneity as $\mathrm{Het}_d(i) = \mathrm{Var}_{j \neq i}[\Delta_d(i,j)]$.

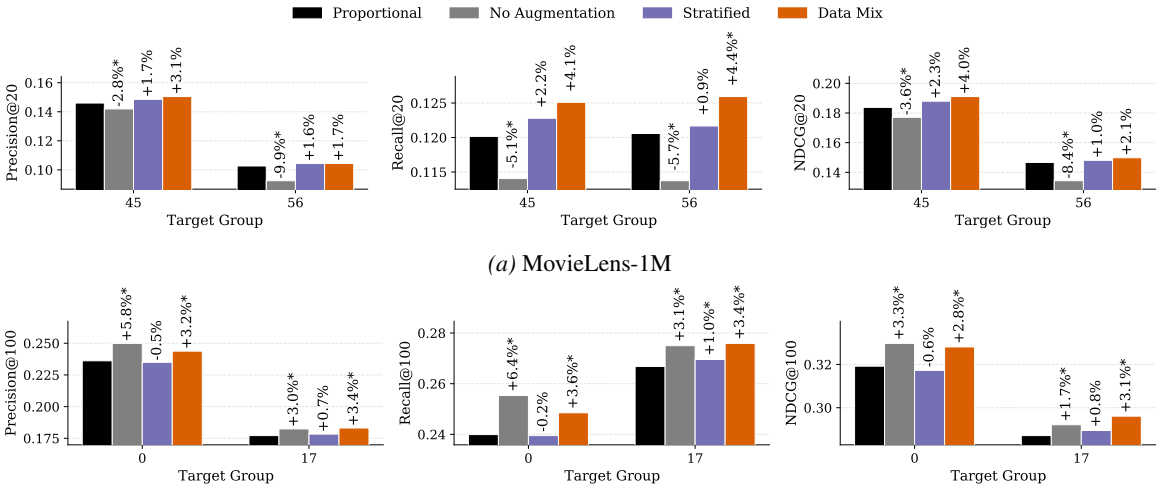
Figure 4 shows that $\mathrm{Het}_d(i)$ decreases as embedding dimension increases, indicating that augmentation groups have increasingly similar d-dimensional item-subspace geometries. When groups are nearly indistinguishable on this dimension-specific measure, changing the mixing weights offers little leverage for improving target-group alignment.

The Stratified vs. Proportional comparison across dimensions further supports this interpretation. We use Stratified as a heuristic lower bound on what reweighting can achieve when useful heterogeneity exists. In Figure 5, relative gains from Stratified appear primarily at low dimensions and diminish as the dimension grows, matching the decline in discrepancy variance. Thus, the data indicate that meaningful mixing gains require representational regimes where groups remain sufficiently distinct.

Result 3: At low embedding dimensions, data mixing leverages group heterogeneity for increased alignment. When we move to a very low-dimensional regime ($d = 4$), the effect of data mixing becomes consistent and substantial. Figure 6 shows that optimized mixes improve Precision@k by roughly 2–4% for selected target groups. All increases for LastFM-Asia are statistically significant. Relative to the $d = 64$ setting, this is a qualitative shift: mixing augmentation groups becomes reliably useful once group differences are sufficiently consequential.

Result 4: Mixing ratios are not necessarily correlated with group similarity. We analyze the learned low-dimensional mixes and find that larger mixing ratios are not simply assigned to the most similar augmentation groups. In Figure 7a, components with higher discrepancy can still receive high weight, while some near-target groups receive low weight. This pattern is consistent with our theory: improving target performance depends on how augmentation groups complement missing structure in the target data, not only on group similarity. Dissimilar groups can be valuable when they rebalance item-class prevalences and support target-group reconstruction.

Figure 7b also shows the optimized values of the budget ratio α_{aug}; in three cases, the ratio is close to 0.5, the lower bound of the sampled region. The relatively small value of the optimal budget ratio is consistent with the strong performance of the No Augmentation baseline; that is, in the low-dimensional setting, reducing the salience of augmentation data increases target-group alignment.



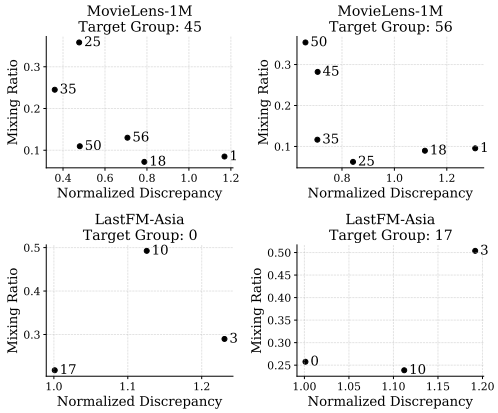
(a) MovieLens-1M

(b) LastFM-Asia

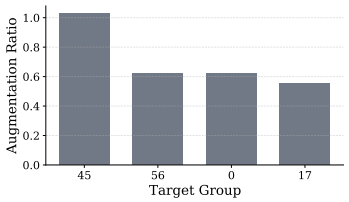
Figure 6. At very low embedding dimensions ($d = 4$), indeed, data mixing reliably improves group alignment relative to Proportional, increasing Precision@k by $2 - 4\%$ in most cases above. For LastFM-Asia, removing augmentation altogether frequently performs the best; however, among the augmented models, data mixing yields the greatest target-group alignment.

6. Conclusion

In general, it is difficult to determine whether and how data from heterogeneous groups can be leveraged to better infer preferences for a target group. Our work shows this problem is tractable in a CF setting, and the result challenges potential intuitions. For instance, initial data-oriented approaches to target-group preference alignment may entail filtering out all data outside the target group or including only data from similar augmentation groups. We show that, in a matrix completion setting, the optimal approach is to add data from complementary groups, specifically those that reduce item-class disparity. However, our empirical results demonstrate that at standard embedding dimensions in these benchmark CF datasets, group differences are not sufficiently pronounced for fine-grained mixing to matter. In these cases, the ideal strategy for data augmentation is to blindly add data without considering groups, i.e., all data are equally beneficial. Our findings demonstrate the need to collect CF datasets exhibiting greater group heterogeneity. Yet, in domains beyond CF where heterogeneous group preferences are captured, our results demonstrate the promise of data mixing for preference alignment.



(a) Mixing Vectors $\boldsymbol{\alpha}$



(b) Augmentation Budget Ratio α_{aug}

Figure 7. Mixing ratios for augmentation groups are not necessarily correlated with similarity to the target group. On top, for select target groups, we plot the optimal mixing vectors $\boldsymbol{\alpha}$. Each point represents one component of the optimal mixing vector. The x-axis is the discrepancy between the augmentation and target item subspaces, and the y-axis is the augmentation group's mixing ratio. For the target age group 56+ in MovieLens-1M, age groups 25 and 35 have similar item subspaces to that of 56+ (low discrepancy) but are not weighted highly in the optimal mixing vector. On the bottom, we also plot the optimal budget ratio α_{aug}.

References

Abdollahpouri, H., Mansoury, M., Burke, R., and Mobasher, B. The unfairness of popularity bias in recommendation, 2019.

Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B., and Malthouse, E. User-centered evaluation of popularity bias in recommender systems. In *UMAP '21*, pp. 119–129, New York, NY, USA, 2021. ACM.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023. doi: 10.1017/pan.2023.2.

Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., and Liang, P. S. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *NeurIPS '22*, pp. 3663–3678, 2022.

Bose, A., Curmei, M., Jiang, D. L., Morgenstern, J. H., Dean, S., Ratliff, L. J., and Fazel, M. Initializing services in interactive ML systems for diverse users. In *NeurIPS '24*, 2024.

Candès, E. J. and Tao, T. The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inf. Theor.*, 56(5):2053–2080, May 2010.

Chen, M. F., Hu, M. Y., Lourie, N., Cho, K., and Re, C. Aioli: A unified optimization framework for language model data mixing. In *ICLR '25*, 2025.

Chen, M. F., Murray, T., Heineman, D., Jordan, M., Hajishirzi, H., Ré, C., Soldaini, L., and Lo, K. Olmix: A framework for data mixing throughout lm development, 2026.

Chen, Y., Chi, Y., Fan, J., and Ma, C. Spectral methods for data science: A statistical perspective. *Found. Trends Mach. Learn.*, 14(5):566–806, October 2021. doi: 10.1561/2200000079.

Dean, S., Curmei, M., Ratliff, L., Morgenstern, J., and Fazel, M. Emergent specialization from participation dynamics and multi-learner retraining. In *AISTATS '24*, pp. 343–351, 2024.

Harper, F. M. and Konstan, J. A. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5 (4), December 2015.

He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., and Wang, M. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR'20*, pp. 639–648, New York, NY, USA, 2020. ACM.

Klimashevskaia, A., Jannach, D., Elahi, M., and Trattner, C. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*, 34 (5):1777–1834, November 2024. ISSN 1573-1391. doi: 10.1007/s11257-024-09406-0.

Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S., Bansal, H., Guha, E., Keh, S., Arora, K., Garg, S., Xin, R., Muennighoff, N., Heckel, R., Mercat, J., Chen, M. F., Gururangan, S., Wortsman, M., Albalak, A., Bitton, Y., Nezhurina, M., Abbas, A., Hsieh, C.-Y., Ghosh, D., Gardner, J., Kilian, M., Zhang, H., Shao, R., Pratt, S., Sanyal, S., Ilharco, G., Daras, G., Marathe, K., Gokaslan, A., Zhang, J., Chandu, K., Nguyen, T., Vasiljevic, I., Kakade, S., Song, S., Sanghavi, S., Faghri, F., Oh, S., Zettlemoyer, L., Lo, K., El-Nouby, A., Pouransari, H., Toshev, A., Wang, S., Groeneveld, D., Soldaini, L., Koh, P. W., Jitsev, J., Kollar, T., Dimakis, A. G., Carmon, Y., Dave, A., Schmidt, L., and Shankar, V. Datacomp-lm: In search of the next generation of training sets for language models. In *NeurIPS '24*, 2024.

Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI '09*, pp. 452–461, Arlington, Virginia, USA, 2009. AUAI Press.

Rozemberczki, B. and Sarkar, R. Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models. In *CIKM '20*, pp. 1325–1334. ACM, 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR '20*, 2020.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect?, 2023.

Shen, J. H., Raji, I. D., and Chen, I. Y. The data addition dilemma. In *MLHC '24*, volume 252. PMLR, 16–17 Aug 2024.

Su, J. and Dean, S. Learning from streaming data when users choose. In *ICML '24*, 2024.

Wang, A., Morgenstern, J., and Dickerson, J. P. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, March 2025a.

Wang, A., Phan, M., Ho, D. E., and Koyejo, S. Fairness through difference awareness: Measuring *Desired* group discrimination in LLMs. In *ACL '25*, pp. 6867–6893, 2025b.

Xie, S. M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., Liang, P., Le, Q. V., Ma, T., and Yu, A. W. Doremi: Optimizing data mixtures speeds up language model pretraining, 2023.

Zhang, L. H., Milli, S., Jusko, K. L., Smith, J., Amos, B., Bouaziz, W., Revel, M., Kussman, J., Sheynin, Y., Titus, L., Radharapu, B., Yu, J., Sarma, V., Rose, K., and Nickel, M. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *ICLR '26*, 2026.

A. Notation

Table 1 details the notation we utilize in this work.

Table 1. Notation

Symbol	Meaning
g	Number of augmentation groups (target group indexed by 0).
$X_0, X_1, \ldots, X_g$	Group-specific binary interaction matrices.
X_{aug}	Augmentation dataset sampled from augmentation groups.
X_α	Aggregate training matrix $\begin{bmatrix} X_0 \\ X_{\text{aug}}(\boldsymbol{\alpha}) \end{bmatrix}$.
$\boldsymbol{\alpha} \in \Delta_{g-1}$	Mixing vector over augmentation groups (simplex).
α_{aug}	Augmentation budget ratio controlling total augmentation size.
n_{aug}	Augmentation sample count, $n_{\text{aug}} = \alpha_{\text{aug}}(n - n_0)$.
$\mathcal{M}$	Target-group alignment metric (evaluation on target users only).
X^*_α	Fully observed (ground-truth) aggregate interaction matrix.
δ	Item-subspace discrepancy, $\|V_\alpha V_\alpha^T - V_0^* V_0^{*T}\|_F^2$.
R_α	High-probability radius for $\sqrt{\delta}$ under missingness randomness.
$\Theta \in \mathbb{R}^{g \times k}$	Augmentation-group class-proportion matrix; Θ_{jc} is class-c share in group j.
$\theta \in \Delta_{k-1}$	Target-group class-proportion vector.
$\sigma_c(\boldsymbol{\alpha})$	Prevalence of class c in the aggregate dataset under mix $\boldsymbol{\alpha}$.
$\sigma_{\max}(\boldsymbol{\alpha}), \sigma_{\min}(\boldsymbol{\alpha})$	Largest and smallest class prevalences.
α^{prop}	Proportional baseline mix vector.
γ	Dirichlet concentration parameter for the Dirichlet sampler.
s	Scale parameter for proportional-perturbation sampler.
a'_j	Group-j interaction scaling factor in the interaction-scaling sampler.
(μ, σ)	Log-normal parameters for interaction-scaling sampler.
$\Delta_d(i, j)$	Normalized dimension-specific discrepancy between groups i and j.
$\text{Het}_d(i)$	Group-heterogeneity score for target group i (variance of $\Delta_d(i, j)$ over $j \neq i$).

B. Proofs

B.1. Proof of Theorem 3.1

We restate the assumptions used by the theorem.

Assumption B.1. Let $n = n_0 + n_{\text{aug}}$ be the total number of users in the aggregated matrix. Assume, without loss of generality, $n \leq m$.

Assumption B.2. The observation probability satisfies

$$p \gtrsim \frac{\|\theta^{-1}\|_\infty^2 \log m}{nm}.$$

Assumption B.3. Within each augmentation group j, sampling follows class proportions: class c contributes $\alpha_j n_{\text{aug}} \Theta_{jc}$ users.

Proof. The argument follows a standard matrix-completion singular-subspace perturbation bound in operator norm of the form

$$\left\| V_\alpha V_\alpha^\top - V_\alpha^* V_\alpha^{*\top} \right\|_2 \lesssim \kappa \sqrt{\frac{\mu r \log m}{np}},$$

with failure probability $\mathcal{O}(m^{-1})$, where r is rank, μ is incoherence, and κ is condition number.

We compute these quantities for the stylized aggregate true matrix

$$X^*_\alpha = \begin{bmatrix} X_0^* \\ X^*_{\text{aug}}(\boldsymbol{\alpha}) \end{bmatrix},$$

where X_α is its partially observed counterpart used by the estimator in Eq. (2).

Rank. Because preferences are generated by k user/item classes, X_α^* has rank $r = k$.

Incoherence. Using

$$\mu = \max\left\{ \frac{n\|U\|_{2,\infty}^2}{r}, \frac{m\|V\|_{2,\infty}^2}{r} \right\},$$

the right-singular term is constant-order under equal item-class sizes. The left-singular term is controlled by the least represented class, yielding

$$\mu = \frac{n}{k\,\sigma_{\min}(\boldsymbol{\alpha})}.$$

Condition number. Each nonzero singular value scales as $\sqrt{\sigma_c(\boldsymbol{\alpha})m/k}$, so

$$\kappa = \frac{\sigma_{\max}(X_\alpha^*)}{\sigma_{\min}^+(X_\alpha^*)} = \sqrt{\frac{\sigma_{\max}(\boldsymbol{\alpha})}{\sigma_{\min}(\boldsymbol{\alpha})}}.$$

Substituting $r = k$, the incoherence scaling, and the condition-number scaling:

$$\kappa\sqrt{\frac{\mu r \log m}{np}} = \sqrt{\frac{\sigma_{\max}(\boldsymbol{\alpha})}{\sigma_{\min}(\boldsymbol{\alpha})}}\sqrt{\frac{\big(n/(k\sigma_{\min}(\boldsymbol{\alpha}))\big)\,k\,\log m}{np}} = \sqrt{\frac{\sigma_{\max}(\boldsymbol{\alpha})\log m}{\sigma_{\min}(\boldsymbol{\alpha})^2\,p}}.$$

Absorbing numerical constants gives

$$\left\|V_\alpha V_\alpha^\top - V_\alpha^* V_\alpha^{*\top}\right\|_2 \lesssim \sqrt{\frac{\sigma_{\max}(\boldsymbol{\alpha})\log m}{\sigma_{\min}(\boldsymbol{\alpha})^2\,p}}$$

with probability at least $1 - \mathcal{O}(m^{-1})$.

Under the stylized setting, all groups share the same right-singular subspace, so $V_\alpha^* V_\alpha^{*\top} = V_0^* V_0^{*T}$. Therefore, using

$$\delta = \left\|V_\alpha V_\alpha^\top - V_0^* V_0^{*T}\right\|_F^2,$$

and the fact that the difference of two rank-r projection matrices has rank at most $2r$, we have $\|A\|_F \le \sqrt{2r}\|A\|_2$ for $A = V_\alpha V_\alpha^\top - V_0^* V_0^{*T}$. Substituting $r = k$ yields

$$\sqrt{\delta} \;\lesssim\; \sqrt{\frac{k\,\sigma_{\max}(\boldsymbol{\alpha})\log m}{\sigma_{\min}(\boldsymbol{\alpha})^2\,p}} \quad \text{with probability } 1 - \mathcal{O}(m^{-1}).$$

Equivalently, the high-probability radius R_α in Eq. (4) is upper bounded by the right-hand side, so the bound is minimized by minimizing the right-hand side:

$$R_\alpha \lesssim \sqrt{\frac{k\,\sigma_{\max}(\boldsymbol{\alpha})\log m}{\sigma_{\min}(\boldsymbol{\alpha})^2\,p}},$$

which proves Theorem 3.1. □

C. Evaluation Methodology

Sampling Mixing Vectors and Augmentation Ratios For each source group, we construct candidate pairs $(\boldsymbol{\alpha}, \alpha_{\text{aug}})$, where $\boldsymbol{\alpha} \in \Delta_{g-1}$ controls relative augmentation-group allocation and $\alpha_{\text{aug}} > 0$ controls augmentation-set size. Each random trial samples $\alpha_{\text{aug}} \sim \text{Unif}(0.5, 2.0)$, and the mixing vector $\boldsymbol{\alpha}$ is drawn from one of three samplers. Trials are allocated evenly across samplers so that no single proposal mechanism dominates the candidate pool.

Dirichlet Sampler The first sampler draws

$$\boldsymbol{\alpha} \sim \text{Dir}(\gamma \mathbf{1}),$$

with concentration parameter $\gamma < 1$ (in our experiments, $\gamma = 0.7$). This sampler is centered at the stratified baseline $\mathbb{E}[\alpha_j] = 1/g$, but with sparse realizations that frequently emphasize a subset of augmentation groups. It therefore explores non-uniform departures from equal weighting while preserving simplex constraints by construction.

Proportional-Perturbation Sampler Let α^{prop} be the proportional baseline, with components proportional to group sizes. We sample a perturbation vector $\delta \in \mathbb{R}^g$ by drawing i.i.d. $\delta_j \sim \text{Unif}(0,1)$, centering it to have zero mean, and scaling by a factor s:

$$\tilde{\boldsymbol{\alpha}} = \alpha^{\text{prop}} + s\big(\delta - \bar{\delta}\mathbf{1}\big).$$

We then project to the nonnegative simplex by clipping negative entries to zero and renormalizing. This sampler is explicitly centered on proportional mixing and performs local exploration around it, yielding nearby alternatives that preserve total mass.

Interaction Scaling Sampler The third sampler is formulated as interaction-level re-scaling by group. For each group j, draw a positive scaling factor

$$a'_j \sim \text{LogNormal}(\mu, \sigma^2),$$

where $a'_j = 1$ implies no re-scaling of interactions from group j, and $a'_j = 2$ implies doubling the weight of each interaction from group j. We then define the mixing vector via size-adjusted normalization:

$$\alpha_j \propto n_j a'_j,$$

Equivalently,

$$\alpha_j = \frac{n_j a'_j}{\sum_{\ell=1}^{g} n_\ell a'_\ell}.$$

In our experiments, we use $\mu = 0$ and $\sigma = 0.5$, so the median scaling factor is 1 while the right tail produces occasional larger scaling factors. Thus, the sampler is centered at no re-scaling but still explores heavy-tailed multiplicative upweighting of selected groups.

Hyperparameter Optimization For each dataset and each model family (LightGCN and MF), we perform k-fold cross-validation over a Cartesian grid of learning rates and regularization coefficients (decay), then choose the setting with the best mean validation Recall@k. Across all runs, the selected regularization coefficient is 10^{-1}. The selected learning rate is 10^{-2} in all settings except MovieLens-1M with LightGCN, where the selected learning rate is 10^{-3}. Training uses a BPR batch size of 10^6 triplets per update, where each triplet consists of a user, one observed (positive) item, and one unobserved (negative) item.

Compute Resources All experiments were run on machines equipped with NVIDIA V100 GPUs with 32GB of memory. The data-mixing validation and final testing pipeline requires approximately 6 hours per target group.

Dataset Metadata Table 2 summarizes the metadata for the target groups in both datasets.

Table 2. Target group metadata

Dataset	Target Group	Num. Users	Num. Interactions
MovieLens-1M	1	222	2.72×10^4
	18	1103	1.84×10^5
	25	2096	3.96×10^5
	35	1193	1.99×10^5
	45	550	8.36×10^4
	50	496	7.25×10^4
	56	380	3.88×10^4
	Total	6040	1.00×10^6
LastFM-Asia	0	1060	5.25×10^5
	3	502	1.75×10^5
	10	1265	5.11×10^5
	17	1493	5.36×10^5
	Total	4320	1.75×10^6