UNDERSTANDING AND MITIGATING BIAS IN MLLM-BASED RECOMMENDER SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) are increasingly adopted as backbones for recommender systems, giving rise to MLLM-based recommender systems (MLLM-RSs). However, bias issues in these systems remain largely underexplored, undermining stakeholders' trust and interests. Two key challenges to solve: How to (1) *identify* the source of residing bias and (2) *mitigate* bias without costly retraining. We propose TRACE (**T**racing and **R**emoving bias via c**A**usal **C**oncept **E**ffects), a novel causally grounded framework to identify and mitigate bias in MLLM-RSs. First, it traces the propagation of bias through three connected levels: from the training sample level, to the modality-specific concept level that captures higher-level factors potentially driving bias, and finally to the recommendation outcome level that models how these concepts causally affect bias. We devise a causally-mediated *chain influence* to connect the three levels, attributing measured bias effects in recommendation outcomes to individual training samples. Second, the resulting chain influence directly guides three efficient bias *mitigation* operators: a removal operator suppresses high-influence bias-driving samples along identified bias directions; a corrective operator reinforces genuine preference signals from corrective pairs within the same directions; and a composed operator integrates both to balance bias removal and preference preservation during parameter updates. Experiments on three datasets and two MLLM backbones under different debiasing cases demonstrate that our proposed methods achieve a superior balance between recommendation fairness and accuracy.[1]

1 INTRODUCTION

Multimodal large language models (MLLMs) (Wu et al., 2023; 2025; Wei et al., 2026; Li et al., 2026b) are actively reshaping recommender systems (Liu et al., 2024b; Ye et al., 2025; Zhang et al., 2025; Wang et al., 2025a;b; 2026; Lei et al., 2026). By incorporating users' interaction history and multimodal item content, e.g., visual images and text descriptions, MLLM-based recommender systems (MLLM-RSs) unify preference modelling and content understanding in a generative backbone.

At the same time, bias remains one of the most persistent pathologies of unfair recommendations (Li et al., 2023b; Wang et al., 2023; Li et al., 2024a) that could hurt users' trust and stakeholders' interests. For example, popularity bias (Zhang et al., 2021) systematically over-exposes head items and starves the long tail, and disparities across user-attribute groups (Ying et al., 2023) raise inequality concerns. MLLM-RSs may further *exaggerate* the problem: bias can be inherited from large-scale MLLM pre-training, injected by supervised fine-tuning (SFT) on skewed interaction data, and amplified by the model's semantic priors (Zhao et al., 2025b). Figure 1 makes this concrete on our own backbones. As shown in Figure 1 (a), training data includes a large proportion of popular items; by training on this, the backbone MLLM-RS model reproduces and re-expresses that skew in its recommendation results. More importantly, such skew mismatches users' preferences, as shown in Figure 1 (b), users whose ground-truth preference lies outside the popular items receive worse serves. Although bias issues have been investigated in the domain of conventional RS (Wang et al., 2023; Li et al., 2024a) or LLM-based RS (Jiang et al., 2024; Gao et al., 2025; Zhao et al., 2025b), it remains largely underexplored for MLLM-RS. Simply adapting existing debiasing methods is

[1]This work was conducted in a company subject to IP regulations. Code will be released shortly after internal approval.

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.

insufficient or even infeasible due to the distinct recommendation paradigm of MLLM-based generation (Li et al., 2024b), additional modality-specific and cross-modal bias interactions (Shang et al., 2024), and the high cost of retraining large backbones (Li et al., 2026a). These limitations call for a framework that can both identify how bias arises inside MLLM-RSs and translate this discovery into targeted mitigation. Developing such a framework raises two coupled challenges.

Challenge 1: How to identify and explain where bias comes from? Observing a disparity in recommendation outcomes reveals that an MLLM-RS is biased, but not *why* the bias arises. A useful diagnosis should trace the bias pathway from the training samples that induce biased behaviour, through the internal concepts in which such information is encoded, to its eventual effect on recommendation outcomes. This is particularly challenging in MLLM-RSs because item and preference information are represented through multiple modalities. For example, bias may be associated with visual signals, textual semantics, historical interaction patterns, or their interactions. Existing methods based on post-hoc saliency (Li et al., 2026a) or predefined causal assumptions (Zhang et al., 2021; Wei et al., 2021) do not provide such an identification ability.

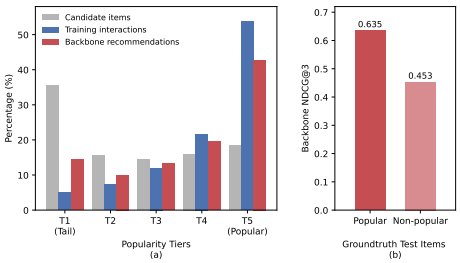


Figure 1: The bias problem in MLLM-RSs.

Challenge 2: How to mitigate the identified bias efficiently while preserving preference modelling? Existing debiasing strategies commonly retrain the model with fairness-aware objectives (Jiang et al., 2024; Li et al., 2025), which is costly for large MLLM backbones and may alter model behaviour beyond the identified source of bias. More importantly, simply suppressing bias-related information can inadvertently remove genuine preference signals that correlate with the same information. The key challenge is therefore how to turn the identified bias pathway into a targeted model update. It should suppress the effects responsible for biased recommendations while retaining the preference information essential for recommendation utility.

We address these two challenges with TRACE (**T**racing and **R**emoving bias via c**A**usal **C**oncept **E**ffects), a causally grounded framework that turns bias identification into targeted mitigation, as illustrated in Figure 2. To address **Challenge 1**, it traces bias through three connected levels. It starts from individual training samples (L_1), identifies modality-specific latent concepts that capture higher-level factors potentially driving bias (L_2), and measures how these concepts causally affect the bias outcome (L_3). These levels are not isolated but connected with sample-to-concept influence and concept-to-bias causal effects, forming a *causally-mediated chain influence*. This provides a unified bias pathway to identify both the samples that drive bias and the internal directions through which their effects are expressed. To address **Challenge 2**, TRACE directly turns this identified pathway into an efficient model update without retraining the backbone. We design two complementary mitigation operators and their composition. The *removal* operator suppresses the effects of the most bias-driving samples specifically along the identified bias directions, while the *corrective* operator reinforces genuine preference signals within the same directions to avoid improving fairness at the expense of personalisation. The *composed* operator provides a controllable balance between bias removal and preference preservation within a one-step model update. Our contributions are as follows:

- We propose TRACE, a causally grounded framework for jointly identifying, explaining, and mitigating bias in MLLM-RSs. It provides a connected three-level view from training samples, through modality-specific latent concepts, to biased recommendation outcomes, exposing how bias is encoded and propagated across multimodal information.
- We connect the three levels through a novel *causally-mediated chain influence*, which attributes measured bias to individual training samples by combining their influence on internal concepts with the causal effects of concepts on bias. Building on the identified bias pathway, we further devise complementary *removal* and *corrective* operators to suppress bias-driving effects while preserving genuine preference signals without retraining the backbone.
- We instantiate TRACE on two MLLM backbones and three recommendation datasets under different types of bias. Extensive experiments demonstrate its effectiveness in identifying meaningful bias pathways and achieving a superior balance between recommendation fairness and accuracy.

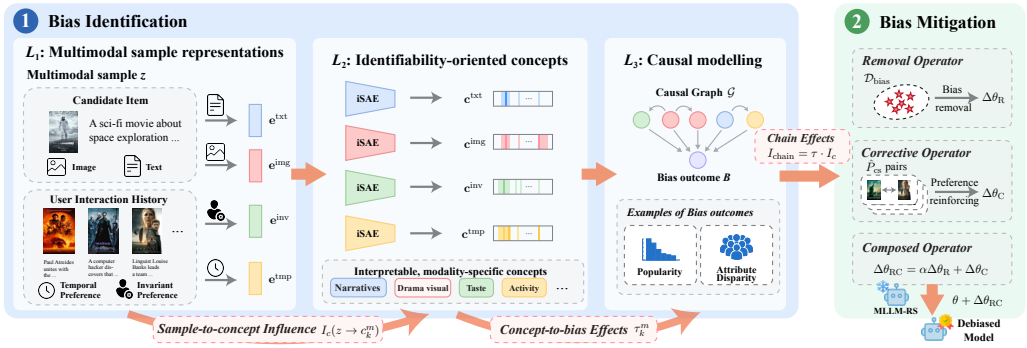


Figure 2: The framework of TRACE. It first identifies bias pathway at three connected levels: from multimodal representations to extracted latent concepts, and then the causal relationship of concepts to bias. It then composes chain influence to compute a composed subspace-projected update to mitigate bias efficiently while preserving preference modelling.

2 Related Work

MLLM-based RS. Casting recommendation as language modelling began with unified text-to-text formulations (Geng et al., 2022; Wu et al., 2024; Lin et al., 2025) and instruction tuning of LLMs on interaction data (Bao et al., 2023; 2025), with follow-up work aligning behaviour patterns to the token space (Liao et al., 2024). MLLM-RSs further extend this paradigm with multimodal item content, either through multimodal personalised prompting (Geng et al., 2023; Liu et al., 2024c; Zhou et al., 2025), SFT (Wang et al., 2025a; Liu et al., 2024b; Wang et al., 2025b) or reinforcement learning (Wang et al., 2026). We follow this paradigm (Ye et al., 2025; Wang et al., 2026), where MLLMs model user preferences from multimodal interaction histories and recommend candidate items, and focus on analysing and correcting their biased behaviours.

Bias and fairness in RS. Bias is a long-standing challenge in RSs (Chen et al., 2023a; Wang et al., 2023), harming equality and interests for users, items, and providers. Based on stakeholders involved, RS bias includes item-side (e.g., popularity bias (Abdollahpouri et al., 2019) where popular items are over-recommended), user-side (Zhao et al., 2025a) (e.g., attribute bias (Deldjoo & Noia, 2025) where different user groups are discriminately treated), and multifaceted bias (Mansoury et al., 2026) with multiple co-existing biases. Conventional debiasing methods employ reweighting (Li et al., 2023a), regularisation (Abdollahpouri et al., 2017), causal interventions (Zhang et al., 2021), or adversarial learning (Ganhör et al., 2022), with recent extensions to LLM-based RSs (Jiang et al., 2024; Li et al., 2026a). However, to the best of our knowledge, no debiasing framework has been proposed for MLLM-RS models, and simple adaptation of existing methods for conventional or LLM-based RSs is insufficient or even infeasible. First, existing debiasing strategies usually rely on predefined causality (Zhang et al., 2021; Wei et al., 2021) or sample-level attribution (Li et al., 2026a; Xu et al., 2022; Kiyohara et al., 2022), and can hardly reveal the internal mechanism of how bias arises in MLLM-RSs, limiting their ability to reveal how bias arises through multimodal interactions and temporal preference dynamics. Second, most debiasing methods are model-specific (Shang et al., 2024) and retraining-based (Jiang et al., 2024; Gao et al., 2025), making them infeasible to adapt to MLLM-RSs due to the systematically different recommendation mechanism and large-scale parameters. This calls for a specially designed debiasing framework for MLLM-RSs.

3 Preliminary

MLLM-based RS. Let $\mathcal{U}$ and $\mathcal{V}$ denote the user and item sets. Each item $v = (\mathbf{x}^{\text{img}}, \mathbf{x}^{\text{txt}}) \in \mathcal{V}$ includes visual content and a textual description. For user $u \in \mathcal{U}$, the chronologically ordered interaction history is $\mathbf{S}_u = [\mathbf{s}_u^1, \cdots, \mathbf{s}_u^t]$ with $\mathbf{s}_u^j \in \mathcal{V}$. Given a candidate item v_c with interaction label $y_u^c \in \{0, 1\}$, an MLLM-RS backbone g_θ receives $(\mathbf{S}_u, v_c)$ in a multimodal prompt (see Appendix E.1.5) and predicts whether the user will interact with v_c by answering `Yes` or `No`. We thus define the decision score:

$$f_\theta(z) = \log \sum_{w \in \mathcal{V}_{\text{Yes}}} \exp(o_\theta(w \mid \mathbf{S}_u, v_c)) - \log \sum_{w \in \mathcal{V}_{\text{No}}} \exp(o_\theta(w \mid \mathbf{S}_u, v_c)), \tag{1}$$

where $o_\theta(\cdot \mid \cdot)$ is the next-token logit, $\mathcal{V}_{\text{Yes}}$ and $\mathcal{V}_{\text{No}}$ collect tokeniser variants of the two answers. The interaction probability is given by $g_\theta(1 \mid \mathbf{S}_u, v_c) = \sigma(f_\theta(z))$, and candidates within the same evaluation group are ranked by $f_\theta(z)$. A training sample is $z = (\mathbf{S}_u, v_c, y_u^c)$, and the backbone is obtained by supervised fine-tuning (SFT) with a LoRA adapter $\theta \in \mathbb{R}^p$ (Ye et al., 2025; Wang et al., 2026; Hu et al., 2022) on $\mathcal{D} = \{z_i\}_{i=1}^n$ with the per-sample loss $\ell(z;\theta) = -\log g_\theta(y_u^c \mid \boldsymbol{S}_u, v_c)$ and empirical risk $R(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(z_i;\theta)$. The fine-tuned adapter $\hat{\theta}$ is then fixed.

Bias setting. We consider a bias setting specified by a measurable bias-relevant variable ρ, such as item popularity for popularity bias or a user attribute for attribute bias. Let $B_\rho(g_\theta; z)$ denote the corresponding bias-related outcome produced by the recommender. The dataset-level bias measure is obtained by an aggregation functional

$$\mathcal{B}_\rho(g_\theta) \;=\; \Phi_\rho\Big(\big\{B_\rho(g_\theta; z) : z \in \mathcal{D}\big\}\Big). \tag{2}$$

It forms an adaptive measure for various types of bias considered. For example, in the case of popularity bias, where tail items are less exposed, $\rho(v)$ represents item popularity and B_ρ can measure the popularity or tail membership of exposed items, whose aggregation yields recommendation-popularity or tail-exposure metrics. For user-attribute bias, where users from different attribute groups receive unequal treatment, B_ρ represents recommendation utility associated with each user and Φ_ρ measures disparity of its conditional expectations across attribute groups.

Our goal is twofold: (1) *identification*, tracing observed bias from recommendation outcomes through interpretable multimodal concepts to influential training samples; and (2) *mitigation*, using the identified pathway to edit the fine-tuned adapter without retraining the MLLM.

4 TRACE FRAMEWORK

4.1 OVERVIEW

As illustrated in Figure 2, our framework includes two key components to respectively *identify* and *mitigate* the bias in MLLM-RSs. First, it connects three levels of bias analysis: from sample-level multimodal representations in L_1, it extracts interpretable modality-specific concepts in L_2, then models how these concepts causally affect the bias outcome in L_3. By connecting them together within a casually-mediated chain influence, it quantifies the interventional effects from samples to bias. Second, this identified pathway is used to compute a removal operator for bias mitigation, along with a corrective operator to reinforce observed preference evidence. Their composition balances bias mitigation and preference preservation through a single parameter update.

4.2 BIAS IDENTIFICATION

L_1: Multimodal sample representations. As the input level, L_1 is built on the extracted multimodal representations. We follow Li et al. (2026a) to uniformly sample a probe set $\mathcal{D}_{\text{probe}} \subset \mathcal{D}$, where each sample has four views: $\mathbf{e}^m = \phi_{\hat{\theta}}^m(z)$, $m \in \mathcal{M} = \{\text{txt}, \text{img}, \text{inv}, \text{tmp}\}$, covering textual and visual content, time-invariant preference and temporal preference, respectively. Let $\mathbf{h}[p] \in \mathbb{R}^h$ be the final layer hidden state at prompt position $p \in P$. The candidate text and image representations are pooled over their corresponding token spans with $\mathbf{e}^{\text{txt}} = \frac{1}{|P_{\text{txt}}|}\sum_{p\in P_{\text{txt}}} \mathbf{h}[p]$ and $\mathbf{e}^{\text{img}} = \frac{1}{|P_{\text{img}}|}\sum_{p\in P_{\text{img}}} \mathbf{h}[p]$. For preference, let $\phi_{\hat{\theta}}(\mathbf{S})$ denote the final-position representation of a history prompt. We construct two complementary views: $\mathbf{e}^{\text{inv}} = \frac{1}{|\mathcal{O}|}\sum_{\pi\in\mathcal{O}} \phi_{\hat{\theta}}(\mathbf{S}_u^\pi)$, where $\mathcal{O}$ contains forward and reverse orderings of the same history, and $\mathbf{e}^{\text{tmp}}(u, j) = \phi_{\hat{\theta}}(\mathbf{S}_u^{1:j})$, $j \in \mathcal{J} = \{1, \lceil t/2 \rceil, t\}$, which captures preference evolution at different stages of the interaction history. Prompt templates are given in Appendix E.1.5.

L_2: Identifiability-oriented multimodal concepts. For each modality m, we aim to extract a high-level concept vector $\mathbf{c}^m \in \mathbb{R}^{K_m}$ from $\mathbf{e}^m$ to identify and explain the bias pathway within the MLLM-RS mechanism. Sparse autoencoders can decompose LLM activations into interpretable features (Huben et al., 2024), but their latents are not identifiable, such that the semantic meaning of an individual SAE feature is not uniquely determined across equivalent representations. We therefore learn concepts with an *identifiability-oriented* SAE (iSAE) that integrates the SAE

objective with an iVAE-style conditional prior (Khemakhem et al., 2020; Mita et al., 2021) anchored on a bias-relevant auxiliary variable a (e.g., the candidate popularity tier for popularity debiasing; the sensitive-attribute group for attribute debiasing). Per modality, an encoder produces $q_\psi(\mathbf{c} \mid \mathbf{e}) = \mathcal{N}\big(\mu_\psi(\mathbf{e}), \mathrm{diag}(\sigma^2_\psi(\mathbf{e}))\big)$ over $\mathbf{c} \in \mathbb{R}^{K_m}$, a decoder D_ξ reconstructs $\mathbf{e}$, and a conditional prior $p_\lambda(\mathbf{c} \mid a)$ is a factorised Gaussian whose parameters depend on a. The resulting objective is

$$\mathcal{J}^m_{\mathrm{iSAE}} = \underbrace{\mathbb{E}\big[\|\mathbf{e} - D_\xi(\tilde{\mathbf{c}})\|_2^2\big]}_{\text{reconstruction}} + \beta \underbrace{\mathbb{E}\big[\|\tilde{\mathbf{c}}\|_1\big]}_{\text{sparsity}} + \gamma \underbrace{\mathbb{E}\big[D_{\mathrm{KL}}\big(q_\psi(\mathbf{c} \mid \mathbf{e}) \,\|\, p_\lambda(\mathbf{c} \mid a)\big)\big]}_{\text{conditional prior (identifiability)}}, \tag{3}$$

with $\tilde{\mathbf{c}} \sim q_\psi(\mathbf{c}|\mathbf{e})$ via reparameterization. Downstream stages use the posterior mean $\hat{\mathbf{c}}^m(z) = \mu_\psi\big(\phi^m_{\hat{\theta}}(z)\big)$. We further encourage consistency across history orderings for invariant preference:

$$\mathcal{L}_{\mathrm{inv}} = \tfrac{1}{|\mathcal{O}|} \textstyle\sum_{\pi \in \mathcal{O}} \big\|\mu_\psi\big(\phi_{\hat{\theta}}(\mathbf{S}^\pi_u)\big) - \bar{\mu}_u\big\|_2^2, \qquad \bar{\mu}_u = \tfrac{1}{|\mathcal{O}|} \textstyle\sum_{\pi \in \mathcal{O}} \mu_\psi\big(\phi_{\hat{\theta}}(\mathbf{S}^\pi_u)\big). \tag{4}$$

The objective for $m = \mathrm{inv}$ thus includes $w\,\mathcal{L}_{\mathrm{inv}}$. For the temporal view, we encode the concept trajectory $\big\{\hat{\mathbf{c}}^{\mathrm{tmp}}(u,j)\big\}_{j \in \mathcal{J}}$ for later analysis of temporal bias propagation.

Remark 1. The auxiliary-conditioned prior provides an identifiability-motivated structural constraint by encouraging concept coordinates whose conditional distributions vary systematically with the observed auxiliary variable. Formal iVAE identifiability requires sufficiently rich variation of the auxiliary-conditioned latent distribution (Khemakhem et al., 2020), which is difficult to satisfy with the limited number of discrete auxiliary values available in practical recommendation data. We therefore do not claim a formal identifiability guarantee for the learned concepts in this setting. Nevertheless, our empirical results in Appendix E.8 show that the learned concepts capture meaningful bias-related structure and support stable measurement of bias-inducing directions.

L_3: interventional effects of concepts on bias. To explore how the learned concepts contribute to the bias outcome B, we compute a structural causal model $\mathcal{G}$ using a score-based causal discovery algorithm. However, graph discovery over $\sum_m K_m$ concepts is both expensive and noisy as many codes carry no bias signal and only add spurious parents. We therefore perform a simple concept selection: a ridge regression of the bias outcome B on all concepts gives coefficients $\hat{w}^m_k$ and $\tilde{s}^m_k = \hat{w}^m_k(q^h_k - q^l_k)$ to measure the linear change in B from moving the concept from its lower to higher activated samples, where q^h_k and q^l_k represent the upper (above 75%) and lower (below 25%) quantiles. The K' concepts with the largest $|\tilde{s}^m_k|$ per modality are selected for causal discovery.

Assumption 1 (Causal-effect model)**.** *The selected concepts and bias outcome B admit an acyclic SCM in which each concept follows $c_j = \mu_j + \sum_{i \in \mathrm{pa}(j)} w_{ji} c_i + \varepsilon_j$, and*

$$B = \mu_B + \sum_{i \in \mathrm{pa}(B)} w_{Bi} c_i + \sum_{(i,i') \in \mathcal{I}_\times} \gamma_{ii'} c_i c_{i'} + \varepsilon_B, \tag{5}$$

where $\mathcal{I}_\times$ contains cross-modal pairs, the noises are mutually independent, and the fitted DAG is correctly specified over the selected variables.

4.3 Chaining the levels: from samples to bias

The preceding levels identify which concepts affect bias. We next trace the effects back to samples.

Assumption 2 (Local influence regularity)**.** *The empirical objective R and sample loss $\ell(z;\theta)$ are twice continuously differentiable in a neighbourhood of the fine-tuned solution $\hat{\theta}$, where $\nabla_\theta R(\hat{\theta}) = 0$ and $\mathbf{H} = \nabla^2_\theta R(\hat{\theta})$ is positive definite.*

Link $L_1 \to L_2$. Let $\bar{c}^m_k(\theta) = \frac{1}{|\mathcal{D}_{\mathrm{probe}}|} \sum_{z' \in \mathcal{D}_{\mathrm{probe}}} \hat{c}^m_k(z';\theta)$ be the probe-averaged activation of concept c^m_k. For $\hat{\theta}_\epsilon = \arg\min_\theta R(\theta) + \epsilon \ell(z;\theta)$, standard influence differentiation gives

$$I_c(z \to c^m_k) := \frac{d\,\bar{c}^m_k(\hat{\theta}_\epsilon)}{d\epsilon}\Big|_{\epsilon=0} = -\nabla_\theta \bar{c}^m_k(\hat{\theta})^\top \mathbf{H}^{-1} \nabla_\theta \ell(z;\hat{\theta}). \tag{6}$$

Link $L_2 \to L_3$. For concept c^m_k, define $F^m_k(x) = \mathbb{E}[B \mid \mathrm{do}(c^m_k = x)]$. Under Assumption 1, descendants of c^m_k are affine in the intervention value x, while B contains only linear and bilinear terms. Hence $F^m_k(x)$ is at most quadratic. For standardised concepts, we define centred causal effect

$$\tau^m_k := F^m_k\big(\tfrac{1}{2}\big) - F^m_k\big(-\tfrac{1}{2}\big) = \frac{\partial F^m_k(x)}{\partial x}\Big|_{x=0}, \tag{7}$$

The quantity τ_k^m, computed by g-computation through the fitted SCM (Robins, 1986), captures both direct and concept-mediated effects on bias. If c_k^m has no directed path to B, then $\tau_k^m = 0$.

Integration. The two links above naturally derive a centred causal bias surrogate $\mathcal{J}_{\text{cen}}(\theta) = \sum_{m\in\mathcal{M}} \sum_{k=1}^{K_m} \tau_k^m \, \bar{c}_k^m(\theta)$. We define the *causally-mediated chain influence* of training sample z as the infinitesimal effect of upweighting z on this surrogate.

Proposition 1 (Centered causal chain influence)**.** *Under Assumptions 1 and 2,*

$$I_{\text{chain}}(z) := \frac{d}{d\epsilon} \mathcal{J}_{\text{cen}}(\hat{\theta}_\epsilon)\Big|_{\epsilon=0} = \sum_{m\in\mathcal{M}} \sum_{k=1}^{K_m} \tau_k^m \, I_c(z \to c_k^m). \tag{8}$$

Hence, under the standard infinitesimal-deletion approximation $\epsilon = -1/n$, $\mathcal{J}_{\text{cen}}(\hat{\theta}_{-1/n}) - \mathcal{J}_{\text{cen}}(\hat{\theta}) = -\frac{1}{n} I_{\text{chain}}(z) + o(n^{-1})$. Moreover, for any non-zero componentwise rescaling and permutation, $\hat{\boldsymbol{c}} = \boldsymbol{\Lambda\Pi c}$, $I_c(z \to \hat{c}_k) = \lambda_k I_c(z \to c_{\pi(k)})$, $\hat{\tau}_k = \lambda_k^{-1} \tau_{\pi(k)}$. Therefore, both $\tau_k I_c(z \to c_k)$ and $\tau_k \nabla_\theta \bar{c}_k^m$ are invariant up to permutation.

Thus, $I_{\text{chain}}(z)$ has an intuitive enterpretation: $L_1 \to L_2$ measures how a sample changes a concept, while $L_2 \to L_3$ measures the causal sensitivity of bias to that concept. Their product attributes the resulting first-order bias effect to the sample.

4.4 Bias Mitigation

Bias removal. Identification provides us both a *target set* and a *direction*. The target set $\mathcal{D}_{\text{bias}} = \text{Top-}K_b\{|I_{\text{chain}}(z)|\}$, $z \in \mathcal{D}$, includes the samples that most strongly drive bias-relevant concepts. The direction is obtained from the bias-concept gradients:

$$\mathbf{G}_{\text{bias}} = \left[\tau_k^m \, \nabla_\theta \bar{c}_k^m(\hat{\theta}) \right]_{(m,k)} \in \mathbb{R}^{p \times K'}, \quad \mathbf{P}_{\text{bias}} = \mathbf{U}_r \mathbf{U}_r^\top, \tag{9}$$

where $\mathbf{G}_{\text{bias}} = \mathbf{U\Sigma V}^\top$ is a singular value decomposition and $\mathbf{U}_r$ collects the top-r left singular vectors. The mitigation update is the projected first-order unlearning step (Koh & Liang, 2017)

$$\Delta\theta_{\text{R}} \;=\; \eta \, \frac{K_b}{n} \, \mathbf{P}_{\text{bias}} \, \mathbf{H}^{-1} \bar{\mathbf{g}}, \qquad \bar{\mathbf{g}} = \frac{1}{K_b} \sum_{z \in \mathcal{D}_{\text{bias}}} \nabla_\theta \ell(z;\theta), \qquad \theta' = \hat{\theta} + \Delta\theta_{\text{R}}, \tag{10}$$

where $\eta > 0$ is a step size. This **removal operator** (**TRACE-R**) is closed-form given the influence state, performed on the LoRA adapter for efficiency.

Preference-preserving correction. Removing bias-driving directions alone may potentially harm the essential preference modelling (as evidenced in Table 1). We therefore construct a second operator using observation pairs in which genuine preference contradicts the biased behaviour:

$$\mathcal{P}_{\text{cs}} \;=\; \big\{(z^+, z^-),\; y^+ = 1,\; y^- = 0,\; \rho(v^+) < \rho(v^-)\big\}. \tag{11}$$

For example, in the case of popularity bias, it finds pairs where the user's *observed* choice z^+ is less popular than the negative item z^-. Ranking these pairs by causal salience $|I_{\text{chain}}(z^+)| + |I_{\text{chain}}(z^-)|$ and keeping the top $K_b/2$ ones as $\hat{\mathcal{P}}_{\text{cs}}$. Their preference direction is

$$\mathbf{g}_{\text{pref}} \;=\; \frac{2}{K_b} \sum_{(z^+,z^-)\in\hat{\mathcal{P}}_{\text{cs}}} \nabla_\theta \big(f_{\hat{\theta}}(z^-) - f_{\hat{\theta}}(z^+) \big). \tag{12}$$

We project it into the same causal bias subspace and define the **corrective operator** (**TRACE-C**):

$$\mathbf{u} \;=\; \mathbf{P}_{\text{bias}} \, \mathbf{g}_{\text{pref}}, \qquad \Delta\theta_{\text{C}} \;=\; -\, D \, \frac{\mathbf{u}}{\|\mathbf{u}\|_2}, \qquad \text{whenever } \mathbf{u} \neq \mathbf{0}. \tag{13}$$

Composed operator. The removal and corrective operators act within the same identified causal subspace and can therefore be combined as

$$\Delta\theta_{\text{RC}}(\alpha) \;=\; \alpha \, \Delta\theta_{\text{R}} \;+\; \Delta\theta_{\text{C}}, \quad \alpha > 0, \quad \theta' = \hat{\theta} + \Delta\theta_{\text{RC}}, \tag{14}$$

with $\alpha \geq 0$. The complete procedure is summarised in Algorithm A.1. This composition possesses the following key properties, with proof given in Appendix D.2:

Proposition 2 (Properties of the mitigation operators). *Let* $\mathcal{S}_{\text{bias}} = \text{span}(\mathbf{U}_r)$, $\Delta\theta_{\text{R}} = \mathbf{P}_{\text{bias}}\Delta_{\text{full}}$, $\mathbf{u} = \mathbf{P}_{\text{bias}}\mathbf{g}_{\text{pref}} \neq \mathbf{0}$, $\Delta\theta_{\text{C}} = -D\frac{\mathbf{u}}{\|\mathbf{u}\|_2}$, $\Delta\theta_{\text{RC}} = \alpha\Delta\theta_{\text{R}} + \Delta\theta_{\text{C}}$. *Then:*

1. ***Removal.*** *$\Delta\theta_{\text{R}}$ is the orthogonal projection of Δ_{full} onto $\mathcal{S}_{\text{bias}}$. Thus $\|\Delta\theta_{\text{R}}\|_2 \leq \|\Delta_{\text{full}}\|_2$. For any functional F, $\nabla F(\hat{\theta}) \in \mathcal{S}_{\text{bias}}$, the projected and full updates have the same first-order effect.*
2. ***Correction.*** *The corrective operator has fixed displacement $\|\Delta\theta_{\text{C}}\|_2 = D$ and is a guaranteed first-order descent direction for the selected preference objective.*
3. ***Composition.*** *All three updates belong to $\mathcal{S}_{\text{bias}}$. Consequently, any functional whose gradient is orthogonal to this subspace is unchanged to first order. Moreover, writing $c = \mathbf{g}_{\text{pref}}^{\top}\Delta\theta_{\text{R}} = \mathbf{u}^{\top}\Delta\theta_{\text{R}}$, the composed update preserves preference-descent whenever $\alpha c < D\|\mathbf{u}\|_2$.*

Bias-agnostic instantiation and co-existing biases. Our framework can be instantiated to address various types of bias through a measurable bias-related variable ρ. For *item-side bias* (e.g., popularity bias), ρ varies across candidate items, such as the popularity of each item. Eq. 11 can therefore be directly constructed from samples with different ρ values to reinforce observed preferences that contradict the biased tendency. For *group-level bias* (e.g., user-attribute bias), ρ identifies the group membership of each user. The same principle is applied by first identifying the worst-served group and then reinforcing its revealed preferences against negative candidates. For multiple co-existing biases, each bias variable ρ_b induces its own identified causal subspace and mitigation update $\Delta\theta_{\rho_b}$, and the resulting updates can be jointly applied as $\Delta\theta = \sum_b \Delta\theta_{\rho_b}$.

5 Experiment

5.1 Experimental Setup

Datasets. To evaluate the debiasing and recommendation performance, we adopt three widely used multimodal recommendation datasets, including **MicroLens** (Ye et al., 2025; Wang et al., 2026), **Amazon Baby** (Ye et al., 2025), and **MovieLens-1M** (Wang et al., 2026). The detailed dataset statistics and preprocessing steps are described in Appendix E.1.1.

Baselines. For a comprehensive comparison, we adopt several representative and state-of-the-art baselines from different classes. Note that, since there are no established debiasing methods targeting MLLM-RSs directly, we adopted a wide range of debiasing methods for both LLM-based and conventional RSs in our settings. (1) For popularity debiasing, we include **RW**, **RR**, **RWRR** (Jiang et al., 2024), **SPRec** (Gao et al., 2025), **FairLRM** (Luo et al., 2026), **FUDLR** (Li et al., 2026a), **PDA** (Zhang et al., 2021), and **MACR** (Wei et al., 2021). (2) For attribute debiasing, we adopt **CFP** (Hua et al., 2024), **LLM-FairRec** (Li et al., 2025), and **FUDLR** (Li et al., 2026a).

This baseline collection covers a wide range of categories, i.e., reweighting and reranking strategies (RW, RR, RWRR), data augmentation (LLM-FairRec) and unlearning (FUDLR), reinforcement learning (SPRec), and prompt adjustment (CFP, FairLRM). We also include the **backbone** as a reference for the original performance. More details on baselines are provided in Appendix E.1.2.

Metrics. We evaluate the debiased performance from three aspects. **1. Accuracy:** **AUC** and **NDCG** are widely used for measuring the accuracy of recommendation results. **2. Fairness:** for popularity debiasing, we measure **TailP** (the proportion of tail items in the recommendation list) and **NPCG** (the NDCG score of test samples with non-popular target items); for attribute debiasing, we measure **ΔNDCG** (the NDCG gap across user groups). **3. Trade-off:** the harmonic mean $H(\cdot,\cdot)$ is used to measure the balance between accuracy and fairness. Specifically, $F_{\text{pop}}@K = H(\text{NDCG}@K, \text{TailP}@K)$ is used for popularity debiasing evaluation, and $F_{\text{attr}}@K = H(\text{NDCG}@K, 1 - 10 \cdot \Delta\text{NDCG}@K)$ is used for attribute debiasing evaluation, accounting for the different directions and scales of value. More details are given in Appendix E.1.3.

Implementation details. We initialise TRACE with two representative MLLM backbones with Qwen2.5-VL-3B-Instruct (Wang et al., 2024; 2026) and LLaVA-NeXT-Mistral-7B (Liu et al., 2024a). All experiments are conducted on a server with 3 NVIDIA A100 GPUs. Implementation details are further given in Appendix E.1.4.

Table 1: Popularity debiasing performance (%) and running time (h) on three datasets. The best results are highlighted in bold, and the second best are underlined. The backbone results are listed as a reference to the original performance. * indicates the improvement is significant with $p < 0.05$.

Dataset	Baseline	AUC ↑	NDCG@3 ↑	NDCG@5 ↑	TailP@3 ↑	TailP@5 ↑	NPCG@3 ↑	NPCG@5 ↑	F_{pop}@3 ↑	F_{pop}@5 ↑	Time ↓
MicroLens	Backbone	86.81 ± 0.03	55.69 ± 0.28	60.57 ± 0.26	14.54 ± 0.29	17.99 ± 0.19	44.46 ± 0.77	49.94 ± 0.38	23.06 ± 0.36	27.74 ± 0.23	74.49
	RW	85.41 ± 0.35	53.48 ± 1.15	58.69 ± 1.08	16.32 ± 0.03	19.26 ± 0.41	44.91 ± 0.58	50.44 ± 0.68	25.00 ± 0.09	29.00 ± 0.34	96.99
	RR	**86.88 ± 0.03**	55.48 ± 0.32	60.40 ± 0.28	14.40 ± 0.28	17.81 ± 0.19	43.94 ± 0.89	49.55 ± 0.43	22.86 ± 0.35	27.50 ± 0.23	2.59
	RWRR	85.88 ± 0.09	54.01 ± 0.40	59.27 ± 0.45	16.12 ± 0.23	19.00 ± 0.39	45.11 ± 0.42	50.67 ± 0.19	24.83 ± 0.25	28.77 ± 0.40	97.15
	SPRec	85.78 ± 0.41	54.27 ± 1.49	59.33 ± 1.18	15.10 ± 1.84	18.69 ± 1.25	42.94 ± 0.41	48.61 ± 0.24	23.56 ± 2.16	28.40 ± 1.33	13.38
	FairLRM	86.09 ± 0.22	52.47 ± 1.24	55.93 ± 1.80	14.83 ± 0.55	19.40 ± 0.65	42.97 ± 0.21	48.59 ± 0.40	23.12 ± 0.57	28.79 ± 0.50	2.23
	FUDLR	86.81 ± 0.02	55.48 ± 0.06	60.38 ± 0.19	14.66 ± 0.28	18.02 ± 0.08	44.31 ± 0.59	49.94 ± 0.24	23.19 ± 0.35	27.76 ± 0.11	14.77
	PDA	86.62 ± 0.04	54.59 ± 0.40	59.73 ± 0.46	15.27 ± 0.30	18.49 ± 0.14	44.16 ± 0.83	49.85 ± 0.67	23.87 ± 0.35	28.23 ± 0.16	**2.19**
	MACR	86.61 ± 0.04	54.61 ± 0.41	59.77 ± 0.42	15.21 ± 0.29	18.46 ± 0.15	44.20 ± 0.72	49.92 ± 0.57	23.79 ± 0.33	28.21 ± 0.19	2.49
	TRACE-R	84.66 ± 0.08	49.49 ± 0.13	55.17 ± 0.30	**17.17 ± 0.58***	**20.12 ± 0.30***	39.43 ± 0.57	45.51 ± 0.65	**25.49 ± 0.65**	**29.48 ± 0.35***	6.27
	TRACE-C	86.84 ± 0.03	**55.94 ± 0.47***	**60.76 ± 0.43***	15.61 ± 0.13	18.67 ± 0.12	45.33 ± 0.65	50.78 ± 0.48	24.41 ± 0.12	28.56 ± 0.16	6.33
	TRACE-RC	86.82 ± 0.02	55.88 ± 0.52	60.69 ± 0.53	15.57 ± 0.08	18.65 ± 0.06	**45.40 ± 0.45***	**50.82 ± 0.43**	24.36 ± 0.15	28.53 ± 0.01	6.47
Baby	Backbone	83.07 ± 0.04	45.82 ± 0.16	52.21 ± 0.17	11.47 ± 0.20	13.34 ± 0.28	26.91 ± 0.27	32.29 ± 0.38	18.34 ± 0.24	21.25 ± 0.34	139.08
	RW	84.32 ± 0.03	**50.73 ± 0.30**	**56.23 ± 0.10**	10.87 ± 0.11	13.33 ± 0.31	26.30 ± 0.82	32.18 ± 0.65	17.90 ± 0.17	21.55 ± 0.40	141.89
	RR	83.14 ± 0.04	45.76 ± 0.17	52.15 ± 0.14	11.35 ± 0.22	13.22 ± 0.28	26.70 ± 0.32	32.08 ± 0.36	18.19 ± 0.28	21.09 ± 0.34	2.97
	RWRR	**84.40 ± 0.03**	50.60 ± 0.33	56.09 ± 0.15	10.62 ± 0.09	13.09 ± 0.37	25.83 ± 0.89	31.69 ± 0.77	17.55 ± 0.13	21.22 ± 0.49	139.86
	SPRec	81.29 ± 0.06	44.06 ± 0.25	50.14 ± 0.15	10.97 ± 0.25	13.29 ± 0.22	26.07 ± 0.60	31.11 ± 0.50	17.56 ± 0.32	21.01 ± 0.27	12.87
	FairLRM	83.66 ± 0.06	47.06 ± 0.23	51.64 ± 0.10	9.69 ± 0.18	12.60 ± 0.11	24.55 ± 0.38	31.35 ± 0.05	16.07 ± 0.24	20.26 ± 0.14	2.23
	FUDLR	83.31 ± 0.08	47.00 ± 0.13	52.99 ± 0.29	11.10 ± 0.16	13.06 ± 0.17	26.77 ± 0.42	31.96 ± 0.39	17.96 ± 0.21	20.95 ± 0.25	19.27
	PDA	82.91 ± 0.04	44.79 ± 0.30	51.35 ± 0.19	11.74 ± 0.24	13.54 ± 0.27	26.69 ± 0.24	31.99 ± 0.35	18.60 ± 0.27	21.42 ± 0.32	**2.19**
	MACR	82.96 ± 0.04	44.96 ± 0.32	51.50 ± 0.16	11.68 ± 0.24	13.50 ± 0.27	26.68 ± 0.27	32.01 ± 0.35	18.54 ± 0.27	21.39 ± 0.33	2.50
	TRACE-R	83.03 ± 0.05	45.40 ± 0.29	51.87 ± 0.22	11.81 ± 0.13	13.48 ± 0.17	27.20 ± 0.40	32.48 ± 0.33	18.74 ± 0.16	21.40 ± 0.20	6.08
	TRACE-C	82.82 ± 0.05	45.13 ± 0.08	51.50 ± 0.07	**12.08 ± 0.11***	**13.81 ± 0.27***	26.89 ± 0.55	32.04 ± 0.52	**19.06 ± 0.15***	**21.78 ± 0.34**	6.18
	TRACE-RC	82.95 ± 0.02	44.98 ± 0.36	51.58 ± 0.40	12.05 ± 0.09	13.56 ± 0.34	**27.47 ± 0.60***	**32.77 ± 0.37***	19.01 ± 0.14	21.47 ± 0.42	6.31
MovieLens	Backbone	85.34 ± 0.14	54.57 ± 1.11	59.11 ± 0.69	12.90 ± 0.69	16.37 ± 0.18	39.92 ± 0.49	45.03 ± 0.50	20.85 ± 0.81	25.64 ± 0.16	9.44
	RW	85.35 ± 0.15	54.22 ± 0.36	**59.00 ± 0.83**	11.15 ± 0.29	14.94 ± 0.23	37.53 ± 0.27	43.16 ± 1.25	18.49 ± 0.38	23.85 ± 0.23	11.14
	RR	85.41 ± 0.14	**54.32 ± 1.02**	58.88 ± 0.53	12.75 ± 0.72	16.17 ± 0.13	39.32 ± 0.52	44.47 ± 0.18	20.64 ± 0.86	25.37 ± 0.12	0.82
	RWRR	**85.43 ± 0.15**	54.02 ± 0.38	58.83 ± 0.87	10.98 ± 0.38	14.75 ± 0.31	37.04 ± 0.34	42.75 ± 1.34	18.25 ± 0.50	23.58 ± 0.34	11.14
	SPRec	83.88 ± 0.19	48.61 ± 0.23	54.21 ± 0.52	15.23 ± 0.39	18.07 ± 0.27	36.29 ± 0.20	41.77 ± 0.37	23.19 ± 0.42	27.10 ± 0.30	9.85
	FairLRM	85.34 ± 0.14	54.57 ± 1.11	59.11 ± 0.69	12.90 ± 0.69	16.37 ± 0.18	39.92 ± 0.49	45.03 ± 0.50	21.10 ± 0.96	25.64 ± 0.22	0.42
	FUDLR	85.02 ± 0.14	53.66 ± 0.91	58.25 ± 1.01	13.65 ± 0.80	16.96 ± 0.07	40.18 ± 0.56	45.40 ± 0.88	21.75 ± 0.93	26.28 ± 0.09	2.69
	PDA	85.25 ± 0.14	53.66 ± 1.04	58.29 ± 0.73	13.21 ± 0.65	16.52 ± 0.25	39.54 ± 0.48	44.60 ± 0.45	21.19 ± 0.75	25.75 ± 0.23	**0.41**
	MACR	85.22 ± 0.14	53.61 ± 1.05	58.22 ± 0.71	13.28 ± 0.69	16.66 ± 0.22	39.64 ± 0.41	44.65 ± 0.46	21.28 ± 0.79	25.90 ± 0.19	0.68
	TRACE-R	84.08 ± 0.17	50.28 ± 0.73	55.75 ± 1.29	14.40 ± 0.44	18.02 ± 0.33	37.47 ± 0.60	43.20 ± 1.39	22.39 ± 0.48	27.24 ± 0.33	4.03
	TRACE-C	85.13 ± 0.15	53.97 ± 0.89	58.34 ± 0.65	13.26 ± 0.43	16.80 ± 0.23	39.83 ± 0.62	44.44 ± 0.66	21.29 ± 0.48	26.08 ± 0.21	4.02
	TRACE-RC	84.31 ± 0.16	53.17 ± 0.40	58.24 ± 0.22	**15.32 ± 0.80**	**18.32 ± 0.90**	**41.15 ± 0.06***	**46.67 ± 0.80***	**23.79 ± 1.00***	**27.87 ± 1.07**	4.23

5.2 BIAS MITIGATION

Popularity bias. We first evaluate TRACE on popularity bias, a representative item-side bias, across three datasets. Several key observations can be drawn from Table 1. (1) Our proposed methods achieve a clear fairness improvement while largely preserving recommendation accuracy, leading to improved balance between debiasing and utility preservation. (2) The three variants exhibit distinct behaviours consistent with their designs. TRACE-R more aggressively increases long-tail exposure, but incurs larger accuracy degradation and can reduce recommendation performance on non-popular items, as observed on MicroLens and MovieLens. In contrast, TRACE-C reinforces preference signals within the identified debiasing directions, resulting in improved utility preservation. Combining the two, TRACE-RC achieves moderate-to-high improvements in tail exposure while attaining the strongest NPCG among the compared methods across all three datasets. Qualitative results in Appendix E.8 also confirm that meaningful concepts are discovered to support targeted debiasing. (3) Among the baselines, reweighting-based retrain methods (RW and RWRR) excel at maintaining and improving recommendation accuracy, but may fail at improving exposure and performance for long-tail items, e.g., on Baby and MovieLens datasets. Another training-based method, SPRec, serves as a strong baseline for lifting tail exposure but does not guarantee accurate recommendations. Similarly, unlearning-based (FUDLR) and causal inference-based methods (PDA and MACR) promote tail exposure and preserve a stable recommendation utility. However, they fail to improve NPCG relative to the backbone, which is more difficult than purely increasing tail exposure. This may be attributed to the overlooking of preference modelling within the debiasing space. In addition, TRACE achieves substantially higher time efficiency than retraining-based methods while remaining comparable to efficient inference-based methods. Overall, our TRACE-RC stands out as a strong and balanced candidate for popularity debiasing in MLLM-RSs.

Attribute bias. As another common bias in RSs, user attribute bias directly harms user experience by providing discriminative performance for users from different attribute groups. We evaluate the performance of attribute bias mitigation on the gender attribute in the MovieLens dataset. From Table 2, we find that our methods significantly improve user fairness by mitigating the performance gap across user groups (ΔNDCG) while preserving reasonable accuracy. In contrast, baselines such as CFP and FUDLR demonstrate limited fairness improvements. This may be attributed to the insufficient discovery and mitigation of the bias sources in MLLM-RSs.

Other bias. We also evaluate **co-existing bias** in Appendix E.3 and **modality bias** in Appendix E.4.

Table 2: User attribute debiasing on the gender attribute in the MovieLens dataset. The best results are highlighted in bold, and the second best are underlined. The backbone results are listed as a reference to the original performance. * indicates the improvement is significant with $p < 0.05$.

Attribute	Baseline	AUC ↑	NDCG@3 ↑	NDCG@5 ↑	ΔNDCG@3 ↓	ΔNDCG@5 ↓	F_{attr}@3 ↑	F_{attr}@5 ↑	Time ↓
Gender	Backbone	85.34 ± 0.14	54.57 ± 1.11	59.11 ± 0.69	2.36 ± 1.39	1.92 ± 0.67	63.31 ± 4.92	68.21 ± 2.72	9.44
	CFP	85.01 ± 0.21	52.31 ± 1.31	57.39 ± 1.12	3.79 ± 2.40	1.90 ± 1.94	55.01 ± 11.48	66.48 ± 7.30	7.10
	LLM-FairRec	84.84 ± 0.05	52.36 ± 0.05	57.27 ± 0.73	2.39 ± 1.32	1.88 ± 0.46	61.75 ± 4.14	67.13 ± 1.11	12.60
	FUDLR	**85.37 ± 0.14**	**54.03 ± 0.68**	**58.42 ± 0.25**	3.92 ± 1.02	2.62 ± 0.75	56.94 ± 3.89	65.11 ± 2.69	4.76
	TRACE-R	84.08 ± 0.17	50.28 ± 0.73	55.75 ± 1.29	2.00 ± 1.80	1.61 ± 1.41	61.22 ± 6.06	66.70 ± 5.16	4.01
	TRACE-C	85.14 ± 0.13	53.95 ± 1.12	58.21 ± 0.88	1.53 ± 1.00	1.31 ± 1.26	65.72 ± 2.58	69.43 ± 3.77	**4.00**
	TRACE-RC	84.88 ± 0.12	53.47 ± 1.00	58.05 ± 0.77	**0.56 ± 0.15***	**1.18 ± 0.36***	**68.25 ± 0.43***	**69.97 ± 0.56***	4.20

5.3 Ablation Study

Table 3: Ablation study of popularity debiasing on MovieLens.

Variant	AUC	NDCG@3	NDCG@5	TailP@3	TailP@5	NPCG@3	NPCG@5
Identification							
AE	84.71	51.87	57.42	14.57	16.99	37.93	43.74
iAE	84.97	53.16	57.79	13.74	17.09	38.75	43.41
SAE	84.28	51.95	57.66	16.23	**19.90**	38.63	45.00
w/o SCM	83.51	51.55	57.15	**16.34**	19.14	38.47	44.46
Mitigation							
TRACE-R	84.08	50.28	55.75	14.40	18.02	37.47	43.20
TRACE-C	**85.13**	**53.97**	**58.34**	13.26	16.80	39.83	44.44
TRACE-RC	84.31	53.17	58.24	15.32	18.32	**41.15**	**46.67**

To validate the effectiveness of our design, we perform ablation on each component of TRACE-RC. In bias identification, (1) replacing iSAE with a vanilla autoencoder (**AE**) removes both sparsity and identifiability constraints, and thus the obtained directions fail to distinguish fairness-oriented gradients; (2) removing sparsity (**iAE**) maintains reasonable accuracy yet fails to improve fairness for tail items (NPCG). (3) replacing iSAE with a sparse autoencoder (**SAE**) removes the conditional prior guidance and weakens accuracy preservation; (4) replacing SCM (**w/o SCM**) with a ridge regression on the bias outcome further weakens preference preserving, leading to a high accuracy cost during debiasing. In bias mitigation, excluding either R or C operator shifts performance towards either preference preservation or debiasing, whereas the full variant, TRACE-RC, achieves a reasonable balance with higher accuracy for tail items. Additional ablation results are given in Appendix E.6.

5.4 Parameter Analysis

We investigate the impacts of key parameters of TRACE-RC, including α, D, β, and γ, on popularity debiasing on MicroLens in Figure 3. (1) **Removal weight** α controls the strength of the bias removal operator. As α increases, tail exposure generally rises while accuracy drops. (2) **Correlation weight** D imposes preference reinforcement. However, the performance becomes less stable at large values of D, with the optimal value lying around 0.25. (3) **Sparsity weight** β controls the density of concetps. Insufficient sparsity leaves the concepts entangled and weakens the projected direction, whereas excessive sparsity may remove concept directions that carry exposure-related information. The optimal value lies near the middle of the stable region. (4) **Conditional-prior weight** γ sets what edit is identified. Weakening the prior decreases the exposure, while over-strengthening it turns the position into an accuracy-oriented edit. The optimal value is around $\gamma = 10^{-2}$. Additional parameter analysis is provided in Appendix E.7.

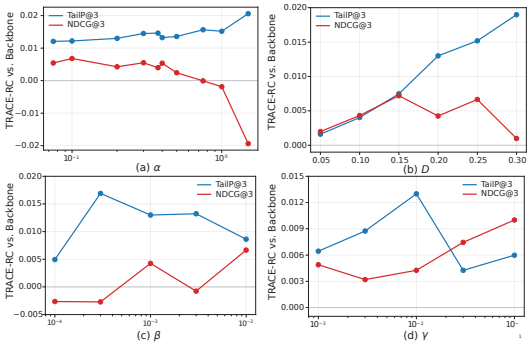


Figure 3: Parameter sensitivity on MicroLens.

6 Conclusion

We presented TRACE, a causally grounded framework for identifying and mitigating bias in MLLM-RSs. It connects multimodal representations, latent concepts, and bias outcomes through a three-level identification process, yielding a chain influence that traces bias back to the training data. Building on this diagnosis, it performs efficient bias mitigation together with a preference-aware corrective step, avoiding costly retraining while preserving utility. Experiments show that TRACE effectively improves fairness while maintaining competitive recommendation accuracy.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing, grammar correction, and code debugging. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, supporting qualitative and thematic data analysis, assisting with translation, cleaning and reformatting datasets. All AI-assisted text and code suggestions were carefully reviewed and tested by the authors. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies bias identification and mitigation in multimodal large-language-model-based recommender systems (MLLM-RSs), with the goal of better understanding how undesirable recommendation bias arises and mitigating unfairness to users, items, or providers. Our experiments focus on popularity bias and user-attribute bias, with additional analyses of co-existing and modality bias. The study is conducted on existing research benchmarks and does not involve the collection of new personal data or direct interaction with human participants. Sensitive attributes, when available in a dataset, are used only for bias analysis and evaluation. In practical systems, their use requires careful consideration of user consent, privacy, applicable regulations, and the social context in which the recommender operates. Human oversight and application-specific evaluation are therefore necessary before applying the proposed mitigation procedure in high-stakes or user-oriented systems.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the datasets, preprocessing procedures, baseline methods, evaluation protocols, and implementation settings in Appendix E. Experiments are conducted on three widely used multimodal recommendation benchmarks, MicroLens, Amazon Baby, and MovieLens-1M, using two representative, open-sourced MLLM backbones, Qwen2.5-VL-3B-Instruct and LLaVA-NeXT-Mistral-7B. We report recommendation utility and bias-related metrics together with their trade-offs, and compare against representative methods covering reweighting, reranking, data augmentation, reinforcement learning, prompt adjustment, causal debiasing, and unlearning. Experiments are conducted on a server equipped with three NVIDIA A100 GPUs. We also analyse the impact of hyperparameters in Appendix E.7 to facilitate reproduction of the complete pipeline. Upon internal approval, we will release the source code and datasets required to reproduce the reported experiments.

REFERENCES

Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys*, pp. 42–46, 2017.

Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Managing popularity bias in recommender systems with personalized re-ranking. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference*, pp. 413–418, 2019.

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B. Grosse. If influence functions are the answer, then what is the question? In *Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2022.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the ACM Conference on Recommender Systems, RecSys*, pp. 1007–1014, 2023.

Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yanchen Luo, Chong Chen, Fuli Feng, and Qi Tian. A bi-step grounding paradigm for large language models in recommendation systems. *ACM Transactions on Recommender Systems*, 3(4):53:1–53:27, 2025.

Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):67:1–67:39, 2023a.

Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Fast model debias with machine unlearning. In *Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems, NeurIPS*, 2023b.

Weixin Chen, Li Chen, Yongxin Ni, and Yuhan Zhao. Causality-inspired fair representation learning for multimodal recommendation. *ACM Transactions on Information Systems*, 43(6):153:1–153:29, 2025.

Yashar Deldjoo and Tommaso Di Noia. Cfairllm: Consumer fairness evaluation in large-language model recommender system. *ACM Transactions on Intelligent Systems and Technology*, 16(6): 142:1–142:31, 2025.

Christian Ganhör, David Penz, Navid Rekabsaz, Oleg Lesota, and Markus Schedl. Unlearning protected user attributes in recommendations with adversarial training. In *The International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, pp. 2142–2147, 2022.

Chongming Gao, Ruijun Chen, Shuai Yuan, Kexin Huang, Yuanqing Yu, and Xiangnan He. Sprec: Self-play to debias llm-based recommendation. In *Proceedings of the ACM on Web Conference, WWW*, pp. 5075–5084, 2025.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, pp. 299–315, 2022. ISBN 9781450392785. doi: 10.1145/3523227.3546767.

Shijie Geng, Juntao Tan, Shuchang Liu, Zuohui Fu, and Yongfeng Zhang. VIP5: towards multimodal foundation models for recommendation. In *Findings of the Association for Computational Linguistics: EMNLP*, volume EMNLP 2023 of *Findings of ACL*, pp. 9606–9620, 2023.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2016.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The International Conference on Learning Representations, ICLR*, 2022.

Wenyue Hua, Yingqiang Ge, Shuyuan Xu, Jianchao Ji, Zelong Li, and Yongfeng Zhang. UP5: unbiased foundation model for fairness-aware recommendation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, EACL*, pp. 1899–1912, 2024.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The International Conference on Learning Representations, ICLR*, 2024.

Meng Jiang, Keqin Bao, Jizhi Zhang, Wenjie Wang, Zhengyi Yang, Fuli Feng, and Xiangnan He. Item-side fairness of large language model-based recommendation system. In *Proceedings of the ACM on Web Conference, WWW*, pp. 4717–4726, 2024.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *The International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217, 2020.

Haruka Kiyohara, Yuta Saito, Tatsuya Matsuhiro, Yusuke Narita, Nobuyuki Shimizu, and Yasuo Yamamoto. Doubly robust off-policy evaluation for ranking policies under the cascade behavior model. In *The ACM International Conference on Web Search and Data Mining, WSDM*, pp. 487–497, 2022.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the International Conference on Machine Learning, ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, 2017.

Zijie Lei, Tao Feng, Zhigang Hua, Yan Xie, Guanyu Lin, Shuang Yang, Ge Liu, and Jiaxuan You. Unirec: Unified multimodal encoding for llm-based recommendations. *Transactions on Machine Learning Research*, 2026, 2026.

Hanzhe Li, Dazhong Shen, Chao Wang, Yuting Liu, and Jingjing Gu. Can llms enhance fairness in recommendation systems? A data augmentation approach. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, pp. 570–580, 2025.

Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, Peng Wu, and Peng Cui. Propensity matters: Measuring and enhancing balancing for recommendation. In *International Conference on Machine Learning, ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20182–20194, 2023a.

Jin Li, Shoujin Wang, Qi Zhang, Longbing Cao, Fang Chen, Xiuzhen Zhang, Dietmar Jannach, and Charu C. Aggarwal. Causal learning for trustworthy recommender systems: A survey. *CoRR*, abs/2402.08241, 2024a.

Jin Li, Huilin Gu, Shoujin Wang, Qi Zhang, Shui Yu, Chen Wang, Xiwei Xu, and Fang Chen. Towards fair large language model-based recommender systems without costly retraining. In *Proceedings of the ACM Web Conference, WWW*, pp. 9675–9686, 2026a.

Jindong Li, Yali Fu, Jiahong Liu, Linxiao Cao, Wei Ji, Menglin Yang, Irwin King, and Ming-Hsuan Yang. Discrete tokenization for multimodal llms: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):9093–9113, 2026b.

Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. Large language models for generative recommendation: A survey and visionary discussions. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING*, pp. 10146–10159, 2024b.

Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5):95:1–95:48, 2023b.

Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. Llara: Large language-recommendation assistant. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, pp. 1785–1795, 2024.

Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. How can recommender systems benefit from large language models: A survey. *ACM Transactions on Information Systems*, 43(2):28:1–28:47, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 26286–26296, 2024a.

Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. Multimodal pretraining, adaptation, and generation for recommendation: A survey. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD*, pp. 6566–6576, 2024b.

Xiaohao Liu, Zhulin Tao, Jiahong Shao, Lifang Yang, and Xianglin Huang. Elimrec: Eliminating single-modal bias in multimedia recommendation. In *The ACM International Conference on Multimedia*, pp. 687–695, 2022.

Yuqing Liu, Yu Wang, Lichao Sun, and Philip S. Yu. Rec-gpt4v: Multimodal recommendation with large vision-language models. *CoRR*, abs/2402.08670, 2024c.

Renqiang Luo, Dong Zhang, Yupeng Gao, Wen Shi, Mingliang Hou, Jiaying Liu, Zhe Wang, and Shuo Yu. Bridging semantic understanding and popularity bias with llms. In *Proceedings of the ACM Web Conference, WWW*, pp. 3656–3665, 2026.

Bo Ma, Hang Li, Zehua Hu, XiaoFan Gui, LuYao Liu, and Simon Lau. Llm4rec: Large language models for multimodal generative recommendation with causal debiasing. *CoRR*, abs/2510.01622, 2025.

Masoud Mansoury, Jin Huang, Mykola Pechenizkiy, Herke van Hoof, and Maarten de Rijke. The unfairness of multifactorial bias in recommendation. *CoRR*, abs/2601.12828, 2026.

Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52, 2015.

Graziano Mita, Maurizio Filippone, and Pietro Michiardi. An identifiable double VAE for disentangled representations. In *Proceedings of the International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7769–7779, 2021.

Yongxin Ni, Yu Cheng, Xiangyan Liu, Junchen Fu, Youhua Li, Xiangnan He, Yongfeng Zhang, and Fajie Yuan. A content-driven micro-video recommendation dataset at scale. In *Proceedings of the ACM International Conference on Information and Knowledge Management, CIKM*, pp. 6486–6491, 2025.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

Yu Shang, Chen Gao, Jiansheng Chen, Depeng Jin, and Yong Li. Improving item-side fairness of multimodal recommendation via modality debiasing. In *Proceedings of the ACM on Web Conference, WWW*, pp. 4697–4705, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024.

Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3):52:1–52:43, 2023.

Yu Wang, Yonghui Yang, Le Wu, Yi Zhang, Fei Liu, and Richang Hong. Multimodal large language models with adaptive preference optimization for sequential recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, pp. 1903–1913, 2026.

Yuxiang Wang, Xin Shi, and Xueqing Zhao. Mllm4rec : multimodal information enhancing LLM for sequential recommendation. *Journal of Intelligent Information Systems*, 63(3):745–761, 2025a.

Zhaoliang Wang, Baisong Liu, Weiming Huang, Tingting Hao, Huiqian Zhou, and Yuxin Guo. Leveraging multimodal large language model for multimodal sequential recommendation. *Scientific Reports*, 15(1):28960, 2025b.

Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *The ACM SIGKDD Conference on Knowledge Discovery and Data Mining, SIGKDD*, pp. 1791–1800, 2021.

Yongxian Wei, Runxi Cheng, Weike Jin, Enneng Yang, Li Shen, Lu Hou, Sinan Du, Chun Yuan, Xiaochun Cao, and Dacheng Tao. Optmerge: Unifying multimodal llm capabilities and modalities via model merging. In *The International Conference on Learning Representations, ICLR*, pp. 61571–61595, 2026.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256, 2023.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. A survey on large language models for recommendation. *World Wide Web (WWW)*, 27(5):60, 2024.

Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal LLM. In *The International Conference on Learning Representations, ICLR*, 2025.

Chen Xu, Jun Xu, Xu Chen, Zhenhua Dong, and Ji-Rong Wen. Dually enhanced propensity score estimation in sequential recommendation. In *Proceedings of the ACM International Conference on Information & Knowledge Management, CIKM*, pp. 2260–2269, 2022.

Yuyang Ye, Zhi Zheng, Yishan Shen, Tianshu Wang, Hengruo Zhang, Peijun Zhu, Runlong Yu, Kai Zhang, and Hui Xiong. Harnessing multimodal large language models for multimodal sequential recommendation. In *AAAI Conference on Artificial Intelligence, AAAI*, pp. 13069–13077, 2025.

Yuxin Ying, Fuzhen Zhuang, Yongchun Zhu, Deqing Wang, and Hongwei Zheng. CAMUS: attribute-aware counterfactual augmentation for minority users in recommendation. In *Proceedings of the ACM Web Conference, WWW*, pp. 1396–1404, 2023.

Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Xiangyu Zhao, Yan Gao, Yao Hu, and Enhong Chen. Notellm-2: Multimodal large representation models for recommendation. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD*, pp. 2815–2826, 2025.

Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal intervention for leveraging popularity bias in recommendation. In *The International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, pp. 11–20, 2021.

Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu C. Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: A survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1):2:1–2:28, 2025a.

Zihuai Zhao, Wenqi Fan, Yao Wu, and Qing Li. Investigating and mitigating stereotype-aware unfairness in llm-based recommendations. *CoRR*, abs/2504.04199, 2025b.

Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. Disentangling user interest and conformity for recommendation with causal embedding. In *The Web Conference, WWW*, pp. 2980–2991, 2021.

Peilin Zhou, Chao Liu, Jing Ren, Xinfeng Zhou, Yueqi Xie, Meng Cao, Zhongtao Rao, You-Liang Huang, Dading Chong, Junling Liu, Jae Boum Kim, Shoujin Wang, Raymond Chi-Wing Wong, and Sunghun Kim. When large vision language models meet multimodal sequential recommendation: An empirical study. In *Proceedings of the ACM on Web Conference, WWW*, pp. 275–292, 2025.

Appendix of TRACE

CONTENTS

A Table of Notations **16**

B Limitations and Future Opportunities **17**

C Algorithm **18**

D Theoretical Analysis **18**

D.1 Proof of Proposition 1 18

D.2 Proof of Proposition 2 19

D.3 Complexity Analysis 22

E More Details about Experiments **22**

E.1 Experimental Details 22

E.2 Full Results of Attribute Bias Mitigation 25

E.3 Multi-bias Mitigation 25

E.4 Modality-Dominance Bias Mitigation 26

E.5 Generalisation on Different Backbones 26

E.6 Additional Ablation Study 27

E.7 Additional Parameter Analysis 28

E.8 Inspection of Identified Bias Concepts 30

E.9 Case Study 33

F Broader Impacts **34**

A TABLE OF NOTATIONS

Notation	Description
$\mathcal{U}, \mathcal{V}$	The user set and the item set, with user $u \in \mathcal{U}$ and item $v = (\mathbf{x}^{\text{img}}, \mathbf{x}^{\text{txt}}) \in \mathcal{V}$.
$\mathbf{S}_u$	The interaction history of u, $\mathbf{S}_u = [\mathbf{s}_u^1, \cdots, \mathbf{s}_u^t]$, and $\mathbf{s}_u^j \in \mathcal{V}$.
$\mathbf{S}_u^{1:j}$	The length-j interaction prefix $\mathbf{S}_u^{1:j} = [\mathbf{s}_u^1, \cdots, \mathbf{s}_u^j]$.
z	The training sample with the interaction history $\mathbf{S}_u$, candidate item v_c, and ground truth y_u^c.
$\mathcal{D}$	The training set for supervised fine-tuning, $\mathcal{D} = \{z_i\}_{i=1}^n$, with n training samples.
g_θ	The backbone MLRS model with parameter θ.
f_θ	The decision score using the Yes/No margin.
$\hat{\theta}$	The fine-tuned LoRA adapter.
l	The per-sample loss for fine-tuning.
R	Empirical risk.
$\mathcal{B}, B$	The bias functional; the bias outcome variable.
$\mathcal{M}$	Modalities involved in our framework.
$\mathbf{h}[p]$	The last-layer hidden state at prompt position p.
$\mathbf{e}$	Extracted representation for each modality.
μ_ψ	The iSAE encoder mean.
D_ξ	The decoder in iSAE.
$p_\lambda(\cdot \mid a)$	The conditional prior with auxiliary a.
$\hat{\boldsymbol{c}}$	Concept code.
K_m	Number of concepts per modality.
$\mathcal{G}$	The derived causal graph.
τ_k^m	Each concept's contribution to bias.
$\mathbf{H}$	The curvature estimation.
I_c, I_{chain}	The sample-to-concept influence; the chained influence.
$\mathcal{D}_{\text{bias}}$	Top-K_b bias-inducing samples.
$\mathbf{G}_{\text{bias}}, \mathbf{P}_{\text{bias}}$	Concept-direction matrix, projector.
$\Delta\theta_{\text{R}}$	Parameter update via the bias removal operator.
$\Delta\theta_{\text{C}}$	Parameter update via the corrective operator.
$\Delta\theta_{\text{RC}}$	Parameter update via the composed operators.

B Limitations and Future Opportunities

We identify several limitations that leave opportunities for future research.

Identifiability with limited auxiliary information. Our identifiable concept learning relies on a bias-relevant auxiliary variable to condition the latent prior. As discussed in Remark 1, formal identifiability requires sufficiently rich variation in the auxiliary-conditioned latent distributions, which can be difficult to satisfy in practical recommendation settings where only limited auxiliary information is available. We therefore do not claim formal identifiability of the learned concepts under these practical settings. An important future direction is to investigate how richer bias auxiliary information can be obtained or constructed, for example, through auxiliary augmentation, continuous or multi-view auxiliary signals, or complementary side information, to better satisfy the conditions required for identifiable concept learning while remaining applicable to real-world recommender systems.

Local assumptions behind influence estimation and mitigation. The sample-to-concept influence in our framework follows standard local influence analysis and assumes that the fine-tuned model is at a stationary solution, together with local smoothness and well-behaved curvature. These conditions may hold only approximately in practice (Bae et al., 2022; Chen et al., 2023b). Moreover, while our damped diagonal Fisher approximation avoids explicitly constructing and inverting the full Hessian and substantially improves computational feasibility, it remains an approximation to the local curvature used by the influence formulation. Thus, both the estimated sample influence and the resulting one-step removal update may become less accurate when the model is away from stationarity or when the local curvature is poorly approximated. Future work could explore influence formulations that relax the stationarity and positive-curvature assumptions, as well as more robust and scalable curvature approximations specialised for large multimodal foundation models.

Modalities, datasets, and foundation models. Our experiments include three multimodal recommendation datasets and two representative MLLM backbones to cover different recommendation domains, model families, and sizes. Moreover, the current multimodal setting mainly focuses on textual and visual item information together with user preference signals. Real-world recommender systems can involve additional modalities, such as audio, video, speech, and other contextual or behavioural signals, which may introduce new forms of intra- and cross-modal bias propagation. Evaluating such settings is currently constrained by the limited availability of recommendation benchmarks containing sufficiently rich multimodal content together with the information required for systematic bias evaluation. Future work can extend existing multimodal recommendation datasets and foundation models supporting a broader range of modalities, and investigate whether new cross-modal bias pathways emerge as the number and diversity of modalities increase.

Scope of bias and fairness evaluation. Fairness is inherently application- and stakeholder-dependent. Our experiments primarily instantiate TRACE for popularity bias and user-attribute bias, with additional analyses of modality and co-existing biases. Although these settings demonstrate the flexibility of the proposed framework, they do not exhaust the biases that may arise in real-world recommendations. Future work could further investigate the generalisability of this framework to diverse and newly emerged biases.

C ALGORITHM

Algorithm A.1 The TRACE framework

Input: the backbone MLRS model θ; the training set $\mathcal{D}$; auxiliary variable a.

Output: the debiased MLRS model θ'.

Identification:

1. Extract per-modality features $\phi_\theta^m(z)$, $m \in \{\mathrm{txt}, \mathrm{img}, \mathrm{inv}, \mathrm{tmp}\}$;
2. Train iSAEs using Eq. 3 conditioned on the auxiliary;
3. Extract concepts $\hat{\mathbf{c}}^m(z) = \mu_\psi(\phi_\theta^m(z))$;
4. Caudal discovery to obtain a DAG $\mathcal{G}$ and measure total derivatives τ.
5. Measure chain influence $I_{\mathrm{chain}}(z) = \sum_k \tau_k I_c(z \to c_k)$ over $\mathcal{D}$.

Mitigation:

6. Compute causal subspace projector $\mathbf{P}_{\mathrm{bias}}$ using Eq. 9;
7. $\Delta\theta_{\mathrm{R}} = -\eta_{\mathrm{R}}\, \mathbf{P}_{\mathrm{bias}} \mathbf{H}^{-1} \sum_{z \in \mathcal{D}_{\mathrm{bias}}} \nabla_\theta \ell(z)$ with $\mathcal{D}_{\mathrm{bias}} = \mathrm{Top}\, K_b |I_{\mathrm{chain}}|$.
8. Select counter-shorcut pairs $\mathcal{P}_{\mathrm{cs}} = \{(z^+, z^-)\}$ via Eq. 11.
9. Compute corrective direction $\boldsymbol{u}_{\mathrm{C}} = \mathbf{P}_{\mathrm{bias}} \nabla_\theta \sum \max(0, m^* - (f(z^+) - f(z^-)))$.
10. Compute composed parameter update $\Delta\theta_{\mathrm{RC}} = \alpha\, \Delta\theta_{\mathrm{R}} - D\, \boldsymbol{u}_{\mathrm{C}} / \|\boldsymbol{u}_{\mathrm{C}}\|$.
11. Apply parameter update via $\theta' \leftarrow \hat{\theta} + \Delta\theta_{\mathrm{RC}}$ for bias mitigation.

D THEORETICAL ANALYSIS

D.1 PROOF OF PROPOSITION 1

Proof of Proposition 1. For brevity, we index each modality-concept pair (m, k) by j and write $\bar{c}_j(\theta)$ and τ_j for its probe-averaged activation and centred causal-effect weight, respectively. By definition,

$$\mathcal{J}_{\mathrm{cen}}(\theta) = \sum_j \tau_j \bar{c}_j(\theta). \tag{D.15}$$

(1) Sample-to-concept influence. Let $\hat{\theta}_\epsilon = \arg\min_\theta \big\{R(\theta) + \epsilon \ell(z;\theta)\big\}$. The first-order optimality condition is

$$\nabla_\theta R(\hat{\theta}_\epsilon) + \epsilon \nabla_\theta \ell(z; \hat{\theta}_\epsilon) = \mathbf{0}. \tag{D.16}$$

Differentiating Eq. D.16 with respect to ϵ and evaluating at $\epsilon = 0$ gives

$$\mathbf{H} \frac{d\hat{\theta}_\epsilon}{d\epsilon}\Big|_{\epsilon=0} + \nabla_\theta \ell(z; \hat{\theta}) = \mathbf{0},$$

where $\mathbf{H} = \nabla_\theta^2 R(\hat{\theta})$. Since $\mathbf{H}$ is positive definite under Assumption 2, we have

$$\frac{d\hat{\theta}_\epsilon}{d\epsilon}\Big|_{\epsilon=0} = -\mathbf{H}^{-1} \nabla_\theta \ell(z; \hat{\theta}). \tag{D.17}$$

Applying the chain rule to each averaged concept activation gives

$$\begin{aligned} \frac{d\,\bar{c}_j(\hat{\theta}_\epsilon)}{d\epsilon}\Big|_{\epsilon=0} &= \nabla_\theta \bar{c}_j(\hat{\theta})^\top \frac{d\hat{\theta}_\epsilon}{d\epsilon}\Big|_{\epsilon=0} && \text{(D.18)} \\ &= -\nabla_\theta \bar{c}_j(\hat{\theta})^\top \mathbf{H}^{-1} \nabla_\theta \ell(z; \hat{\theta}) \\ &= I_c(z \to c_j). && \text{(D.19)} \end{aligned}$$

(2) Chain influence. Since the SCM and hence the centred causal-effect weights τ_j are fixed when differentiating with respect to the infinitesimal training sample weight ϵ, we have

$$\begin{aligned} \frac{d}{d\epsilon} \mathcal{J}_{\mathrm{cen}}(\hat{\theta}_\epsilon)\Big|_{\epsilon=0} &= \sum_j \tau_j \frac{d\,\bar{c}_j(\hat{\theta}_\epsilon)}{d\epsilon}\Big|_{\epsilon=0} && \text{(D.20)} \\ &= \sum_j \tau_j I_c(z \to c_j). && \text{(D.21)} \end{aligned}$$

Restoring the modality and concept indices yields

$$I_{\text{chain}}(z) = \sum_{m \in \mathcal{M}} \sum_{k=1}^{K_m} \tau_k^m I_c(z \to c_k^m),$$

which proves Eq. 8.

(3) Infinitesimal deletion. For the empirical-risk convention $R(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(z_i;\theta)$, deleting one sample is represented to first order by $\epsilon = -1/n$. A first-order Taylor expansion around $\epsilon = 0$ therefore gives

$$\mathcal{J}_{\text{cen}}(\hat{\theta}_{-1/n}) - \mathcal{J}_{\text{cen}}(\hat{\theta}) = -\frac{1}{n} I_{\text{chain}}(z) + o(n^{-1}). \tag{D.22}$$

(4) Invariance to equivalent concept coordinates. Consider a non-zero component-wise rescaling and permutation of the centred concept coordinates,

$$\hat{\boldsymbol{c}} = \boldsymbol{\Lambda}\boldsymbol{\Pi}\boldsymbol{c},$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_K)$ with $\lambda_k \neq 0$ and $\boldsymbol{\Pi}$ is a permutation matrix. For coordinate k,

$$\hat{c}_k = \lambda_k c_{\pi(k)}, \qquad \bar{\hat{c}}_k(\theta) = \lambda_k \bar{c}_{\pi(k)}(\theta).$$

Consequently,

$$I_c(z \to \hat{c}_k) = \lambda_k I_c(z \to c_{\pi(k)}). \tag{D.23}$$

Let

$$F_j(x) = \mathbb{E}[B \mid \text{do}(c_j = x)].$$

Under the reparameterisation $\hat{c}_k = \lambda_k c_{\pi(k)}$, intervening on $\hat{c}_k = \hat{x}$ is equivalent to intervening on $c_{\pi(k)} = \hat{x}/\lambda_k$. Hence, the corresponding interventional response satisfies

$$\hat{F}_k(\hat{x}) = F_{\pi(k)}\left(\frac{\hat{x}}{\lambda_k}\right).$$

By the chain rule,

$$\hat{\tau}_k = \hat{F}_k'(0) = \lambda_k^{-1} F_{\pi(k)}'(0) = \lambda_k^{-1} \tau_{\pi(k)}. \tag{D.24}$$

Combining Eqs. D.23 and D.24, the scaling cancels:

$$\hat{\tau}_k I_c(z \to \hat{c}_k) = \tau_{\pi(k)} I_c(z \to c_{\pi(k)}).$$

The only remaining change is the permutation, and therefore $I_{\text{chain}}(z)$ is invariant.

Finally,

$$\nabla_\theta \bar{\hat{c}}_k = \lambda_k \nabla_\theta \bar{c}_{\pi(k)},$$

so similarly

$$\hat{\tau}_k \nabla_\theta \bar{\hat{c}}_k = \tau_{\pi(k)} \nabla_\theta \bar{c}_{\pi(k)}.$$

Hence, the columns of $\mathbf{G}_{\text{bias}}$ are unchanged up to permutation, proving the stated invariance of the causally weighted concept-gradient representation. □

D.2 Proof of Proposition 2

Proof of Proposition 2. Let

$$\mathcal{S}_{\text{bias}} = \text{span}(\mathbf{U}_r), \qquad \mathbf{P}_{\text{bias}} = \mathbf{U}_r \mathbf{U}_r^\top.$$

Since the columns of $\mathbf{U}_r$ are orthonormal, $\mathbf{P}_{\text{bias}}$ is the orthogonal projector onto $\mathcal{S}_{\text{bias}}$ and therefore satisfies

$$\mathbf{P}_{\text{bias}}^\top = \mathbf{P}_{\text{bias}}, \qquad \mathbf{P}_{\text{bias}}^2 = \mathbf{P}_{\text{bias}}.$$

(1) Properties of the removal operator. By definition,

$$\Delta\theta_{\mathrm{R}} = \mathbf{P}_{\mathrm{bias}}\Delta_{\mathrm{full}}.$$

For any $\delta \in \mathcal{S}_{\mathrm{bias}}$, we have $\mathbf{P}_{\mathrm{bias}}\delta = \delta$. Decompose

$$\Delta_{\mathrm{full}} = \mathbf{P}_{\mathrm{bias}}\Delta_{\mathrm{full}} + (\mathbf{I} - \mathbf{P}_{\mathrm{bias}})\Delta_{\mathrm{full}} = \Delta\theta_{\mathrm{R}} + (\mathbf{I} - \mathbf{P}_{\mathrm{bias}})\Delta_{\mathrm{full}},$$

where the two terms are orthogonal. Hence

$$\|\delta - \Delta_{\mathrm{full}}\|_2^2 = \|\delta - \Delta\theta_{\mathrm{R}} - (\mathbf{I} - \mathbf{P}_{\mathrm{bias}})\Delta_{\mathrm{full}}\|_2^2 \tag{D.25}$$

$$= \|\delta - \Delta\theta_{\mathrm{R}}\|_2^2 + \|(\mathbf{I} - \mathbf{P}_{\mathrm{bias}})\Delta_{\mathrm{full}}\|_2^2. \tag{D.26}$$

The second term is independent of δ, so the minimum is attained when $\delta = \Delta\theta_{\mathrm{R}}$, i.e.,

$$\Delta\theta_{\mathrm{R}} = \arg\min_{\delta \in \mathcal{S}_{\mathrm{bias}}} \|\delta - \Delta_{\mathrm{full}}\|_2.$$

The same orthogonal decomposition gives

$$\|\Delta_{\mathrm{full}}\|_2^2 = \|\Delta\theta_{\mathrm{R}}\|_2^2 + \|(\mathbf{I} - \mathbf{P}_{\mathrm{bias}})\Delta_{\mathrm{full}}\|_2^2,$$

and thus

$$\|\Delta\theta_{\mathrm{R}}\|_2 \le \|\Delta_{\mathrm{full}}\|_2.$$

Now let F satisfy $\nabla F(\hat{\theta}) \in \mathcal{S}_{\mathrm{bias}}$. Since an orthogonal projector acts as the identity on its range,

$$\mathbf{P}_{\mathrm{bias}}\nabla F(\hat{\theta}) = \nabla F(\hat{\theta}).$$

Using $\mathbf{P}_{\mathrm{bias}}^\top = \mathbf{P}_{\mathrm{bias}}$,

$$\nabla F(\hat{\theta})^\top \Delta\theta_{\mathrm{R}} = \nabla F(\hat{\theta})^\top \mathbf{P}_{\mathrm{bias}}\Delta_{\mathrm{full}} \tag{D.27}$$

$$= \left(\mathbf{P}_{\mathrm{bias}}\nabla F(\hat{\theta})\right)^\top \Delta_{\mathrm{full}} \tag{D.28}$$

$$= \nabla F(\hat{\theta})^\top \Delta_{\mathrm{full}}. \tag{D.29}$$

Thus, projection preserves the first-order effect of the full update on any functional whose gradient lies in the retained bias-mechanism subspace.

(2) Properties of the corrective operator. Let

$$\mathbf{u} = \mathbf{P}_{\mathrm{bias}}\mathbf{g}_{\mathrm{pref}} \neq \mathbf{0}, \qquad \Delta\theta_{\mathrm{C}} = -D\frac{\mathbf{u}}{\|\mathbf{u}\|_2}.$$

Its norm is immediately

$$\|\Delta\theta_{\mathrm{C}}\|_2 = D.$$

Moreover,

$$\mathbf{g}_{\mathrm{pref}}^\top \mathbf{u} = \mathbf{g}_{\mathrm{pref}}^\top \mathbf{P}_{\mathrm{bias}}\mathbf{g}_{\mathrm{pref}} \tag{D.30}$$

$$= \mathbf{g}_{\mathrm{pref}}^\top \mathbf{P}_{\mathrm{bias}}^\top \mathbf{P}_{\mathrm{bias}}\mathbf{g}_{\mathrm{pref}} \tag{D.31}$$

$$= \|\mathbf{P}_{\mathrm{bias}}\mathbf{g}_{\mathrm{pref}}\|_2^2 = \|\mathbf{u}\|_2^2, \tag{D.32}$$

where we used $\mathbf{P}_{\mathrm{bias}}^\top = \mathbf{P}_{\mathrm{bias}}$ and $\mathbf{P}_{\mathrm{bias}}^2 = \mathbf{P}_{\mathrm{bias}}$. Therefore

$$\mathbf{g}_{\mathrm{pref}}^\top \Delta\theta_{\mathrm{C}} = -D\,\frac{\mathbf{g}_{\mathrm{pref}}^\top \mathbf{u}}{\|\mathbf{u}\|_2} \tag{D.33}$$

$$= -D\|\mathbf{u}\|_2 \tag{D.34}$$

$$= -D\|\mathbf{P}_{\mathrm{bias}}\mathbf{g}_{\mathrm{pref}}\|_2 < 0. \tag{D.35}$$

Hence $\Delta\theta_{\mathrm{C}}$ is a strict first-order descent direction for the selected pair objective whenever $\mathbf{P}_{\mathrm{bias}}\mathbf{g}_{\mathrm{pref}} \neq \mathbf{0}$.

(3) Subspace locality and the composed descent condition. Because

$$\Delta\theta_{\mathrm{R}} = \mathbf{P}_{\mathrm{bias}}\Delta_{\mathrm{full}} \in \mathcal{S}_{\mathrm{bias}}$$

and

$$\Delta\theta_{\mathrm{C}} = -D\frac{\mathbf{P}_{\mathrm{bias}}\mathbf{g}_{\mathrm{pref}}}{\|\mathbf{P}_{\mathrm{bias}}\mathbf{g}_{\mathrm{pref}}\|_2} \in \mathcal{S}_{\mathrm{bias}},$$

their composition

$$\Delta\theta_{\mathrm{RC}} = \alpha\Delta\theta_{\mathrm{R}} + \Delta\theta_{\mathrm{C}}$$

also belongs to $\mathcal{S}_{\mathrm{bias}}$.

Let F have an L_F-Lipschitz gradient and satisfy

$$\nabla F(\hat{\theta}) \perp \mathcal{S}_{\mathrm{bias}}.$$

Then, for any

$$\Delta \in \{\Delta\theta_{\mathrm{R}}, \Delta\theta_{\mathrm{C}}, \Delta\theta_{\mathrm{RC}}\},$$

we have

$$\nabla F(\hat{\theta})^\top \Delta = 0.$$

The standard smoothness inequality gives

$$\left|F(\hat{\theta}+\Delta) - F(\hat{\theta}) - \nabla F(\hat{\theta})^\top\Delta\right| \le \frac{L_F}{2}\|\Delta\|_2^2.$$

Since the linear term vanishes,

$$\left|F(\hat{\theta}+\Delta) - F(\hat{\theta})\right| \le \frac{L_F}{2}\|\Delta\|_2^2.$$

Thus, all three operators have no first-order effect on functionals whose gradients are orthogonal to $\mathcal{S}_{\mathrm{bias}}$.

Next define

$$c = \mathbf{g}_{\mathrm{pref}}^\top \Delta\theta_{\mathrm{R}}.$$

Since $\Delta\theta_{\mathrm{R}} \in \mathcal{S}_{\mathrm{bias}}$, we have $\mathbf{P}_{\mathrm{bias}}\Delta\theta_{\mathrm{R}} = \Delta\theta_{\mathrm{R}}$, and hence

$$\begin{aligned} c &= \mathbf{g}_{\mathrm{pref}}^\top \Delta\theta_{\mathrm{R}} && \text{(D.36)}\\ &= \mathbf{g}_{\mathrm{pref}}^\top \mathbf{P}_{\mathrm{bias}}\Delta\theta_{\mathrm{R}} && \text{(D.37)}\\ &= (\mathbf{P}_{\mathrm{bias}}\mathbf{g}_{\mathrm{pref}})^\top \Delta\theta_{\mathrm{R}} = \mathbf{u}^\top\Delta\theta_{\mathrm{R}}. && \text{(D.38)} \end{aligned}$$

Therefore,

$$\begin{aligned} \mathbf{g}_{\mathrm{pref}}^\top \Delta\theta_{\mathrm{RC}} &= \alpha\mathbf{g}_{\mathrm{pref}}^\top \Delta\theta_{\mathrm{R}} + \mathbf{g}_{\mathrm{pref}}^\top \Delta\theta_{\mathrm{C}} && \text{(D.39)}\\ &= \alpha c - D\|\mathbf{u}\|_2. && \text{(D.40)} \end{aligned}$$

The composed update remains a strict first-order descent direction for the pair objective whenever

$$\alpha c < D\|\mathbf{u}\|_2. \tag{D.41}$$

If $c \le 0$, this condition holds automatically for every $\alpha \ge 0$.

When $c > 0$ and $\Delta\theta_{\mathrm{R}} \neq \mathbf{0}$, let ω denote the angle between $\mathbf{u}$ and $\Delta\theta_{\mathrm{R}}$. Then

$$c = \mathbf{u}^\top\Delta\theta_{\mathrm{R}} = \|\mathbf{u}\|_2\|\Delta\theta_{\mathrm{R}}\|_2\cos\omega.$$

Since $\mathbf{u} \neq \mathbf{0}$, dividing the descent condition by $\|\mathbf{u}\|_2$ gives

$$\alpha\|\Delta\theta_{\mathrm{R}}\|_2\cos\omega < D.$$

If $\Delta\theta_{\mathrm{R}} = \mathbf{0}$, the condition is trivially satisfied.

(4) First-order additivity of the composition. Let F have an L_F-Lipschitz gradient. Applying the same smoothness inequality to $\Delta\theta_{\mathrm{RC}}$ gives

$$\left|F(\hat{\theta}+\Delta\theta_{\mathrm{RC}})-F(\hat{\theta})-\nabla F(\hat{\theta})^{\top}\Delta\theta_{\mathrm{RC}}\right| \le \frac{L_F}{2}\|\Delta\theta_{\mathrm{RC}}\|_2^2.$$

Using

$$\Delta\theta_{\mathrm{RC}} = \alpha\Delta\theta_{\mathrm{R}} + \Delta\theta_{\mathrm{C}},$$

the linear term decomposes as

$$\nabla F(\hat{\theta})^{\top}\Delta\theta_{\mathrm{RC}} = \alpha\nabla F(\hat{\theta})^{\top}\Delta\theta_{\mathrm{R}} + \nabla F(\hat{\theta})^{\top}\Delta\theta_{\mathrm{C}}.$$

Therefore

$$\left|F(\hat{\theta}+\Delta\theta_{\mathrm{RC}})-F(\hat{\theta})-\alpha\nabla F(\hat{\theta})^{\top}\Delta\theta_{\mathrm{R}}-\nabla F(\hat{\theta})^{\top}\Delta\theta_{\mathrm{C}}\right| \le \frac{L_F}{2}\|\Delta\theta_{\mathrm{RC}}\|_2^2,$$

which proves the stated first-order additivity of the removal and corrective effects. □

D.3 Complexity Analysis

TRACE operates on a uniformly sampled probe set $\mathcal{D}_{\mathrm{probe}} \subset \mathcal{D}$. Let $n_p = |\mathcal{D}_{\mathrm{probe}}|$, p be the number of trainable LoRA parameters, d be the representation dimension, $K = \sum_m K_m$ the total number of learned concepts, and K' the number retained for causal analysis, where typically $n_p \ll n$ and $K' \ll p$. We denote C_{grad} the cost of one LoRA-restricted gradient evaluation.

Bias identification. Given the extracted d-dimensional probe representations, training the modality-specific concept learners for E_{SAE} epochs costs $O(E_{\mathrm{SAE}} n_p d K)$. The subsequent causal discovery requires $C_{\mathrm{causal}}(n_p, K')$, which depends on the adopted discovery algorithm.

Influence estimation. We avoid forming or inverting the full $p \times p$ Hessian. Instead, we use a diagonal Fisher approximation estimated from the probe samples, together with damping $\lambda > 0$. This requires one gradient sweep over $\mathcal{D}_{\mathrm{probe}}$, $O(n_p C_{\mathrm{grad}})$, while applying $(\mathbf{F} + \lambda\mathbf{I})^{-1}$ requires only $O(p)$ time and $O(p)$ memory.

Bias-mechanism subspace and mitigation. Constructing $\mathbf{G}_{\mathrm{bias}}$ requires at most K' probe-averaged concept-gradient evaluations, giving $O(K' n_p C_{\mathrm{grad}})$ backbone computation. Since $K' \ll p$, the thin SVD of $\mathbf{G}_{\mathrm{bias}}$ costs $O(pK'^2)$. For any vector $\boldsymbol{x}$, $\mathbf{P}_{\mathrm{bias}}\boldsymbol{x} = \mathbf{U}_r(\mathbf{U}_r^{\top}\boldsymbol{x})$, which requires only $O(pr)$ operations.

Once the top-K_b bias-driving samples or corrective pairs are identified, constructing the removal and corrective gradients requires at most $O(K_b C_{\mathrm{grad}})$ additional computation, followed by $O(pr)$ projection and one-step parameter updates.

Collecting the dominant terms, the overall complexity is

$$O\left(E_{\mathrm{SAE}} n_p d K + K' n_p C_{\mathrm{grad}} + pK'^2 + C_{\mathrm{causal}}(n_p, K')\right). \tag{D.42}$$

The lower-order Fisher-estimation, chain-scoring, mitigation, and projection costs are absorbed into the terms above.

The working memory is $O(n_p d + pK' + p)$ corresponding primarily to the cached probe representations, the bias-mechanism matrix, and LoRA-scale gradient/Fisher vectors. Since $n_p \ll n$ and all identification and mitigation stages operate on the fixed probe subset, the additional backbone computation scales with n_p rather than the full training-set size n.

E More Details about Experiments

E.1 Experimental Details

E.1.1 Datasets

We adopt three widely used multimodal recommendation datasets with statistics listed in Table 5.

Table 5: Dataset statistics.

Dataset	MicroLens	Amazon Baby	MovieLens-1M
#User	25,411	41,081	6,040
#Item	20,276	14,393	3,952
#Interaction	223,263	400,876	1,000,209
Density	2.1×10^{-4}	6.8×10^{-4}	4.2×10^{-2}

- **MicroLens** (Ni et al., 2025) is a content-driven micro-video recommendation dataset. We follow Ye et al. (2025); Wang et al. (2026) in using titles and images as the textual and visual modalities.
- **Amazon Baby** (McAuley et al., 2015) is a product recommendation dataset collected from the Baby category of Amazon, containing user reviews and product metadata. We use product titles and images as the textual and visual modalities.
- **MovieLens-1M** (Harper & Konstan, 2016) is a widely used movie recommendation dataset containing approximately one million user ratings. We use movie titles and posters as the textual and visual modalities.

We split each dataset into training, validation, and test sets in an 8:1:1 ratio. Following Wang et al. (2026), we keep the most recent 6 interactions in each sequence, where the last one is used as the prediction target. Additionally, we follow Ye et al. (2025) to implement a 1:1 ratio for negative sampling during training and a 1:20 ratio for evaluation.

E.1.2 BASELINES

We collect a wide range of representative and state-of-the-art debiasing baselines for a comprehensive comparison. Note that, to the best of our knowledge, there is no existing debiasing framework designed for MLRSs. Thus, we adopt and adapt debiasing methods from LLM-based RSs and conventional RSs to our settings for a fair comparison.

Popularity debiasing:

1. **RW** (Jiang et al., 2024) is a training-side reweighting method to upweight targets whose popularity group is under-represented.
2. **RR** (Jiang et al., 2024) is an inference-side re-ranking method that adjusts scores in proportion to a fitted per-group unfairness profile.
3. **RWRR** (Jiang et al., 2024) is a combined version of both RW and RR.
4. **SPRec** (Gao et al., 2025) is a self-play debiasing method built on preference optimisation to push down over-recommended items.
5. **FairLRM** (Luo et al., 2026) is a prompt-based popularity debiasing method that directly asks the model to compose a list that skews toward tail items for niche users.
6. **FUDLR** (Li et al., 2026a) is a machine unlearning-based debiasing framework for LLM-based RSs. It scores each training sample by its influence on bias and then unlearns the most bias-inducing samples through parameter updates.
7. **PDA** (Zhang et al., 2021) is a classic causal-based popularity debiasing method. It casts item popularity as a confounder of the user-item match and removes it by a do-calculus adjustment.
8. **MACR** (Wei et al., 2021) is also a representative causal-based debiasing method that applies counterfactual inference to estimate and remove the bias effect on the ranking score.

Attribute debiasing:

- **CFP** (Zhang et al., 2021) uses counterfactually-fair prompting for user-attribute fairness. A continuous prefix is prepended to the frozen recommender and trained adversarially against

a discriminator that tries to read the sensitive attribute off the prefix-conditioned representation, so that the recommendation becomes insensitive to that attribute.

- **LLM-FairRec** (Li et al., 2025) is a data augmentation method that asks an LLM to produce fairness-aware candidate ranking for worst-served users and retrains the recommender with augmented data.
- **FUDLR** (Li et al., 2026a), as a general debiasing framework for LLM-based RSs, it also supports attribute debiasing by adopting a demographic-parity fairness objective.

Note that some of the existing debiasing methods, which are designed for factorisation-machine-based RS models, e.g., Zheng et al. (2021) and Chen et al. (2025), cannot be adapted to MLLM-RSs due to their intrinsically different generative recommendation mechanism. Meanwhile, although Ma et al. (2025) discussed a debiasing pipeline for multimodal generative recommendations, it is not included in the comparison due to insufficient design details for implementation.

E.1.3 Evaluation

We evaluate the debiasing performance from three perspectives: accuracy-oriented, fairness-oriented, and combined metrics.

Accuracy. **AUC** measures the probability that the positive item outranks a sampled negative, and **NDCG**@K evaluates the model's Top-K ranking ability. A higher score indicates better recommendation accuracy.

Popularity fairness. Items are grouped into five tiers by their training interaction count. Tier 5 is the *popular* tier, and Tier 1 is the *tail*. We use **TailP**@K to measure the proportion of tail items in the Top-K recommendation list. More importantly, we use **NPCG**@K to compute NDCG@K on test samples whose target item is non-popular. A popularity debiasing method that blindly recommends tail items to users without considering genuine user preferences will lead to a low NPCG score. These two fairness-oriented metrics are the lower the better.

Attribute fairness. Users are split into two groups by their sensitive attribute (e.g., gender, age, or occupation). We use Δ**NDCG**@K to measure the absolute gap across user groups. A lower score indicates that users from different groups receive a more equitable treatment.

Combined scores. We adopt the harmonic mean $H(\cdot,\cdot)$ to measure the balance between accuracy and fairness. Specifically, we compute $F_{\text{pop}}@K = H(\text{NDCG@}K, \text{TailP@}K)$ for popularity debiasing evaluation, and $F_{\text{attr}}@K = H(\text{NDCG@}K, 1 - 10 \cdot \Delta\text{NDCG@}K)$ for attribute debiasing evaluation, accounting for the different directions and scales of metric values.

In our experiments, we report results with the common choice $K \in \{3, 5\}$ (Ye et al., 2025; Wang et al., 2026). The averaged scores and standard deviations are reported from multiple runs.

E.1.4 Implementation Details

For a fair comparison, all methods are run on a server with 3 NVIDIA A100 GPUs. We adopt Qwen-2.5-VL-3B-Instruct[2] (Wang et al., 2024; 2026) as the MLLM backbone and follow Ye et al. (2025) to perform SFT on recommendation tasks with LoRA (rank 8). Each training example is the prompt in Appendix E.1.5. Training uses a cosine schedule at 10^{-4} for five epochs. We also adopt a second MLLM backbone, LLaVA-NeXT-Mistral-7B[3] (Liu et al., 2024a), with results shown in Appendix E.5. The backbone is fine-tuned with LoRA on the cross-entropy of the Yes/No answer token. At inference, the decision score is the long-sum-exp margin over the tokeniser variants of Yes and No (Eq. 1), and candidates are ranked by this margin.

E.1.5 Prompt Template

Each training/evaluation sample includes the user history and the candidate item into a single multimodal prompt:

[2] https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct
[3] https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

Table 6: User attribute debiasing on three sensitive attributes in the MovieLens dataset. The best results are highlighted in bold, and the second best are underlined. The backbone results are listed as a reference to the original performance. * indicates the improvement is significant with $p < 0.05$.

Attribute	Baseline	AUC ↑	NDCG@3 ↑	NDCG@5 ↑	ΔNDCG@3 ↓	ΔNDCG@5 ↓	F_{attr}@3 ↑	F_{attr}@5 ↑	Time ↓
Gender	Backbone	85.34 ± 0.14	54.57 ± 1.11	59.11 ± 0.69	2.36 ± 1.39	1.92 ± 0.67	63.31 ± 4.92	68.21 ± 2.72	9.44
	CFP	85.01 ± 0.21	52.31 ± 1.31	57.39 ± 1.12	3.79 ± 2.40	1.90 ± 1.94	55.01 ± 11.48	66.48 ± 7.30	7.10
	LLM-FairRec	84.84 ± 0.05	52.36 ± 0.05	57.27 ± 0.73	2.39 ± 1.32	1.88 ± 0.46	61.75 ± 4.14	67.13 ± 1.11	12.60
	FUDLR	**85.37 ± 0.14**	**54.03 ± 0.68**	**58.42 ± 0.25**	3.92 ± 1.02	2.62 ± 0.75	56.94 ± 3.89	65.11 ± 2.69	4.76
	TRACE-R	84.08 ± 0.17	50.28 ± 0.73	55.75 ± 1.29	2.00 ± 1.80	1.61 ± 1.41	61.22 ± 6.06	66.70 ± 5.16	4.01
	TRACE-C	85.14 ± 0.13	53.95 ± 1.12	58.21 ± 0.88	1.53 ± 1.00	1.31 ± 1.26	65.72 ± 2.58	69.43 ± 3.77	**4.00**
	TRACE-RC	84.88 ± 0.12	53.47 ± 1.00	58.05 ± 0.77	**0.56 ± 0.15***	**1.18 ± 0.36***	**68.25 ± 0.43***	**69.97 ± 0.56***	4.20
Age	Backbone	85.34 ± 0.14	54.57 ± 1.11	59.11 ± 0.69	2.86 ± 2.00	2.68 ± 1.33	61.04 ± 7.71	65.10 ± 5.47	9.44
	CFP	85.02 ± 0.27	52.57 ± 0.87	57.51 ± 1.15	3.85 ± 1.97	2.80 ± 0.91	55.56 ± 9.22	63.78 ± 3.97	6.82
	LLM-FairRec	84.86 ± 0.06	52.35 ± 0.06	57.48 ± 0.89	2.96 ± 0.51	2.19 ± 0.03	60.02 ± 1.83	66.21 ± 0.71	13.52
	FUDLR	84.92 ± 0.19	53.13 ± 0.02	57.27 ± 0.33	3.01 ± 0.23	2.19 ± 0.92	60.38 ± 0.85	65.98 ± 3.52	**4.80**
	TRACE-R	80.83 ± 0.08	43.86 ± 0.39	49.94 ± 0.19	3.02 ± 0.26	2.39 ± 0.11	53.86 ± 1.08	60.29 ± 0.22	**5.03**
	TRACE-C	**85.10 ± 0.14**	**53.33 ± 0.92**	**57.94 ± 1.02***	2.27 ± 0.89	2.29 ± 1.17	**63.04 ± 3.62**	66.04 ± 4.97	5.06
	TRACE-RC	83.65 ± 0.14	48.19 ± 0.71	53.92 ± 0.45	**2.01 ± 0.57***	**1.25 ± 0.76***	60.09 ± 2.08	**66.66 ± 2.52**	5.34
Occupation	Backbone	85.34 ± 0.14	54.57 ± 1.11	59.11 ± 0.69	2.33 ± 1.29	2.08 ± 0.72	63.53 ± 4.72	67.62 ± 2.94	9.44
	CFP	85.08 ± 0.17	52.31 ± 1.31	57.47 ± 1.22	3.36 ± 1.18	2.35 ± 1.55	58.21 ± 4.62	65.15 ± 5.91	6.85
	LLM-FairRec	84.92 ± 0.11	52.57 ± 0.39	57.68 ± 0.72	3.29 ± 1.78	2.31 ± 2.20	58.21 ± 7.55	64.88 ± 8.90	16.37
	FUDLR	84.94 ± 0.15	53.43 ± 1.06	58.44 ± 0.98	3.87 ± 1.52	2.39 ± 0.68	56.58 ± 6.69	66.02 ± 2.53	4.71
	TRACE-R	**85.57 ± 0.20**	53.91 ± 0.95	58.63 ± 0.85	1.82 ± 1.03	**1.05 ± 0.85**	64.77 ± 2.81	**70.73 ± 1.96***	**4.36**
	TRACE-C	85.23 ± 0.13	**54.14 ± 1.23***	58.67 ± 0.74	3.26 ± 1.49	2.33 ± 0.85	59.60 ± 6.01	66.34 ± 3.37	4.39
	TRACE-RC	85.24 ± 0.13	54.01 ± 0.62	**58.68 ± 0.63**	**0.77 ± 0.30***	0.73 ± 0.47*	**68.13 ± 1.32***	71.82 ± 0.89	4.67

```
The user has interacted with the following {item type}s,
ordered from oldest to newest:
1.  <image> "{title 1}"
⋮
t.  <image> "{title t}"
Candidate {item type}:
<image> "{title c}"
Will the user interact with the candidate {item type} next?
Answer only Yes or No.
```

E.2 Full Results of Attribute Bias Mitigation

Table 6 extends the attribute-bias evaluation in the main paper from gender to all three available user attributes in MovieLens, including gender, age, and occupation. The results further demonstrate the applicability of TRACE to different forms of group-level bias. For gender, TRACE-RC substantially reduces ΔNDCG@3 and achieves the highest F_{attr}. A similar pattern is observed for age: TRACE-RC obtains the lowest performance gaps, while TRACE-C better preserves the original recommendation utility. For occupation, TRACE-RC again achieves the lowest ΔNDCG@3 and competitive ΔNDCG@5, together with the strong fairness-utility balance measured by F_{attr}. In comparison, existing methods generally preserve recommendation accuracy but do not consistently reduce inter-group performance disparities across different attributes. These results complement the gender results reported in the main paper and show that the proposed framework is not tied to a particular sensitive attribute.

E.3 Multi-bias Mitigation

As our framework can be easily extended to a multi-bias case (Section 4.4), we evaluate a simultaneous debiasing performance on both popularity and gender attribute bias in Table 7. Note that only FUDLR is used for comparison, as it is originally designed to support multi-bias mitigation.

The results exhibit a pattern consistent with the single-bias experiments. TRACE-R strongly promotes long-tail exposure, but incurs a noticeable reduction in recommendation accuracy. TRACE-C better preserves utility while providing moderate improvements across the two bias dimensions.

Table 7: Co-existing bias mitigation on the MovieLens dataset. The best results are highlighted in bold, and the second best are underlined. The backbone results are listed as a reference to the original performance. * indicates the improvement is significant with $p < 0.05$.

Baseline	AUC ↑	NDCG@3 ↑	NDCG@5 ↑	TailP@3 ↑	TailP@5 ↑	ΔNDCG@3 ↓	ΔNDCG@5 ↓	F_{pop}@3 ↑	F_{pop}@5 ↑	F_{attr}@3 ↑	F_{attr}@5 ↑	Time ↓
Backbone	85.34 ± 0.14	54.57 ± 1.11	59.11 ± 0.69	12.90 ± 0.69	16.37 ± 0.18	2.36 ± 1.39	1.92 ± 0.67	20.85 ± 0.81	25.64 ± 0.16	63.31 ± 4.92	68.21 ± 2.72	9.44
FUDLR	**85.35 ± 0.14**	**54.53 ± 1.06**	**59.21 ± 0.62**	12.97 ± 0.81	16.29 ± 0.23	2.81 ± 1.62	2.20 ± 0.48	20.94 ± 0.98	25.55 ± 0.26	61.56 ± 5.98	67.28 ± 2.08	4.90
TRACE-R	83.83 ± 0.16	51.24 ± 1.37	56.22 ± 0.78	**15.67 ± 0.97***	**19.25 ± 0.50***	1.52 ± 2.01	2.28 ± 1.91	**23.98 ± 0.98***	**28.67 ± 0.51***	63.26 ± 6.43	64.42 ± 7.15	4.03
TRACE-C	84.96 ± 0.23	53.34 ± 1.51	57.83 ± 1.48	13.47 ± 0.81	17.12 ± 0.14	1.60 ± 1.54	1.57 ± 1.45	21.48 ± 0.90	26.41 ± 0.14	64.79 ± 4.18	68.21 ± 4.40	**4.02**
TRACE-RC	84.81 ± 0.12	53.76 ± 0.81	58.44 ± 1.49	13.60 ± 1.44	17.33 ± 0.40	**1.13 ± 0.67**	**0.91 ± 0.74***	21.69 ± 1.77	26.73 ± 0.32	**66.92 ± 2.54**	**71.11 ± 3.37***	4.23

Table 8: Modality-dominance debiasing on the MicroLens dataset. The best results are highlighted in bold, and the second best are underlined. The backbone results are listed as a reference to the original performance. * indicates the improvement is significant with $p < 0.05$.

	AUC ↑	NDCG@3 ↑	NDCG@5 ↑	WI-N@3 ↑	WI-N@5 ↑	F_{mod}@3 ↑	F_{mod}@5 ↑
Backbone	86.81 ± 0.03	55.69 ± 0.28	60.57 ± 0.26	50.04 ± 0.32	56.04 ± 0.70	52.71 ± 0.21	58.21 ± 0.34
Salience	86.26 ± 0.43	54.06 ± 1.14	59.23 ± 1.00	49.04 ± 2.66	54.54 ± 1.94	51.42 ± 1.96	56.78 ± 1.50
EliMRec	86.65 ± 0.04	54.39 ± 0.38	59.53 ± 0.45	48.97 ± 0.51	54.92 ± 0.58	51.54 ± 0.27	57.13 ± 0.28
MD	86.68 ± 0.05	54.66 ± 0.36	59.84 ± 0.40	49.54 ± 0.51	55.47 ± 0.58	51.98 ± 0.31	57.57 ± 0.16
TRACE-R	84.66 ± 0.08	49.49 ± 0.13	55.17 ± 0.30	42.33 ± 0.78	49.32 ± 0.08	45.63 ± 0.45	52.08 ± 0.15
TRACE-C	86.77 ± 0.04	**55.71 ± 0.42***	**60.58 ± 0.30***	**51.77 ± 1.75**	**57.14 ± 0.72***	**53.66 ± 1.09***	**58.81 ± 0.52***
TRACE-RC	**86.78 ± 0.04***	55.59 ± 0.36	60.43 ± 0.34	51.71 ± 1.19	57.11 ± 0.32	53.58 ± 0.74	58.72 ± 0.32

Combining the two operators, TRACE-RC achieves simultaneous improvements in popularity and attribute fairness while retaining competitive recommendation accuracy. In contrast, FUDLR closely preserves the original recommendation accuracy but does not improve fairness consistently. This indicates that the sample attribution is insufficient to identify the root causes of bias in MLLM-RSs, thereby hindering thorough bias mitigation.

E.4 Modality-Dominance Bias Mitigation

Beyond conventional popularity and user-attribute biases, we investigate modality-dominance bias, a bias specific to multimodal recommendation where the model may disproportionately rely on a prevalent modality and consequently disadvantage items whose information is weak in that modality (Shang et al., 2024). We adopt two representative modality-debiasing approaches, **EliMRec** (Liu et al., 2022) and **MD** (Shang et al., 2024), for comparison. We additionally implement **Salience**, a simple baseline that adjusts the recommendation score based on the salience of the dominant modality. To quantify the performance on disadvantaged items, we report WI-N@K, i.e., NDCG@K calculated over test samples with weak-image target items. We further report $F_{\text{mod}} = H(\text{NDCG@}K, \text{WI-N@}K)$ as the harmonic mean between overall recommendation utility and weak-modality performance.

As shown in Table 8, existing debiasing methods do not improve WI-N over the original backbone and are accompanied by varying degrees of utility degradation. In contrast, TRACE-C and TRACE-RC improve weak-image recommendation while almost fully preserving the original recommendation performance. In particular, TRACE-C increases WI-N and simultaneously achieves the highest F_{mod}. The similar performance of TRACE-RC further confirms this improvement. In contrast, TRACE-R alone performs poorly for modality-dominance bias, as simply suppressing bias-associated information may remove useful multimodal preference signals. Overall, these results demonstrate that TRACE can be instantiated beyond conventional item- and user-side biases to address bias arising from multimodal information itself.

E.5 Generalisation on Different Backbones

We further examine whether the effectiveness of TRACE generalises beyond the backbone used in the main experiments by replacing it with LLaVA and repeating both popularity- and attribute-bias mitigation. Tables 9 and 10 report the corresponding results.

Table 9: Popularity debiasing performance (%) and running time (h) on MicroLens with LLaVA backbone. The best results are highlighted in bold, and the second best are underlined. The backbone results are listed as a reference to the original performance. * indicates the improvement is significant with $p < 0.05$.

	AUC ↑	NDCG@3 ↑	NDCG@5 ↑	TailP@3 ↑	TailP@5 ↑	NPCG@3 ↑	NPCG@5 ↑	F_{pop}@3 ↑	F_{pop}@5 ↑
Backbone	87.05 ± 0.08	56.33 ± 0.71	61.13 ± 0.03	9.60 ± 0.23	13.68 ± 0.23	37.44 ± 1.13	42.92 ± 0.29	16.41 ± 0.37	22.35 ± 0.31
RW	87.22 ± 0.07	57.02 ± 1.46	**61.80 ± 1.29**	10.39 ± 0.19	14.05 ± 0.31	37.59 ± 2.37	43.68 ± 1.75	17.58 ± 0.22	22.89 ± 0.34
RR	87.15 ± 0.06	56.34 ± 0.37	61.24 ± 0.40	9.35 ± 0.50	13.04 ± 0.20	36.76 ± 0.85	42.71 ± 0.26	16.03 ± 0.75	21.49 ± 0.28
RWRR	**87.39 ± 0.06**	56.90 ± 1.24	61.61 ± 1.24	9.40 ± 0.45	13.10 ± 0.38	36.65 ± 1.65	42.70 ± 1.56	16.13 ± 0.68	21.60 ± 0.45
SPRec	85.14 ± 0.15	51.84 ± 0.13	57.17 ± 0.48	10.24 ± 0.04	16.01 ± 0.16	33.04 ± 0.28	40.13 ± 0.67	17.10 ± 0.05	25.01 ± 0.15
FairLRM	87.03 ± 0.06	56.51 ± 0.60	61.41 ± 0.48	9.95 ± 0.63	13.76 ± 0.23	37.34 ± 0.82	43.25 ± 0.61	16.92 ± 0.92	22.49 ± 0.32
FUDLR	87.06 ± 0.10	**57.46 ± 0.15**	61.78 ± 0.49	9.35 ± 0.69	13.69 ± 0.93	37.91 ± 1.32	43.09 ± 1.01	16.07 ± 1.03	22.40 ± 1.26
PDA	86.82 ± 0.06	55.56 ± 0.83	60.41 ± 0.81	10.52 ± 0.54	14.45 ± 0.17	37.53 ± 0.85	43.02 ± 0.68	17.69 ± 0.80	23.32 ± 0.17
MACR	86.89 ± 0.06	55.66 ± 0.86	60.59 ± 0.84	10.32 ± 0.63	14.23 ± 0.19	37.33 ± 0.86	42.93 ± 0.84	17.41 ± 0.93	23.04 ± 0.23
TRACE-R	86.94 ± 0.23	56.61 ± 0.58	61.27 ± 0.80	10.63 ± 0.74	14.32 ± 0.39	37.53 ± 1.68	43.39 ± 1.35	17.90 ± 1.07	23.21 ± 0.56
TRACE-C	86.68 ± 0.24	56.65 ± 0.29	61.08 ± 0.23	**12.34 ± 1.34***	**16.29 ± 0.98**	39.64 ± 0.71	44.74 ± 0.40	**20.25 ± 1.84***	**25.71 ± 1.22***
TRACE-RC	87.06 ± 0.07	57.37 ± 0.67	61.98 ± 0.37	11.98 ± 0.56	15.67 ± 0.51	**40.68 ± 1.11***	**45.78 ± 0.90***	19.81 ± 0.80	25.02 ± 0.67

Table 10: User attribute debiasing on the gender attribute in the MovieLens dataset with LLaVA backbone. The best results are highlighted in bold, and the second best are underlined. The backbone results are listed as a reference to the original performance. * indicates the improvement is significant with $p < 0.05$.

	AUC ↑	NDCG@3 ↑	NDCG@5 ↑	ΔNDCG@3 ↓	ΔNDCG@5 ↓	F_{attr}@3 ↑	F_{attr}@5 ↑
Backbone	87.05 ± 0.07	56.36 ± 0.75	61.56 ± 0.57	2.09 ± 2.63	2.50 ± 3.28	64.99 ± 9.77	66.02 ± 14.05
LLM-FairRec	86.70 ± 0.01	**56.82 ± 0.13**	**61.36 ± 0.46**	2.89 ± 1.93	2.83 ± 1.16	62.55 ± 7.65	65.89 ± 4.69
FUDLR	**87.03 ± 0.06**	57.38 ± 0.02	61.71 ± 0.42	2.07 ± 2.13	2.26 ± 3.32	65.73 ± 8.31	66.10 ± 15.75
TRACE-R	86.89 ± 0.03	56.08 ± 0.23	60.71 ± 0.22	0.54 ± 0.32	**0.26 ± 0.27***	70.42 ± 0.71	**74.80 ± 0.97***
TRACE-C	**87.03 ± 0.05***	56.37 ± 0.65	61.33 ± 0.43	1.45 ± 0.12	2.05 ± 0.36	67.95 ± 0.84	69.22 ± 1.08
TRACE-RC	86.74 ± 0.04	56.56 ± 0.79	60.91 ± 0.67	**0.12 ± 0.06**	0.57 ± 0.45	**71.92 ± 0.47**	73.99 ± 0.89

For popularity bias, TRACE-C provides a strong improvement in long-tail recommendation under the LLaVA backbone, resulting in the highest F_{pop} at both cutoffs. More importantly, these gains are obtained while maintaining comparable AUC and NDCG. The gender-debiasing results show a similarly clear fairness improvement. TRACE-RC reduces ΔNDCG@3 and achieves the highest F_{attr}@3, while TRACE-R obtains the lowest ΔNDCG@5 and highest F_{attr}@5. By comparison, LLM-FairRec and FUDLR largely retain the original inter-group disparity. Taken together, the results across popularity and attribute biases demonstrate that the identification and mitigation mechanisms of TRACE remain effective when instantiated on a different multimodal backbone, rather than depending on a particular MLLM architecture.

E.6 Additional Ablation Study

Table 11 provides additional ablations for gender attribute debiasing on MovieLens. The results highlight the importance of jointly learning identifiable and sparse concepts and causally connecting them to the bias outcome. Replacing the proposed concept representation with AE, iAE, or SAE consistently weakens the final fairness-utility trade-off. Among these alternatives, iAE performs relatively better, suggesting the importance of concept identifiability for defining meaningful interventions. Nevertheless, without the sparse concept structure used by TRACE, its overall performance remains below the complete framework. Removing the SCM can reduce ΔNDCG to some extent, but causes a substantial degradation in recommendation accuracy. This indicates that identifying bias-related concepts alone is insufficient. Explicitly modelling their causal relationship with the bias outcome is important for distinguishing bias-driving information from useful preference information.

The mitigation ablations further reveal complementary roles of the two proposed operators. TRACE-R provides direct suppression of identified bias-driving information but may sacrifice recommendation utility, whereas TRACE-C substantially better preserves accuracy while improving fairness.

Table 11: Ablation study of gender attribute debiasing on MovieLens.

Variant	AUC ↑	NDCG@3 ↑	NDCG@5 ↑	ΔNDCG@3 ↓	ΔNDCG@5 ↓	F_{attr}@3 ↑	F_{attr}@5 ↑
Identification							
AE	84.90	53.00	57.36	3.66	3.62	57.72	60.41
iAE	<u>85.06</u>	<u>53.77</u>	57.98	2.19	2.39	63.70	65.83
SAE	84.73	51.34	56.70	3.25	2.39	58.34	64.98
w/o SCM	83.90	48.55	54.17	<u>1.13</u>	1.80	62.52	67.73
Mitigation							
TRACE-R	84.08	50.28	55.75	2.00	1.61	61.22	66.70
TRACE-C	**85.14**	**53.95**	**58.21**	1.53	<u>1.31</u>	<u>65.72</u>	<u>69.43</u>
TRACE-RC	84.88	53.47	<u>58.05</u>	**0.56**	**1.18**	**68.25**	**69.97**

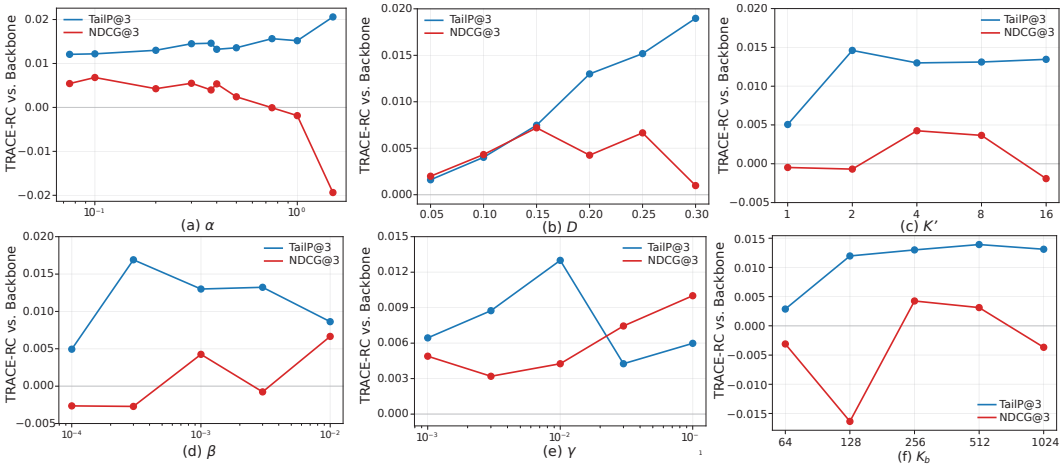


Figure 4: Parameter sensitivity of popularity debiasing on MicroLens.

Their combination in TRACE-RC achieves the smallest group disparities and the highest F_{attr} at both cutoffs. Together, the identification and mitigation ablations support the complete design of TRACE: identifiable sparse concept learning provides a structured intervention space; causal modelling determines how these concepts contribute to bias, and the complementary mitigation operators perform bias reduction while preserving useful recommendation signals.

E.7 Additional Parameter Analysis

We investigate the sensitivity of TRACE-RC to its key parameters, including the removal weight α, correlation weight D, sparsity weight β, conditional-prior weight γ, the number of selected concepts K', and the number of bias-driving samples K_b. Figure 4 and Figure 5 report the results for popularity debiasing on MicroLens and gender attribute debiasing on MovieLens, respectively. All values are shown as performance changes relative to the original backbone. Thus, values above (below) zero indicate an increase (decrease) after mitigation.

Popularity debiasing. As shown in Figure 4, the parameters generally exhibit a clear trade-off between bias mitigation and recommendation utility. (a) **Removal weight α.** Increasing α strengthens the removal of the identified bias direction. TailP@3 generally increases with α, indicating progressively stronger exposure correction. In contrast, NDCG@3 eventually decreases when α becomes large, with a particularly sharp degradation at the largest value. This confirms that overly aggressive removal may suppress preference-relevant information together with the bias signal. (b) **Correlation weight D.** Increasing D generally strengthens tail exposure, while its effect on NDCG@3 is non-monotonic. Moderate values provide a favourable trade-off between the two objectives, whereas further increasing D continues to improve TailP@3 but no longer consistently preserves accuracy. (c) **Number of selected concepts K'.** Using only one concept yields a relatively weak improvement in tail exposure. Increasing K' substantially strengthens the debiasing effect, after which TailP@3 remains relatively stable. NDCG@3, however, varies across K', suggesting that incorporating additional concept directions can introduce both preference-relevant and preference-irrelevant effects.

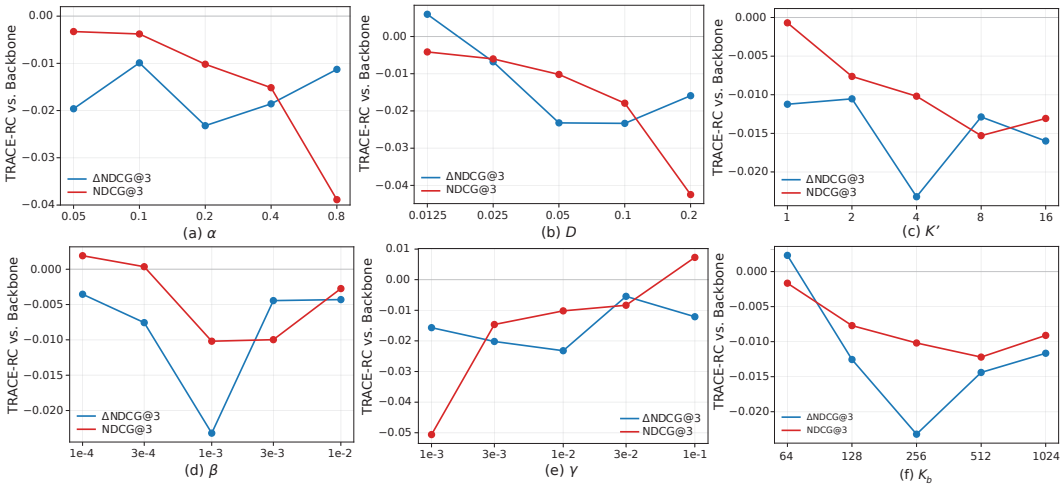


Figure 5: Parameter sensitivity of gender attribute debiasing on MovieLens.

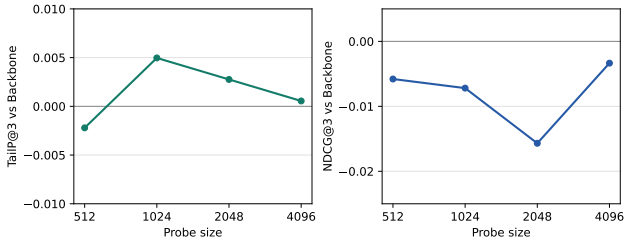


Figure 6: Impact of the size of probe set on MovieLens.

The results indicate that selecting a moderate number of concepts is sufficient for effective correction. (d) **Sparsity weight β.** Very small β provides insufficient sparsity for isolating distinct concept directions, while overly large β may suppress useful exposure-related concepts. The results therefore favour an intermediate sparsity level, where tail exposure is substantially improved without a large accuracy loss. (e) **Conditional-prior weight γ.** The best balance is observed at an intermediate γ. Increasing γ from a small value first strengthens the identified debiasing direction and improves TailP@3, whereas a larger value shifts the resulting edit away from exposure improvement and toward accuracy preservation. This demonstrates the importance of appropriately weighting the bias-relevant conditional prior when identifying the editable concept space. (f) **Number of bias-driving samples K_b.** TailP@3 improves rapidly as K_b increases from a small value and then largely saturates, indicating that a moderate number of highly influential samples already captures most of the target popularity-bias signal. Meanwhile, NDCG@3 shows a non-monotonic response, suggesting that including more samples does not necessarily improve the utility-fairness trade-off.

Attribute debiasing. Figure 5 further evaluates the same parameters for gender attribute debiasing. (a) Increasing the **removal weight** α generally reduces ΔNDCG@3, demonstrating stronger mitigation of the inter-group performance gap, but a large α also causes a noticeable decrease in NDCG@3. (b) A similar pattern is observed for the **correlation weight** D: stronger correction generally reduces the group disparity, while excessive reinforcement incurs increasing accuracy degradation. (c) For the **number of selected concepts** K', the disparity reduction becomes considerably stronger when moving beyond a very small concept set, while the corresponding accuracy also drops. (d) For the **sparsity weight** β, intermediate values produce the largest reduction in ΔNDCG@3, whereas both weaker and stronger sparsity lead to smaller fairness improvements. This supports the role of sparsity in isolating editable bias-related concepts without excessively discarding useful information. (e) The **conditional-prior weight** γ exhibits a clear trade-off. Increasing γ substantially recovers recommendation accuracy, but simultaneously weakens the reduction in ΔNDCG@3. It suggests that γ controls how strongly the learned concept representation is anchored to the bias-relevant auxiliary variable and consequently affects the fairness-utility balance of the resulting edit. (f) Increasing the **number of bias-driving samples** K_b initially strengthens disparity reduction,

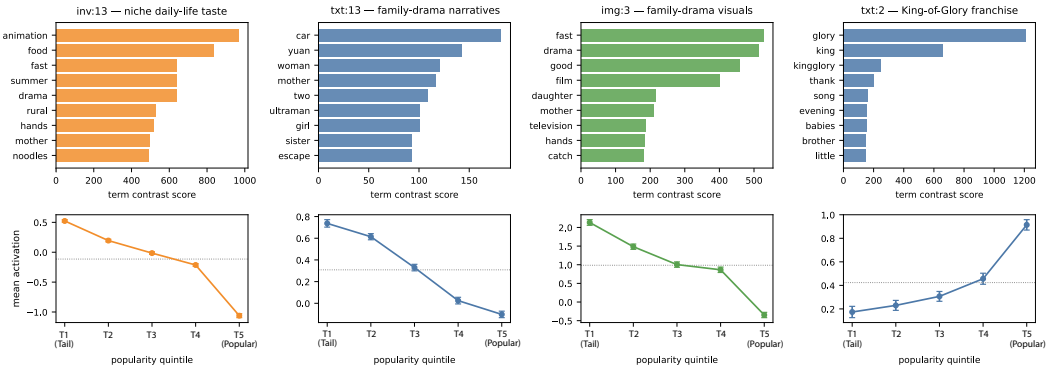


Figure 7: Concept validation on MicroLens. Top: high-vs-low activation term contrasts for four concepts spanning the subspace. Bottom: mean activation by popularity quintile. The dotted line is the global mean.

with the largest reduction occurring at an intermediate value. Further increasing K_b weakens this gain while partially recovering accuracy, suggesting that lower-ranked samples contain weaker or less specific bias signals.

Probe-set size. We further examine the impact of the probe-set size n_p on MovieLens in Figure 6. The results show that increasing n_p does not monotonically improve debiasing performance. A very small probe set ($n_p = 512$) provides insufficient coverage and slightly decreases TailP@3, whereas increasing it to 1,024 yields the larger tail-exposure improvement with a moderate accuracy loss. Further enlarging the probe set gradually reduces the TailP gain, while $n_p = 4{,}096$ better preserves NDCG@3 but provides limited exposure improvement. Overall, a moderate probe size provides a favourable balance between debiasing effectiveness and utility, while also supporting the computational efficiency of performing identification on a small subset rather than the full training set.

Overall sensitivity. Across the two bias settings, the results reveal a consistent pattern. Parameters controlling the *strength* of the intervention (α, D, K', and K_b) generally strengthen debiasing up to a moderate range but may sacrifice recommendation utility when the intervention becomes overly broad or aggressive. Meanwhile, β and γ affect the quality and bias relevance of the learned concept space, and intermediate values generally provide a more favourable balance.

E.8 Inspection of Identified Bias Concepts

Beyond evaluating the downstream debiasing performance, we further inspect the intermediate outputs of the identification module to understand what information is captured by the learned concepts and how these concepts are associated with the bias outcome. We focus on popularity bias on MicroLens and examine the identified concepts from three complementary perspectives: their semantic content, their activation patterns across item-popularity levels, and their causal relationship in the discovered SCM. The semantic names used below are post-hoc descriptions assigned according to the representative high-activation samples/terms for ease of interpretation.

Semantic coherence of the learned concepts. Figure 7 visualises four representative concepts spanning different modalities and preference components. For each concept, we contrast terms associated with high- and low-activation samples to obtain a human-interpretable view of its semantics. The resulting concepts exhibit coherent rather than arbitrary patterns. For example, preference-invariant concept `inv:13` is characterised by terms such as “animation”, “food”, “summer”, “rural”, and “noodles”, suggesting a broad *niche daily-life taste*. In the candidate-text space, `txt:13` is dominated by terms related to family roles and narrative elements, such as “woman”, “mother”, “girl”, and “sister”, and is therefore interpreted as a *family-drama narrative* concept. A related candidate-image concept, `img:3`, is associated with drama/film and family-related visual content,

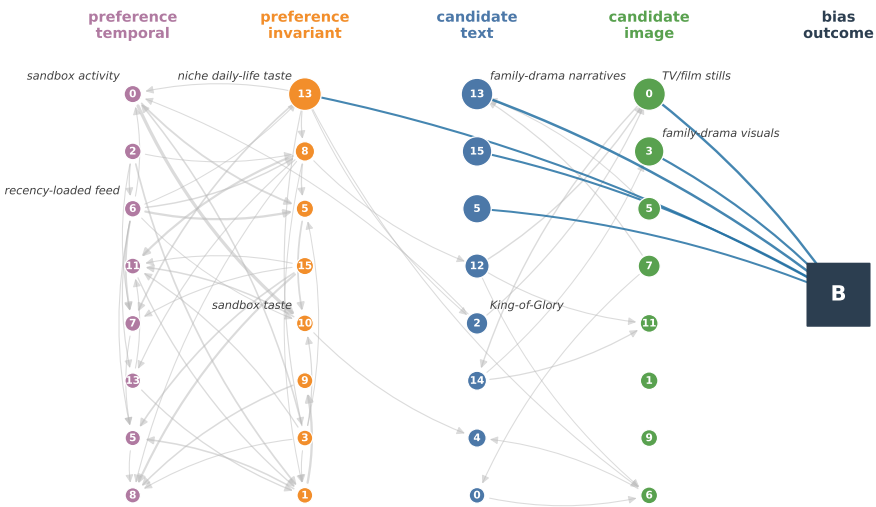


Figure 8: The discovered SCM on MicroLens, where node size $\propto |\tau|$, grey edges represent concept $\rightarrow$ concept with width $\propto$ weight, blue edges are the weight direct parents of the bias outcome B.



Figure 9: Concept validation on MicroLens. Top: high-vs-low activation term contrasts for four concepts spanning the subspace. Bottom: mean activation by popularity quintile. The dotted line is the global mean.

providing a cross-modal counterpart that we describe as *family-drama visuals*. In contrast, `txt:2` is strongly characterised by terms such as "glory", "king", and "kingglory", revealing a substantially different semantic direction related to the *King-of-Glory franchise*.

These examples indicate that the sparse concept space separates different factors qualitatively rather than merely encoding a single popularity dimension. It contains preference-related concepts as well as textual and visual concepts, providing an interpretable intermediate space in which the source and propagation of recommendation bias can be further examined.

Concept activation is associated with popularity. The bottom row of Figure 7 reports the mean activation of each concept across five item-popularity quintiles, from the tail (T1) to the most popular group (T5). The dotted line denotes the global mean activation of the corresponding concept. Clear and heterogeneous popularity-dependent patterns emerge. For `inv:13`, `txt:13`, and `img:3`, activation generally decreases from tail to popular items. In particular, the activation of `inv:13` changes from clearly positive for tail items to strongly negative for the most popular group, consistent with its interpretation as a niche-oriented preference direction. In contrast, `txt:2` exhibits the opposite pattern: its activation increases toward the popular quintiles and rises sharply for T5, consistent with a franchise-related concept that is concentrated among more popular content.

Cross-modal relationship of the discovered bias pathways. Figure 8 visualises the SCM learned over the identified concepts and the popularity-bias outcome B. The discovered structure reveals that the bias outcome is associated with multiple components rather than a single modality. Direct pathways originate from preference-invariant, textual, and visual concepts, while additional concepts can contribute indirectly through cross-concept dependencies. This provides an important rationale for the multi-level and cross-modal design of TRACE: popularity bias can be represented through both preference signals and semantic/visual characteristics, and these factors may interact before affecting the final bias outcome.

The graph also provides semantically meaningful examples of such pathways. The preference-invariant *niche daily-life taste* concept appears among the influential preference-side concepts, while family-drama-related factors are recovered independently from candidate text and image representations. Their presence in the discovered graph suggests that the SCM can capture cross-modal semantic factors that participate in the learned popularity structure. Meanwhile, the *King-of-Glory* textual concept provides an example of a concept whose activation is concentrated toward popular items, further illustrating that the identified bias space contains both tail- and popularity-associated semantic directions.

Visual inspection of modality-specific concepts. We additionally inspect the original images corresponding to samples with the highest and lowest activation for representative visual concepts, as shown in Figure 9. For the family-drama visual concept `img:3`, high-activation examples consistently contain TV/drama scenes, including close-up characters and interpersonal interactions. By contrast, low-activation examples predominantly contain game interfaces, animated characters, or action-oriented game scenes. This qualitative separation agrees with the term-based interpretation of the same concept and provides direct evidence that its activation corresponds to recognisable visual content rather than an arbitrary latent coordinate.

Temporal propagation. We further inspect how preference-related concepts evolve along the user history and propagate into the more stable preference representation. To this end, we divide each interaction sequence into early, middle, and late thirds and compare the activation of temporal concepts across the three stages. Some concepts remain nearly unchanged over time. For example, $\mathrm{tmp}{:}15$ has mean activations $2.56/2.59/2.52$, suggesting that it captures a relatively persistent preference pattern. In contrast, other concepts exhibit clear temporal dynamics. Concept $\mathrm{tmp}{:}6$ increases from -2.02 to 4.85 to 7.63, indicating that this signal becomes progressively stronger in more recent interactions.

The discovered SCM further shows that these temporal signals can propagate into time-invariant preference concepts rather than directly affecting the bias outcome. Among the strongest outgoing relations from the temporal block are $\mathrm{tmp}{:}0 \rightarrow \mathrm{inv}{:}10$, linking recent sandbox-gaming activity to a stable sandbox-related preference. In the discovered graph, the direct parents of the bias outcome B come from the invariant and candidate blocks, while temporal concepts influence B through these intermediate preference representations. This provides a temporal interpretation of the identified bias pathway: short-term interaction patterns can gradually reshape the user's stable preference representation, which then participates in the downstream bias-related decision together with candidate-side concepts. This observation motivates identifying both temporal and time-invariant preference concepts in the identification space, so that the resulting mitigation can target user-side propagation pathways in addition to candidate-side popularity signals.

Overall interpretation. Taken together, these analyses provide a coherent view of the identification process. First, the learned sparse dimensions correspond to recognisable preference, textual, and visual semantics. Second, their activations exhibit distinct relationships with item popularity, including both tail-associated and popular-associated directions. Third, the discovered SCM connects a subset of these concepts across preference and candidate modalities to the bias outcome, exposing both direct and mediated pathways. Finally, inspection of representative images independently confirms the semantic meaning of visual concepts. These observations support the central motivation of TRACE: bias in MLLM-RSs is not merely reflected in the final recommendation scores, but is distributed across semantically meaningful concepts whose propagation can be explicitly identified and subsequently targeted for mitigation.

Table 12: Debiasing cases for the popularity axis. Each row is one test sample: the user's history items (with popularity tier), the target item, and the first-ranked candidate of the backbone, of a baseline RR, and of the proposed TRACE-RC. The number in parentheses is the rank of the target item given by the method. Popular tier items are marked in orange, middle tier items are in grey, and tail items are in blue.

Case	User history (oldest → newest)	Target item	Backbone: top-1 (rank)	Baseline: top-1 (rank)	TRACE-RC: top-1 (rank)
MicroLens					
Tail-loyal history	Intern pranks staff with bad-luck stickers; Star hides in an anti-fan community; "This mother is too cruel" summer film; Naughty boy, catching up on a good show; Talk-show special	"Laugh and tears" warm-heart drama recommendation	Fans abused at a female idol's concert (2)	Fans abused at a female idol's concert (2)	**Target item** (1)
Mixed history	Foodie kebab challenge; Badass League; Minecraft: SapNap mutiny; Correcting apprentices' homework; "Ray lion" fan tag video	Annai son / Anlei fan meme, "serious ooc apology"	Kingsman / Borderlands 3 mission (3)	Kingsman / Borderlands 3 mission (3)	**Target item** (1)
Preference match	Minecraft animated soundtrack; Duel-arena strange operations; Youth "watergate" counter attitude; Liu Huaqiang BGM electric guitar; "Draw next time, brothers!" #Naruto mobile	"I made ···" tower-defence game	"The most tolerant player in Genshin" (4)	"We live in order not to regret" #Genshin (4)	"Want to lose your skills?" #Naruto mobile game (3)
Amazon Baby					
Tail-loyal history	BusyBee stroller wagon; Cute-pig night light; Century modular stroller; Skip Hop baby nest; Graco Nautilus car seat	AutoBrush Ellie the Elephant silicone teether	Glow-in-the-dark ceiling star stickers (2)	Glow-in-the-dark ceiling star stickers (2)	**Target item** (1)
Mixed history	Ely's & Co. swaddle blanket; Boon Lawn bottle drying rack; Baby Merlin's Magic Sleepsuit; Simyke WiFi baby monitor; RUVALINO diaper-bag backpack	Baby Delight Alpine portable bouncer, 0–6 months	Frida Baby MediFrida pacifier medicine dispenser (2)	Frida Baby MediFrida pacifier medicine dispenser (2)	**Target item** (1)
Latent preference	Dr. Brown's anti-colic bottles; Tommee Tippee bottle; Tommee Tippee formula bottle; Munchkin dishwasher basket; Chicco 360 hook-on chair	Chicco KeyFit 30 infant car seat and base	Luvable Friends hooded towel (2)	Luvable Friends hooded towel (2)	**Target item** (1)
MovieLens-1M					
Mixed preference	Tommy Boy; Tora! Tora! Tora!; I.Q.; Grumpy Old Men; Grumpier Old Men	Out to Sea (1997; Lemmon–Matthau)	Who Framed Roger Rabbit? (2)	Who Framed Roger Rabbit? (2)	**Target item** (1)
1990s comedy history	The Butcher's Wife; Batman Forever; The Waterboy; Robin Hood: Men in Tights; A Night at the Roxbury	Air Bud	Inspector Gadget (4)	Inspector Gadget (4)	**Target item** (1)
Preference match	Annie Hall; To Kill a Mockingbird; Amadeus; Chinatown; One Flew Over the Cuckoo's Nest	The Spanish Prisoner (1997)	Crumb (1994) (4)	Crumb (1994) (4)	Carnival of Souls (1962) (3)

Table 13: Debiasing cases for the gender attribute axis on MovieLens. On this test split the backbone serves male users worse. Each row is one test sample: the user's history items, the target item, and the first-ranked candidate of the backbone, of the baseline FUDLR / CFP, and of the proposed TRACE-RC. The number in parentheses is the rank of the target item given by the method.

Case	User history (oldest → newest)	Target item	Backbone: top-1 (rank)	FUDLR / CFP: top-1 (rank)	TRACE-C: top-1 (rank)
Male user; period-drama target	Splash; The Mask; Devil in a Blue Dress; Dead Again; Jumanji	Anna and the King (1999)	Down by Law (1986) (3)	Down by Law (3) / Down by Law (2)	**Target item** (1)
Male user; thriller history	Halloween; The Texas Chainsaw Massacre; The Silence of the Lambs; Aliens; L.A. Confidential	The Usual Suspects	Instinct (1999) (2)	Instinct (2) / Instinct (2)	**Target item** (1)
Female user; taste-consistent	The Scarlet Letter (1926); Dangerous Liaisons; Who's Afraid of Virginia Woolf?; Fight Club; Braveheart	The Next Best Thing (2000)	Wild Wild West (3)	Wild Wild West (3) / Wild Wild West (2)	Barry Lyndon (1975) (2)

E.9 Case Study

To qualitatively examine how the identified bias mitigation changes individual recommendations, we present representative cases for popularity and user-attribute debiasing in Tables 12 and 13,

respectively. Rather than only examining whether the recommended item changes, we compare the popularity composition and semantic preference reflected in users' histories, targets, and top-ranked predictions. The cases illustrate how TRACE modifies biased ranking decisions while retaining preference-relevant information.

Popularity debiasing. Table 12 presents representative examples from all three datasets. Several observations can be made.

First, TRACE-RC can correct cases where a tail target is suppressed by a more popular candidate. For example, in the *tail-loyal* MicroLens case, most historical interactions and the target belong to the tail tier, yet both the backbone and baseline rank a middle-tier item first and place the target second. After mitigation, TRACE-RC promotes the tail target to rank 1. Similar behaviour appears on Amazon Baby, where the tail target is promoted over a middle-tier candidate, and on MovieLens, where the tail item *Out to Sea* is promoted over the popular *Who Framed Roger Rabbit?*. These examples show that the original ranking can favour relatively popular candidates even when less popular items are genuinely preferred with the observed interaction history, while TRACE-RC can alleviate such popularity-driven ranking effects.

Second, the correction is not restricted to users with uniformly tail-oriented histories. In the mixed-history examples, the histories contain items from different popularity tiers, yet the target tail items are still ranked below more popular candidates by the backbone and baseline. TRACE-RC instead ranks the targets first.

Third, debiasing does not amount to blindly promoting every tail item. The preference-match cases provide useful counterexamples. For MicroLens, although the history is dominated by popular items, TRACE-RC selects a middle-tier Naruto-related item that is semantically consistent with a recent Naruto interaction rather than simply selecting a tail candidate. Similarly, in the MovieLens case with a history of well-known films, the selected candidate changes without forcing the target to rank first. These cases are consistent with the role of preference reinforcement in TRACE-RC.

User-attribute debiasing. Table 13 further provides representative examples for gender debiasing on MovieLens. In this evaluation split, the backbone shows a larger recommendation error for male users, motivating mitigation of the performance gap across gender groups.

For the first two male-user cases, the backbone and baseline methods fail to rank the ground-truth target first. In contrast, TRACE promotes the target to rank 1. These examples illustrate how mitigating the identified attribute-associated directions and reinforcing preference modelling can improve rankings for samples from the disadvantaged group. In the female-user example, TRACE changes the top-ranked item from *Wild Wild West* to *Barry Lyndon* and improves the target from rank 3 to rank 2. Together with the two male examples, this suggests that the intervention modifies attribute-associated bias while preserving individual preference information.

Overall, the case studies provide qualitative evidence complementary to the aggregate evaluation. For popularity bias, TRACE can recover relevant less-popular items. For attribute bias, it can effectively improve inaccurate rankings associated with the disadvantaged group. These observations are consistent with the design of TRACE, where identified bias-related directions guide the mitigation while preference reinforcement helps preserve recommendation utility.

F Broader Impacts

Recommender systems increasingly mediate how people discover information, products, entertainment, services, and opportunities. Consequently, bias in recommendations can have impacts extending beyond differences in metric scores. For users, biased recommendations may result in unequal quality of service, restricted exposure to relevant content, or systematic disparities across demographic or other user groups. For item providers and content creators, bias can disproportionately allocate visibility and opportunities to already popular items while marginalising long-tail providers. At scale, such effects can lead to economic consequences through unequal traffic, engagement, and commercial opportunities, and can reinforce existing social inequalities through repeated exposure and feedback loops. Depending on the application and jurisdiction, unfair treatment associated with sensitive user attributes may also raise legal, regulatory, and compliance concerns. Understanding

and mitigating recommendation bias is therefore important not only for recommendation quality, but also for protecting the interests of multiple stakeholders and supporting trustworthy deployment of recommender systems.

These concerns become particularly important for MLLM-based recommender systems. By jointly modelling user interactions and multimodal item content, MLLM-RSs introduce complex pathways through which biases may be encoded, combined across modalities, and ultimately reflected in recommendation outcomes. Their increasing model scale and representation complexity can also make such mechanisms difficult to diagnose using output-level fairness measurements alone. This work aims to contribute toward more transparent and fair MLLM-RSs by tracing measured bias from training samples through modality-specific latent concepts to recommendation outcomes. This can provide model-level evidence about where a measured bias originates and, in turn, support more targeted mitigation rather than treating the model as a black box.

The proposed mitigation operators further provide a training-free mechanism for intervening on identified bias pathways while explicitly considering preference preservation. Avoiding complete model retraining can reduce the computational cost of investigating and correcting undesirable behaviour in large multimodal recommenders, potentially making iterative bias auditing and mitigation more practical. More broadly, connecting bias identification, explanation, and mitigation may help practitioners examine not only whether a recommender exhibits a particular measured bias, but also how that bias is represented and which parts of the training data contribute to it.

At the same time, the framework should not be interpreted as providing a universal definition or guarantee of fairness. We therefore view TRACE as a diagnostic and intervention tool to support, rather than replace, human judgment. Real-world deployment should combine model-level analysis with stakeholder-informed fairness objectives, privacy protection, appropriate governance, and continuous monitoring for unintended consequences.