Geodesic Expert Routing for Unbiased Knowledge Distillation in Recommendation

Xuan Zhang[1] , **Rongchuan Wei**[1] , **Chunyu Wei**[2,*] , **Hongxing Yuan**[1,*] and **Yushun Fan**[1]
[1]Department of Automation, Tsinghua University, Beijing National Research Center for Information Science and Technology, China
[2]School of Information, Renmin University of China, China
{xuan-zha23, wrc20, yuanhx24}@mails.tsinghua.edu.cn, weichunyu@ruc.edu.cn, fanyus@tsinghua.edu.cn

Abstract

Knowledge distillation has become a prevalent technique for deploying efficient recommender systems, enabling lightweight student models to approximate the performance of larger teachers. However, we identify a critical issue: distillation systematically amplifies popularity bias, as student models inherit and intensify the popularity-driven shortcuts encoded in teachers trained on interaction data dominated by popular items. To address this limitation, we propose GUIDE (Geodesic aware Unbiased Instructive Distillation with Experts), a collaborative distillation framework that incorporates domain-specific debiasing experts alongside the global teacher. GUIDE tackles two key challenges in this paradigm. First, for expert routing, we introduce Spherical Expert Alignment, which conducts expert-student matching on the spherical manifold with geodesic distance optimization, eliminating magnitude-induced bias and ensuring stable gradient flow. Second, for context fusion, a Meta-Debiasing Gate is designed to dynamically arbitrate teacher-expert influence using real-time context via end-to-end amortized meta-learning. Extensive experiments on multiple real-world datasets demonstrate that GUIDE significantly mitigates popularity bias while preserving recommendation accuracy, with state-of-the-art trade-offs among efficiency, accuracy, and fairness.

1 Introduction

Recommender systems have become indispensable infrastructure for modern online platforms, serving as the primary mechanism through which users discover products, content, and services aligned with their preferences [Deldjoo *et al.*, 2024]. By filtering the overwhelming volume of available options into personalized suggestions, these systems significantly enhance user experience across diverse domains such as e-commerce, streaming media, and social networks [Park *et al.*, 2026; Gao *et al.*, 2023]. As data volume and item diversity continue to scale exponentially, system designers face an increasingly critical trade-off between recommendation accuracy and computational efficiency [Zhang *et al.*, 2019].

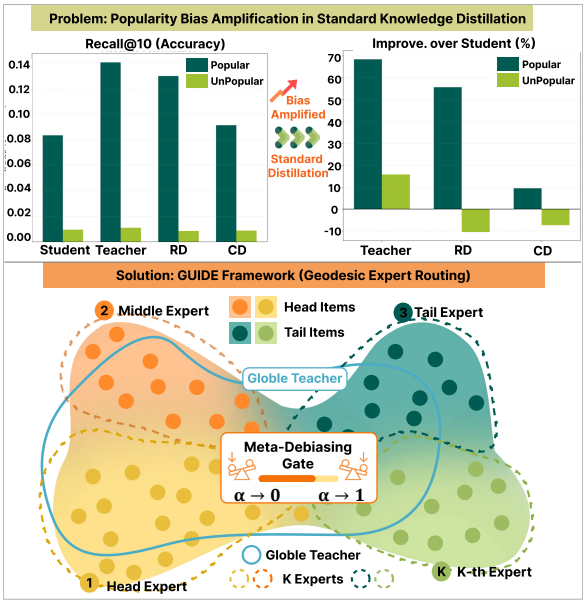

Figure 1: (Top) Recall@10 comparison on popular/unpopular items. Relative improvements over the base Student (directly learned from the data) reveal that standard distillation amplifies popularity bias. (Bottom) The proposed GUIDE dynamically fuses knowledge from global and expert teachers to balance accuracy and fairness.

Knowledge Distillation (KD) has emerged as a widely adopted paradigm to address this efficiency challenge [Kang *et al.*, 2020]. By transferring knowledge from a computationally expensive teacher model to a lightweight student model, KD enables the deployment of compact recommenders that retain much of the predictive capability of their larger counterparts while satisfying stringent latency requirements [Lee *et al.*, 2019; Kweon *et al.*, 2021]. This approach has demonstrated remarkable success across various recommendation architectures, from collaborative filtering to sequential models [Sarwar *et al.*, 2001; Kang *et al.*, 2021]. However, we identify a critical yet underexplored issue: *knowledge distillation systematically amplifies popularity bias in the stu-*

*Corresponding author
The code and data are available at: https://github.com/zx19971219/GUIDE

dent model. Popularity bias, where models disproportionately favor frequently-interacted items at the expense of long-tail content, is already pervasive in recommender systems trained on implicit feedback [Oestreicher-Singer and Sundararajan, 2012; Abdollahpouri *et al.*, 2019]. Our empirical investigation reveals that this bias is not merely preserved but substantially magnified during the distillation process. We argue that the root cause lies in the reliance on a single global teacher trained on complete interaction data that is inherently dominated by popularity patterns. Figure 1 shows KD improvements are driven by popular items at the expense of unpopular ones. By mimicking such a teacher, the student internalizes and amplifies these popularity-driven shortcuts, resulting in reduced exposure diversity and exacerbated unfairness toward niche items [Lee *et al.*, 2019; Tang and Wang, 2018].

To address this limitation, we propose **GUIDE** (**G**eodesic-aware **U**nbiased **I**nstructive **D**istillation with **E**xperts), a novel framework that augments conventional single-teacher distillation with collaborative guidance from domain-specific debiasing experts. Our key insight is that while the global teacher provides comprehensive knowledge essential for accuracy, targeted experts trained on strategically partitioned data subsets can serve as corrective signals that counteract popularity bias. Specifically, we partition the original interaction data into multiple subsets based on item popularity strata and train specialized expert models in parallel. During distillation, the student model learns not only from the global teacher but also from these experts, enabling balanced knowledge acquisition that promotes both accuracy and fairness. However, transitioning from single-teacher distillation to this collaborative teacher-expert paradigm introduces two fundamental challenges:

- **Expert Routing**: How to select the most appropriate expert for each prediction context? A natural approach measures expert-student similarity via cosine or inner product in Euclidean space [Koren *et al.*, 2009]. However, embedding norms encode popularity-correlated intensity information, as popular items naturally exhibit larger magnitudes due to more frequent gradient updates [Zhu *et al.*, 2021; Wang *et al.*, 2017]. This causes the student to be attracted to high-norm experts rather than semantically aligned ones. While post-hoc ℓ_2 normalization offers an implicit remedy, it remains geometrically imprecise and optimization-unfriendly.
- **Context Fusion**: How to dynamically balance knowledge from the teacher and experts based on varying user-item characteristics? Static fusion weights are fundamentally inadequate, since susceptibility to popularity bias varies substantially across user-item contexts depending on user engagement diversity, item popularity characteristics, and interaction patterns [Han *et al.*, 2024; Abdollahpouri *et al.*, 2021]. Determining appropriate debiasing intensity requires real-time contextual reasoning rather than a fixed arbitration strategy.

To tackle **Expert Routing**, we propose *Spherical Expert Alignment*, which conducts expert-student matching directly on the spherical manifold $\mathcal{S}^{d-1}$, ensuring that all comparisons reflect pure directional similarity free from magnitude contamination. Crucially, we optimize geodesic distance rather than cosine angle: cosine-based gradients scale with $\sin(\theta)$ and vanish as $\theta \to 0$, whereas geodesic gradients remain linear in θ, enabling stable optimization even when student and expert representations are closely aligned.

To address **Context Fusion**, we design a *Meta-Debiasing Gate*, a lightweight gating network that ingests contextual features such as user history diversity, item popularity percentile, and interaction density. The gate outputs a dynamic fusion weight $\alpha \in [0, 1]$ that arbitrates the balance between teacher and expert influence. It is trained end-to-end via amortized meta-learning to explicitly optimize for both accuracy and debiasing effectiveness.

Our main contributions are summarized as follows:

- We identify the phenomenon of popularity bias amplification in knowledge distillation and propose GUIDE, a collaborative distillation framework that integrates domain-specific debiasing experts alongside the global teacher to enable accuracy-fairness balanced knowledge transfer.
- We introduce Spherical Expert Alignment for geometry-aware expert routing and Meta-Debiasing Gate for context-aware dynamic fusion, addressing the challenges of expert selection and adaptive knowledge arbitration respectively.
- Extensive experiments on real-world datasets demonstrate that GUIDE significantly mitigates popularity bias while maintaining competitive accuracy, achieving state-of-the-art trade-offs among efficiency, accuracy, and fairness.

2 RELATED WORK

2.1 Knowledge Distillation for Recommendation

Knowledge Distillation (KD) enhances recommender systems by transferring knowledge from intermediate layers [Romero *et al.*, 2015], privileged features [Xu *et al.*, 2020], or structured graphs [Zhang *et al.*, 2020]. Beyond efficiency [Zhu *et al.*, 2020] and collaborative learning [Gou *et al.*, 2024], KD is widely applied to mitigate biases. For instance, [Liu *et al.*, 2020] and UKDRec [Yang *et al.*, 2025] leverage counterfactual data and multi-teacher frameworks for debiasing. UNKD [Chen *et al.*, 2023a] and [Ding *et al.*, 2022] address popularity bias and model fusion, respectively. Research also optimizes distillation losses, ranging from ranking-based top-N selection [Tang and Wang, 2018] and instance construction [Lee *et al.*, 2019] to list-level losses [Kang *et al.*, 2020], extending to heterogeneous graph integration [Tao *et al.*, 2022; Wang *et al.*, 2021].

2.2 Bias in Recommendation

Recommendation bias arises from user behavior and feedback loops [Chen *et al.*, 2023b]. Mitigation strategies typically employ IPS [Joachims *et al.*, 2018; Li *et al.*, 2023; Schnabel *et al.*, 2016], Doubly Robust methods [Song *et al.*, 2023], or causal graphs [Liu *et al.*, 2021; Liu *et al.*, 2023; He *et al.*, 2022] on biased data [Ning *et al.*, 2024; Zhang and Shen, 2023], or leverage unbiased data via meta-learning [Chen *et al.*, 2021] and distillation [Liu *et al.*, 2020]. Beyond data, architectures [Lin *et al.*, 2019] and optimizers [Tang *et*

al., 2020] contribute to bias. Notably, KD itself can aggravate popularity bias [Chen *et al.*, 2023a]: a teacher model trained on skewed data transfers inherent biases to the student, undermining debiasing efforts [Yang *et al.*, 2025].

3 Preliminaries

3.1 Problem Formulation

Let $\mathcal{U} = \{u_1, u_2, \ldots, u_M\}$ denote the set of M users and $\mathcal{I} = \{i_1, i_2, \ldots, i_N\}$ denote the set of N items. The user-item interaction matrix $\mathbf{R} \in \{0,1\}^{M \times N}$ encodes implicit feedback, where $r_{ui} = 1$ indicates that user u has interacted with item i, and $r_{ui} = 0$ otherwise. For each user u, we denote the set of interacted items as $\mathcal{I}_u^+ = \{i \in \mathcal{I} \mid r_{ui} = 1\}$. The primary objective of a recommender system is to learn a scoring function $f : \mathcal{U} \times \mathcal{I} \to \mathbb{R}$ that predicts user preferences over unobserved items.

3.2 Knowledge Distillation in Recommendation

Knowledge Distillation (KD) aims to transfer knowledge from a high-capacity teacher model $\mathcal{T}$ to a lightweight student model $\mathcal{S}$. Given a user-item pair (u, i), the teacher produces a soft prediction $\hat{y}_{ui}^{\mathcal{T}}$ that encodes rich relational knowledge beyond binary labels. The student is trained to mimic these soft targets while maintaining computational efficiency. The distillation objective typically minimizes the divergence between teacher and student outputs:

$$\mathcal{L}_{KD} = \sum_{(u,i) \in \mathcal{D}} \ell\left(\hat{y}_{ui}^{\mathcal{S}}, \hat{y}_{ui}^{\mathcal{T}}\right), \tag{1}$$

where $\ell(\cdot, \cdot)$ denotes a divergence measure such as Mean Squared Error or KL divergence, and $\mathcal{D}$ is the training set.

3.3 Spherical Manifold

The $(d-1)$-dimensional unit spherical manifold $\mathcal{S}^{d-1}$ is defined as the set of all d-dimensional vectors with unit ℓ_2 norm:

$$\mathcal{S}^{d-1} = \left\{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\right\}. \tag{2}$$

The spherical manifold provides an embedding space where vector magnitudes are normalized, enabling the model to focus purely on directional relationships. For any vector $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, the projection onto $\mathcal{S}^{d-1}$ is given by:

$$\pi(\mathbf{x}) = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}. \tag{3}$$

The intrinsic distance on the spherical manifold is the geodesic distance, representing the length of the shortest arc connecting two points. For $\mathbf{x}, \mathbf{y} \in \mathcal{S}^{d-1}$, the distance is:

$$d_{\text{geo}}(\mathbf{x}, \mathbf{y}) = \arccos\left(\mathbf{x}^\top \mathbf{y}\right). \tag{4}$$

4 Methodology

Figure 2 illustrates the overall framework of GUIDE. Given user-item interaction data, GUIDE first partitions the item set into K popularity-based subsets to construct domain-specific expert teachers. To address the **Expert Routing** challenge, we project both users and expert centroids onto a spherical manifold, where expert relevance is quantified via geodesic distance. To address the **Context Fusion** challenge, a Meta-Debiasing Gate dynamically arbitrates between expert and global teacher knowledge based on real-time contextual features. The final distillation signal is derived as a context-aware ensemble tailored to each user-item interaction. We detail each component in the following subsections.

4.1 Domain-Specific Expert Construction

Relying on a single global teacher causes popularity bias amplification in knowledge distillation. A global teacher trained on the complete interaction data inevitably inherits the long-tail distribution, leading to overfitting on popular items while under-representing niche content. To mitigate this limitation, we construct multiple domain-specific experts, each specializing in a distinct popularity stratum.

Popularity-Based Data Partitioning

We quantify the popularity of each item i by its interaction frequency $d_i = |\{u \in \mathcal{U} \mid r_{ui} = 1\}|$. Based on these frequency values, we partition the item set $\mathcal{I}$ into K disjoint subsets $\{\mathcal{I}^{(1)}, \mathcal{I}^{(2)}, \ldots, \mathcal{I}^{(K)}\}$, where each subset corresponds to a distinct popularity tier. Specifically, we sort all items by d_i in descending order and divide them into K groups of approximately equal size, yielding a spectrum from head (high-popularity) to tail (low-popularity) items. Each item subset $\mathcal{I}^{(k)}$ induces a corresponding interaction subset:

$$\mathcal{D}^{(k)} = \left\{(u, i) \mid i \in \mathcal{I}^{(k)}, r_{ui} = 1\right\}. \tag{5}$$

This partitioning creates isolated training environments that shield tail item learning from head item interference.

Expert Teacher Training

For each partition $\mathcal{D}^{(k)}$, we train a dedicated expert teacher $\mathcal{T}^{(k)}$ exclusively on interactions within $\mathcal{I}^{(k)}$. This constraint ensures that each expert captures the intrinsic collaborative patterns specific to its popularity domain without being dominated by global popularity shortcuts. Formally, the expert $\mathcal{T}^{(k)}$ is optimized via:

$$\mathcal{T}^{(k)} = \arg\min_{\theta} \sum_{(u,i,j) \in \mathcal{D}^{(k)}} \mathcal{L}_{\text{BPR}}\left(f_\theta(u, i), f_\theta(u, j)\right), \tag{6}$$

where j denotes a negative sample drawn from $\mathcal{I}^{(k)}$. In parallel, we train a global teacher $\mathcal{T}^{(g)}$ on the complete interaction data $\mathcal{D}$ to preserve comprehensive collaborative signals.

Remark. The expert construction procedure is model-agnostic: any recommendation architecture (e.g., matrix factorization, neural collaborative filtering, or sequential models) can serve as the backbone for both expert and student models. This flexibility ensures broad applicability across diverse recommendation scenarios.

4.2 Spherical Expert Alignment

With K expert teachers constructed, the next challenge lies in dynamically routing each user to the most semantically aligned expert. Conventional approaches measure expert-student similarity via cosine or inner product in Euclidean space. However, as discussed in Section 1, embedding magnitudes encode popularity-correlated intensity information

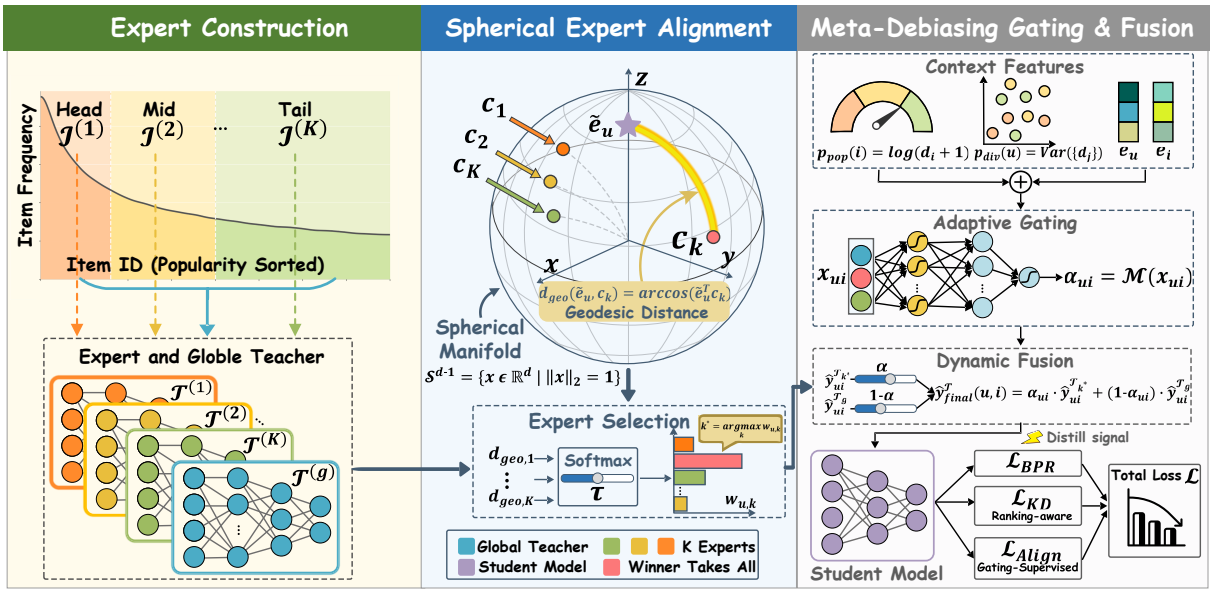

Figure 2: **Overall of GUIDE framework.** It consists of three key parts: (1) Expert Construction, which trains specialized teachers on popularity-based data partitions; (2) Spherical Expert Alignment, which routes users to experts via geodesic distance on a spherical manifold to eliminate magnitude bias; and (3) Meta-Debiasing Gating & Fusion, which adaptively fuses expert and global knowledge based on interaction context for personalized distillation.

that confounds directional similarity. We address this challenge through Spherical Expert Alignment, which conducts matching directly on the spherical manifold $\mathcal{S}^{d-1}$.

Expert Centroid Representation

To enable efficient expert routing, we represent each expert $\mathcal{T}^{(k)}$ by a semantic centroid that summarizes its domain knowledge. Let $\mathbf{E}^{(k)} = \{\mathbf{e}_i \mid i \in \mathcal{I}^{(k)}\}$ denote the set of item embeddings learned by expert $\mathcal{T}^{(k)}$. We compute the expert centroid $\mathbf{c}_k$ by aggregating these embeddings and projecting onto the spherical manifold:

$$\mathbf{c}_k = \pi\left(\frac{1}{|\mathcal{I}^{(k)}|}\sum_{i\in\mathcal{I}^{(k)}}\mathbf{e}_i\right) = \frac{\sum_{i\in\mathcal{I}^{(k)}}\mathbf{e}_i}{\left\|\sum_{i\in\mathcal{I}^{(k)}}\mathbf{e}_i\right\|_2}, \tag{7}$$

where $\mathbf{c}_k \in \mathcal{S}^{d-1}$ serves as the representative prototype for the k-th expert domain.

User Projection and Geodesic Matching

For any user u with embedding $\mathbf{e}_u$ learned by the student model, we project it onto the same spherical manifold:

$$\tilde{\mathbf{e}}_u = \pi(\mathbf{e}_u) = \frac{\mathbf{e}_u}{\|\mathbf{e}_u\|_2}. \tag{8}$$

The affinity between user u and expert k is then quantified by the geodesic distance on $\mathcal{S}^{d-1}$:

$$d_{\text{geo}}(\tilde{\mathbf{e}}_u, \mathbf{c}_k) = \arccos\left(\tilde{\mathbf{e}}_u^\top \mathbf{c}_k\right). \tag{9}$$

This geodesic formulation offers two key advantages over cosine similarity. First, constraining representations to the unit sphere strictly decouples embedding magnitude—often correlated with item popularity—from semantic direction, ensuring pure content-based routing. Second, unlike cosine gradients which scale with $\sin(\theta)$ and vanish as $\theta \to 0$, the geodesic gradient $\nabla_v \mathcal{L}_{geo} \propto \frac{\partial\theta^2}{\partial\theta}\frac{\partial\theta}{\partial\cos\theta}\nabla_v \cos\theta$ introduces a $1/\sin(\theta)$ term that perfectly cancels the projection-induced $\sin(\theta)$ decay. This yields stable gradients linear in θ (i.e., $\nabla\mathcal{L} \propto 2\theta$) even for closely aligned representations.

Expert Selection via Soft Routing

Based on geodesic distances, we compute a probability distribution over experts for each user u:

$$w_{u,k} = \frac{\exp\left(-\tau \cdot d_{\text{geo}}(\tilde{\mathbf{e}}_u, \mathbf{c}_k)\right)}{\sum_{j=1}^{K}\exp\left(-\tau \cdot d_{\text{geo}}(\tilde{\mathbf{e}}_u, \mathbf{c}_j)\right)}, \tag{10}$$

where $\tau > 0$ is a temperature parameter controlling the sharpness of the distribution. To minimize noise from less relevant experts, we adopt a winner-takes-all strategy that selects the expert with the highest probability:

$$k^* = \arg\max_{k\in\{1,\ldots,K\}} w_{u,k}. \tag{11}$$

The selected expert $\mathcal{T}^{(k^*)}$ provides the specialized soft label $\hat{y}_{ui}^{\mathcal{T}_{k^*}}$ for the current user-item pair.

4.3 Meta-Debiasing Gate

While Spherical Expert Alignment identifies the optimal domain-specific expert for each user, relying solely on expert knowledge may be suboptimal. Expert teachers, trained on isolated popularity strata, may lack the broader collaborative signals captured by the global teacher. Conversely, the global teacher provides robust general knowledge but inherits popularity bias. To dynamically balance these complementary knowledge sources, we design a Meta-Debiasing Gate that adaptively arbitrates their influence based on real-time contextual features.

Context Feature Extraction

The optimal balance between expert and global teacher knowledge depends on the specific characteristics of each user-item interaction. We extract a context feature vector $\mathbf{x}_{ui}$ that captures two key factors:

Item Popularity. We quantify the global exposure of candidate item i using its log-normalized interaction frequency:

$$p_{\text{pop}}(i) = \log(d_i + 1). \tag{12}$$

This feature informs the gate about the item's position in the popularity spectrum.

User Diversity. We measure the diversity of user u's historical interactions by computing the variance of popularity values across interacted items:

$$p_{\text{div}}(u) = \text{Var}\left(\{d_j \mid j \in \mathcal{I}_u^+\}\right). \tag{13}$$

This feature indicates the user's tendency to explore across different popularity tiers. The context vector concatenates these features with learned user and item embeddings, where $\oplus$ denotes concatenation:

$$\mathbf{x}_{ui} = \left[\mathbf{e}_u \oplus \mathbf{e}_i \oplus p_{\text{pop}}(i) \oplus p_{\text{div}}(u)\right]. \tag{14}$$

Adaptive Gating Mechanism

The Meta-Debiasing Gate $\mathcal{M}$ uses a lightweight multi-layer perceptron (MLP) with sigmoid activation, mapping the context vector to a scalar fusion coefficient $\alpha_{ui} \in (0, 1)$:

$$\alpha_{ui} = \mathcal{M}(\mathbf{x}_{ui}) = \sigma\left(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \mathbf{x}_{ui} + \mathbf{b}_1) + \mathbf{b}_2\right), \tag{15}$$

where $\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2$ are learnable parameters, and $\sigma(\cdot)$ denotes the sigmoid function.

The coefficient α_{ui} quantifies the necessity of debiasing for the current context: higher values indicate that the model should prioritize expert knowledge (e.g., for long-tail items or users with diverse histories), while lower values suggest relying on the global teacher for stability.

Dynamic Knowledge Fusion

The final distillation target is synthesized as a convex combination of soft labels from the selected expert $\mathcal{T}^{(k^*)}$ and the global teacher $\mathcal{T}^{(g)}$:

$$\hat{y}_{\text{final}}^{\mathcal{T}}(u, i) = \alpha_{ui} \cdot \hat{y}_{ui}^{\mathcal{T}_{k^*}} + (1 - \alpha_{ui}) \cdot \hat{y}_{ui}^{\mathcal{T}_g}. \tag{16}$$

This dynamic fusion ensures that the distillation signal is adaptively tailored to each user-item interaction, effectively resolving the trade-off between accuracy and fairness.

4.4 Optimization

We formulate a multi-objective loss function that jointly optimizes recommendation accuracy, knowledge transfer, and gate calibration:

$$\mathcal{L} = \mathcal{L}_{\text{BPR}} + \lambda_{\text{KD}}\mathcal{L}_{\text{KD}} + \lambda_{\text{Align}}\mathcal{L}_{\text{Align}} + \lambda_{\text{Reg}}\|\Theta\|_2^2, \tag{17}$$

where λ_{KD}, λ_{Align}, and λ_{Reg} are hyperparameters balancing each component.

Recommendation Loss. We employ the BPR loss to ensure the student captures fundamental collaborative signals:

$$\mathcal{L}_{\text{BPR}} = \sum_{(u,i,j)\in\mathcal{D}} -\ln \sigma\left(\hat{y}_{ui}^{\mathcal{S}} - \hat{y}_{uj}^{\mathcal{S}}\right), \tag{18}$$

where (u, i, j) denotes a triplet with positive item $i \in \mathcal{I}_u^+$ and negative item $j \notin \mathcal{I}_u^+$.

Ranking-Aware Distillation Loss. Standard pointwise distillation may fail to preserve the nuanced ranking relationships essential for recommendation. We propose a ranking-aware objective that aligns the preference margins between student and teacher:

$$\mathcal{L}_{\text{KD}} = \sum_{(u,i,j)\in\mathcal{D}} \left(\Delta p_{uij}^{\mathcal{S}} - \Delta p_{uij}^{\mathcal{T}}\right)^2, \tag{19}$$

where $\Delta p_{uij}^{\mathcal{S}} = \sigma(\hat{y}_{ui}^{\mathcal{S}}) - \sigma(\hat{y}_{uj}^{\mathcal{S}})$ and $\Delta p_{uij}^{\mathcal{T}} = \sigma(\hat{y}_{\text{final}}^{\mathcal{T}}(u, i)) - \sigma(\hat{y}_{\text{final}}^{\mathcal{T}}(u, j))$ represent the preference margins from the student and fused teacher, respectively.

Gate Alignment Loss. To prevent the gate coefficient α_{ui} from collapsing to trivial solutions, we introduce an explicit supervisory signal. Let $t_i = \mathbb{I}(i \in \mathcal{I}_{\text{tail}})$ be an indicator for tail items. We employ binary cross-entropy to guide the gate:

$$\mathcal{L}_{\text{Align}} = -\sum_{(u,i)\in\mathcal{D}} \left[t_i \ln(\alpha_{ui}) + (1 - t_i)\ln(1 - \alpha_{ui})\right]. \tag{20}$$

This objective encourages the gate to prioritize expert knowledge ($\alpha_{ui} \to 1$) for tail items while retaining global knowledge ($\alpha_{ui} \to 0$) for popular items.

4.5 Complexity Analysis

Training Complexity. Given batch size B, embedding dimension d, and K experts, the primary computational costs include: (1) student forward pass: $O(Bd)$; (2) geodesic distance computation: $O(BKd)$; and (3) Meta-Debiasing Gate: $O(Bd^2)$. Since K is typically small (e.g., $K \leq 5$), and the winner-takes-all strategy activates only the selected expert and global teacher, the overall training complexity remains linear in batch size.

Inference Complexity. During inference, all auxiliary components (expert teachers, spherical projection, Meta-Debiasing Gate) are discarded. Only the lightweight student model is deployed, yielding inference complexity identical to the base architecture with no additional latency.

5 Experiments

We conduct comprehensive experiments to evaluate GUIDE, aiming to answer the following research questions:

- **RQ1:** How does GUIDE perform compared to state-of-the-art knowledge distillation methods for recommendation?
- **RQ2:** What are the individual contributions of each proposed component?
- **RQ3:** How sensitive is GUIDE to key hyperparameters?
- **RQ4:** How effective is GUIDE in handling popularity bias and enhancing item fairness?

5.1 Experimental Setup

Datasets. We evaluate our model on three public benchmarks: **CiteULike** (scholarly articles), **Amazon-Movie**, and **Amazon-Game** (e-commerce). Following standard preprocessing, we retain users with at least 20 interactions and split

Table 1: Dataset statistics after preprocessing.

Statistic	CiteULike	Amazon-Movie	Amazon-Game
#Users	5,219	123,960	24,303
#Items	25,181	50,052	10,672
#Interactions	125,580	1,697,533	231,780
Sparsity	99.90%	99.97%	99.91%

the data into training, validation, and test sets using an 8:1:1 ratio. Detailed statistics are reported in Table 1.

Baselines. We compare GUIDE against five state-of-the-art distillation methods: (1) *Ranking-based*: **RD** [Tang and Wang, 2018] and **CD** [Lee *et al.*, 2019] transfer ranking knowledge via position-aware weighting and rank-aware sampling; (2) *Structure-aware*: **DE-RRD** [Kang *et al.*, 2020] and **HTD** [Kang *et al.*, 2021] distill latent knowledge and topological structures; and (3) *Bias-aware*: **UNKD** [Chen *et al.*, 2023a] mitigates popularity bias via stratified distillation.

Evaluation Metrics. We adopt two widely-used top-K metrics: **Recall@**K (R@K), which measures the proportion of relevant items retrieved, and **NDCG@**K (N@K), which accounts for ranking position via discounted cumulative gain. We report results for $K \in \{10, 20\}$.

Implementation Details. All hyperparameters are tuned based on NDCG@20 on the validation set. The teacher uses 3-layer LightGCN, for students, we evaluate 2-layer GCN and BPRMF. The learning rate is searched over $\{0.0002, 0.001, 0.005, 0.01\}$. The distillation coefficient λ_{KD} is selected from $\{0.01, 0.02, ..., 0.9, 0.99\}$, and the alignment coefficient λ_{Align} from $\{0.0005, 0.001, ..., 0.9, 0.99\}$. The number of experts K is set to 5 by default. All experiments are repeated 5 times with different random seeds, and we report the mean and standard deviation.

5.2 Overall Performance (RQ1)

Table 2 presents the overall performance comparison. From these results, we draw the following observations:

GUIDE consistently outperforms all baselines across datasets and metrics. The improvements are particularly pronounced on the sparser Amazon-Game dataset, demonstrating GUIDE's robustness in challenging scenarios where popularity bias is most detrimental. Remarkably, GUIDE surpasses the teacher on several metrics, indicating that filtering popularity noise yields superior generalization.

Ranking-based methods exhibit limited effectiveness. RD yields the poorest performance among all methods, as simple ranking alignment neglects the rich semantic relationships essential for recommendation. This underscores the necessity of more sophisticated knowledge transfer mechanisms.

Structure-aware methods provide substantial improvements over ranking-based approaches. CD, HTD, and UNKD significantly outperform RD by incorporating structural knowledge. Notably, CD achieves competitive results by distilling collaborative similarity distributions, while UNKD leverages unobserved feedback for enhanced supervision.

The advantages of GUIDE stem from its two core innovations. Spherical Expert Alignment eliminates magnitude-induced bias in expert routing, enabling semantically meaningful knowledge selection. The Meta-Debiasing Gate further adapts the fusion strategy to each context, ensuring debiasing preserves accuracy on well-represented items.

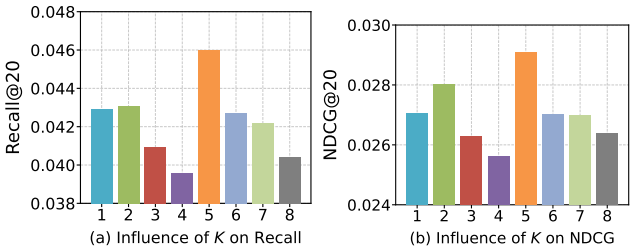

Figure 3: Effect of the number of experts K on Amazon-Game.

5.3 Ablation Study (RQ2)

To validate the contribution of each component, we compare GUIDE against four ablated variants on Amazon-Game:

- **w/o Spherical**: Replaces spherical manifold matching with Euclidean distance-based expert selection.
- **w/o Meta-Gate**: Substitutes the adaptive gate with a fixed fusion weight ($\alpha = 0.5$).
- **w/o Ranking KD**: Replaces the ranking-aware distillation loss with standard pointwise MSE.
- **w/o Alignment**: Removes the gate alignment loss $\mathcal{L}_{\text{Align}}$.

As shown in Table 3, the full GUIDE model consistently outperforms all variants, confirming the necessity of each component: (i) *w/o Spherical* shows that Euclidean embeddings are prone to magnitude-induced bias, emphasizing the need for directional alignment; (ii) *w/o Meta-Gate* confirms that static fusion is insufficient for handling user-item heterogeneity; (iii) *w/o Ranking KD* proves that pointwise distillation fails to capture nuanced preference margins; and (iv) *w/o Alignment* indicates that explicit gate supervision is vital to avoid trivial solutions.

5.4 Hyperparameter Sensitivity (RQ3)

We examine three key hyperparameters: (1) **Number of experts** K: As shown in Figure 3, performance peaks around $K = 5$; too few experts fail to capture popularity diversity, while excessive K introduces noise due to data sparsity. (2) **Distillation coefficient** λ_{KD}: Figure 4 (Left) shows optimal result at 0.8, balancing knowledge transfer and student adaptation. (3) **Alignment coefficient** λ_{Align}: Figure 4 (Right) indicates that $\lambda_{\text{Align}} = 0.01$ is optimal, while larger values may conflict with the primary ranking objective.

5.5 Debiasing Effectiveness (RQ4)

To evaluate the effectiveness of GUIDE in addressing popularity bias and enhancing fairness, we conduct three complementary analyses.

Addressing Popularity Bias and Fairness. We evaluate GUIDE on Amazon-Game (stratified into Head/Tail groups) and CiteULike (fairness metrics). Table 4 presents the results. GUIDE achieves substantial improvements on Tail items (+6.21%) while simultaneously improving Head performance (+2.75%). On CiteULike, GUIDE significantly

Table 2: Performance comparison on three datasets (Student: LightGCN). Best student results are in **bold**, second-best are underlined. "Improv." denotes the relative improvement of GUIDE over the best baseline. Blue shading indicates teacher results.

Dataset	Metric	Teacher	RD	CD	DE-RRD	HTD	UNKD	GUIDE	Improv.
CiteULike	R@10	0.0964	$0.062_{\pm.006}$	$0.086_{\pm.005}$	$0.062_{\pm.001}$	$0.089_{\pm.002}$	$0.093_{\pm.001}$	$\mathbf{0.097}_{\pm.001}$	+4.3%
	R@20	0.1456	$0.085_{\pm.009}$	$0.129_{\pm.006}$	$0.092_{\pm.003}$	$0.140_{\pm.001}$	$0.132_{\pm.004}$	$\mathbf{0.145}_{\pm.002}$	+3.6%
	N@10	0.0788	$0.051_{\pm.005}$	$0.075_{\pm.002}$	$0.050_{\pm.002}$	$0.076_{\pm.001}$	$0.078_{\pm.000}$	$\mathbf{0.079}_{\pm.000}$	+1.3%
	N@20	0.0937	$0.058_{\pm.006}$	$0.087_{\pm.003}$	$0.060_{\pm.001}$	$0.091_{\pm.001}$	$0.089_{\pm.002}$	$\mathbf{0.094}_{\pm.001}$	+3.3%
Amazon-Movie	R@10	0.0284	$0.020_{\pm.001}$	$0.025_{\pm.003}$	$0.022_{\pm.001}$	$0.027_{\pm.002}$	$0.027_{\pm.001}$	$\mathbf{0.029}_{\pm.001}$	+7.4%
	R@20	0.0487	$0.033_{\pm.003}$	$0.044_{\pm.004}$	$0.037_{\pm.001}$	$0.045_{\pm.002}$	$0.046_{\pm.001}$	$\mathbf{0.049}_{\pm.000}$	+6.5%
	N@10	0.0313	$0.023_{\pm.002}$	$0.028_{\pm.002}$	$0.024_{\pm.001}$	$0.030_{\pm.001}$	$0.030_{\pm.001}$	$\mathbf{0.031}_{\pm.002}$	+3.3%
	N@20	0.0394	$0.028_{\pm.005}$	$0.036_{\pm.003}$	$0.030_{\pm.000}$	$0.037_{\pm.001}$	$0.037_{\pm.002}$	$\mathbf{0.039}_{\pm.001}$	+5.4%
Amazon-Game	R@10	0.0282	$0.018_{\pm.001}$	$0.025_{\pm.002}$	$0.020_{\pm.002}$	$0.026_{\pm.001}$	$0.025_{\pm.001}$	$\mathbf{0.029}_{\pm.001}$	+11.5%
	R@20	0.0460	$0.028_{\pm.003}$	$0.043_{\pm.003}$	$0.028_{\pm.001}$	$0.040_{\pm.002}$	$0.041_{\pm.002}$	$\mathbf{0.046}_{\pm.002}$	+6.9%
	N@10	0.0232	$0.016_{\pm.001}$	$0.020_{\pm.002}$	$0.017_{\pm.000}$	$0.021_{\pm.001}$	$0.021_{\pm.001}$	$\mathbf{0.023}_{\pm.000}$	+9.5%
	N@20	0.0299	$0.020_{\pm.001}$	$0.027_{\pm.002}$	$0.022_{\pm.000}$	$0.026_{\pm.001}$	$0.027_{\pm.001}$	$\mathbf{0.029}_{\pm.002}$	+7.4%

Table 3: Ablation study on Amazon-Game. Each row removes one component from the full GUIDE model.

Variant	R@10	R@20	N@10	N@20
w/o Spherical	0.0259	0.0422	0.0206	0.0266
w/o Meta-Gate	0.0259	0.0419	0.0209	0.0270
w/o Ranking KD	0.0258	0.0420	0.0206	0.0267
w/o Alignment	0.0261	0.0419	0.0210	0.0270
GUIDE (Full)	**0.0288**	**0.0465**	**0.0230**	**0.0299**

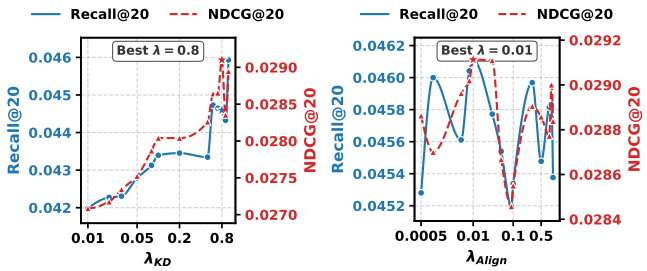

Figure 4: Sensitivity to λ_{KD} and λ_{Align} on Amazon-Game.

Table 4: Performance on Amazon-Game (Head/Tail) and fairness metrics on CiteULike.

	Amazon-Game			CiteULike			
Model	Head	Tail	Δ Tail	Gini ↓	Coverage ↑	Tail Ratio ↑	R@20
UNKD	0.0509	0.00177	–	0.9679	0.1234	0.0845	0.1331
GUIDE	**0.0523**	**0.00188**	**+6.21%**	**0.9239**	**0.2521**	**0.0876**	**0.1447**

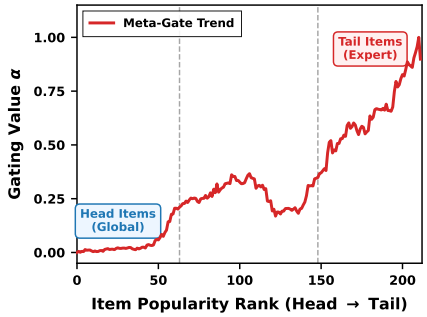

Figure 5: Distribution of gating coefficient α across item popularity. The gate adaptively shifts from global teacher ($\alpha \to 0$) for popular items to expert knowledge ($\alpha \to 1$) for tail items.

lowers the Gini index and doubles Item Coverage. These results across datasets prove that GUIDE effectively handles popularity bias and improves fairness while maintaining superior overall accuracy.

Visualization of Adaptive Gating. To verify that the Meta-Debiasing Gate learns meaningful context-aware strategies, we visualize the distribution of gating coefficients α_{ui} across item popularity on the Amazon-Game test set. As shown in Figure 5, a clear trend emerges: $\alpha \approx 0$ for popular (Head) items, indicating reliance on the global teacher, while $\alpha \approx 1$ for long-tail (Tail) items, shifting toward expert knowledge. This confirms that the gate learns a principled popularity-aware arbitration strategy rather than random assignment, validating the effectiveness of the alignment loss $\mathcal{L}_{\text{Align}}$.

Case Study: Cold-Start Item Recommendation. To evaluate performance under extreme sparsity, we analyze a cold-start interaction (User #8137, Item #10872) in Amazon-Game. While the LightGCN teacher fails to recommend the item (Rank $>$ 200) due to absent collaborative signals, GUIDE achieves Rank 1 with a high gating coefficient ($\alpha_{ui} = 0.84$). This confirms that GUIDE adaptively prioritizes specialized expert knowledge to compensate for data scarcity, effectively enabling zero-shot cold-start recommendations.

6 Conclusion

Existing KD approaches, using standard Euclidean teachers, often suffer from bias amplification. We propose GUIDE, projecting diverse expert knowledge onto a Spherical Manifold. This design eliminates magnitude-induced noise via distinct subspace experts, while our Meta-Debiasing Gate intelligently arbitrates between global and specialized supervision to adapt to dynamic user preferences. Experiments confirm that GUIDE consistently outperforms competing methods in both recommendation accuracy and bias mitigation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62173199 and No.62506366).

References

[Abdollahpouri *et al.*, 2019] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Managing popularity bias in recommender systems with personalized re-ranking. In Roman Barták and Keith W. Brawner, editors, *Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference, Sarasota, Florida, USA, May 19-22 2019*, pages 413–418. AAAI Press, 2019.

[Abdollahpouri *et al.*, 2021] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, Bamshad Mobasher, and Edward C. Malthouse. User-centered evaluation of popularity bias in recommender systems. In Judith Masthoff, Eelco Herder, Nava Tintarev, and Marko Tkalcic, editors, *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, Utrecht, The Netherlands, June, 21-25, 2021*, pages 119–129. ACM, 2021.

[Chen *et al.*, 2021] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. Autodebias: Learning to debias for recommendation. In *SIGIR*, pages 21–30. ACM, 2021.

[Chen *et al.*, 2023a] Gang Chen, Jiawei Chen, Fuli Feng, Sheng Zhou, and Xiangnan He. Unbiased knowledge distillation for recommendation. In *WSDM*, pages 976–984. ACM, 2023.

[Chen *et al.*, 2023b] 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 Trans. Inf. Syst.*, 41(3):67:1–67:39, 2023.

[Deldjoo *et al.*, 2024] Yashar Deldjoo, Zhankui He, Julian J. McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. A review of modern recommender systems using generative models (gen-recsys). In Ricardo Baeza-Yates and Francesco Bonchi, editors, *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6448–6458. ACM, 2024.

[Ding *et al.*, 2022] Sihao Ding, Fuli Feng, Xiangnan He, Jinqiu Jin, Wenjie Wang, Yong Liao, and Yongdong Zhang. Interpolative distillation for unifying biased and debiased recommendation. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 40–49. ACM, 2022.

[Gao *et al.*, 2023] Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, and Yong Li. A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *Trans. Recomm. Syst.*, 1(1):1–51, 2023.

[Gou *et al.*, 2024] Jianping Gou, Liyuan Sun, Baosheng Yu, Lan Du, Kotagiri Ramamohanarao, and Dacheng Tao. Collaborative knowledge distillation via multiknowledge transfer. *IEEE Trans. Neural Networks Learn. Syst.*, 35(5):6718–6730, 2024.

[Han *et al.*, 2024] Yishan Han, Biao Xu, Yao Wang, and Shanxing Gao. Towards popularity-aware recommendation: A multi-behavior enhanced framework with orthogonality constraint. *CoRR*, abs/2412.19172, 2024.

[He *et al.*, 2022] Ming He, Xinlei Hu, Changshu Li, Xin Chen, and Jiwen Wang. Mitigating confounding bias for recommendation via counterfactual inference. In *ECML/PKDD (1)*, volume 13713 of *Lecture Notes in Computer Science*, pages 524–540. Springer, 2022.

[Joachims *et al.*, 2018] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5284–5288. ijcai.org, 2018.

[Kang *et al.*, 2020] SeongKu Kang, Junyoung Hwang, Wonbin Kweon, and Hwanjo Yu. DE-RRD: A knowledge distillation framework for recommender system. In *CIKM*, pages 605–614. ACM, 2020.

[Kang *et al.*, 2021] SeongKu Kang, Junyoung Hwang, Wonbin Kweon, and Hwanjo Yu. Topology distillation for recommender system. In *KDD*, pages 829–839. ACM, 2021.

[Koren *et al.*, 2009] Yehuda Koren, Robert M. Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[Kweon *et al.*, 2021] Wonbin Kweon, SeongKu Kang, and Hwanjo Yu. Bidirectional distillation for top-k recommender system. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3861–3871. ACM / IW3C2, 2021.

[Lee *et al.*, 2019] Jae-woong Lee, Minjin Choi, Jongwuk Lee, and Hyunjung Shim. Collaborative distillation for top-n recommendation. In *ICDM*, pages 369–378. IEEE, 2019.

[Li *et al.*, 2023] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, and Peng Wu. Balancing unobserved confounding with a few unbiased ratings in debiased recommendations. In *WWW*, pages 1305–1313. ACM, 2023.

[Lin *et al.*, 2019] Kun Lin, Nasim Sonboli, Bamshad Mobasher, and Robin Burke. Crank up the volume: Preference bias amplification in collaborative recommendation. In *RMSE@RecSys*, volume 2440 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019.

[Liu *et al.*, 2020] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. A general knowledge distillation framework for counterfactual

recommendation via uniform data. In *SIGIR*, pages 831–840. ACM, 2020.

[Liu *et al.*, 2021] Dugang Liu, Pengxiang Cheng, Hong Zhu, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. Mitigating confounding bias in recommendation via information bottleneck. In *RecSys*, pages 351–360. ACM, 2021.

[Liu *et al.*, 2023] Dugang Liu, Pengxiang Cheng, Hong Zhu, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. Debiased representation learning in recommendation via information bottleneck. *Trans. Recomm. Syst.*, 1(1):1–27, 2023.

[Ning *et al.*, 2024] Wentao Ning, Reynold Cheng, Xiao Yan, Ben Kao, Nan Huo, Nur Al Hasan Haldar, and Bo Tang. Debiasing recommendation with personal popularity. In *WWW*, pages 3400–3409. ACM, 2024.

[Oestreicher-Singer and Sundararajan, 2012] Gal Oestreicher-Singer and Arun Sundararajan. Recommendation networks and the long tail of electronic commerce. *MIS Q.*, 36(1):65–83, 2012.

[Park *et al.*, 2026] Jongwon Park, Minku Kang, Wooseok Sim, Soyoung Lee, and Hogun Park. Federated recommender system with data valuation for e-commerce platform. *Expert Syst. Appl.*, 298:129695, 2026.

[Romero *et al.*, 2015] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR (Poster)*, 2015.

[Sarwar *et al.*, 2001] Badrul Munir Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In Vincent Y. Shen, Nobuo Saito, Michael R. Lyu, and Mary Ellen Zurko, editors, *Proceedings of the Tenth International World Wide Web Conference, WWW 10, Hong Kong, China, May 1-5, 2001*, pages 285–295. ACM, 2001.

[Schnabel *et al.*, 2016] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1670–1679. JMLR.org, 2016.

[Song *et al.*, 2023] Zijie Song, Jiawei Chen, Sheng Zhou, Qihao Shi, Yan Feng, Chun Chen, and Can Wang. CDR: conservative doubly robust learning for debiased recommendation. In *CIKM*, pages 2321–2330. ACM, 2023.

[Tang and Wang, 2018] Jiaxi Tang and Ke Wang. Ranking distillation: Learning compact ranking models with high performance for recommender system. In Yike Guo and Faisal Farooq, editors, *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 2289–2298. ACM, 2018.

[Tang *et al.*, 2020] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020.

[Tao *et al.*, 2022] Ye Tao, Ying Li, Su Zhang, Zhirong Hou, and Zhonghai Wu. Revisiting graph based social recommendation: A distillation enhanced social graph network. In *WWW*, pages 2830–2838. ACM, 2022.

[Wang *et al.*, 2017] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L_2 hypersphere embedding for face verification. In Qiong Liu, Rainer Lienhart, Haohong Wang, Sheng-Wei "Kuan-Ta" Chen, Susanne Boll, Yi-Ping Phoebe Chen, Gerald Friedland, Jia Li, and Shuicheng Yan, editors, *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1041–1049. ACM, 2017.

[Wang *et al.*, 2021] Shuai Wang, Kun Zhang, Le Wu, Haiping Ma, Richang Hong, and Meng Wang. Privileged graph distillation for cold start recommendation. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1187–1196. ACM, 2021.

[Xu *et al.*, 2020] Chen Xu, Quan Li, Junfeng Ge, Jinyang Gao, Xiaoyong Yang, Changhua Pei, Fei Sun, Jian Wu, Hanxiao Sun, and Wenwu Ou. Privileged features distillation at taobao recommendations. In *KDD*, pages 2590–2598. ACM, 2020.

[Yang *et al.*, 2025] Xinxin Yang, Xinwei Li, Zhen Liu, Yafan Yuan, and Yannan Wang. Multi-teacher knowledge distillation for debiasing recommendation with uniform data. *Expert Syst. Appl.*, 273:126808, 2025.

[Zhang and Shen, 2023] Fan Zhang and Qijie Shen. A model-agnostic popularity debias training framework for click-through rate prediction in recommender system. In *SIGIR*, pages 1760–1764. ACM, 2023.

[Zhang *et al.*, 2019] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.*, 52(1):5:1–5:38, 2019.

[Zhang *et al.*, 2020] Yuan Zhang, Xiaoran Xu, Hanning Zhou, and Yan Zhang. Distilling structured knowledge into embeddings for explainable and accurate recommendation. In *WSDM*, pages 735–743. ACM, 2020.

[Zhu *et al.*, 2020] Jieming Zhu, Jinyang Liu, Weiqi Li, Jincai Lai, Xiuqiang He, Liang Chen, and Zibin Zheng. Ensembled CTR prediction via knowledge distillation. In *CIKM*, pages 2941–2958. ACM, 2020.

[Zhu *et al.*, 2021] Ziwei Zhu, Yun He, Xing Zhao, Yin Zhang, Jianling Wang, and James Caverlee. Popularity-opportunity bias in collaborative filtering. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, pages 85–93. ACM, 2021.