CROSS: Feedback-Oriented Multi-Modal Dynamic Alignment in Recommendation Systems

YANG LI, Institute of Artificial Intelligence, Xiamen University, Xiamen, China
JUNPENG DU, School of Informatics, Xiamen University, Xiamen, China
CHENZHAN WANG, School of Big Data & Software Engineering, Chongqing University, Chongqing, China
ZUNLONG LIU, School of Information Science and Engineering, Shandong Normal University, Jinan, China
XIAOMIN ZHU, Strategic Assessments and Consultation Institute, Academy of Military Science, Beijing, China
CHEN LIN*, School of Informatics, Xiamen University, Xiamen, China and National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China

Aligning the multi-modal content and ID embeddings is crucial in multi-modal recommendation systems. Existing solutions typically adopt a bidirectional alignment paradigm. Our prior work, FETTLE, challenges this paradigm by proposing a one-way directional alignment at the item level, thus reducing the negative impact of low-quality modalities. However, FETTLE leaves two open questions: (1) when is one-way directional alignment optimal, and (2) how to incorporate collaborative signals to enhance alignment? We present CROSS (feedba**C**k-o**R**iented multi-m**O**dal alignment in recommendation **S**y**S**tem), a plug-and-play framework that extends FETTLE by introducing three major advancements. First, we introduce `Dynamic Item-Level Alignment`, which dynamically calibrates the 'strength' of each modality via a variance-based compensation mechanism, mitigating the risk of overshadowing weaker modalities in the early stages of training. Second, we develop `Multi-grained Collaborative Alignment`, which introduces a medium-granularity alignment strategy based on neighboring items that share similar user feedback profiles. This neighbor-level alignment effectively balances noisy user interactions and excessive smoothing across items. Third, we conduct extensive experiments on more real-world datasets and show that CROSS significantly boosts the performance of both collaborative filtering (CF) models and multi-modal recommendation (MRS) approaches, achieving 21.52%–70.78% average improvement on CF backbones and 8.70%–20.73% on MRS backbones. Compared with FETTLE, CROSS achieves additional improvements of 3.82%–5.24%.

Additional Key Words and Phrases: Recommender Systems, Multi-modal Recommendation, Multi-Modality Alignment

*corresponding author

Authors' Contact Information: Yang Li, Institute of Artificial Intelligence, Xiamen University, Xiamen, China; e-mail: goatxy@stu.xmu.edu.cn; Junpeng Du, School of Informatics, Xiamen University, Xiamen, China; e-mail: junpdu@stu.xmu.edu.cn; Chenzhan Wang, School of Big Data & Software Engineering, Chongqing University, Chongqing, Chongqing, China; e-mail: chenzhan_wang@163.com; Zunlong Liu, School of Information Science and Engineering, Shandong Normal University, Jinan, China; e-mail: cpuzzzq@163.com; Xiaomin Zhu, Strategic Assessments and Consultation Institute, Academy of Military Science, Beijing, Beijing, China; e-mail: xmzhu@nudt.edu.cn; Chen Lin, School of Informatics, Xiamen University, Xiamen, Fujian, China and National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, Fujian, China; e-mail: chenlin@xmu.edu.cn.

ACM 2770-6699/2025/5-ART
https://doi.org/10.1145/3734527

1 Introduction

Multi-modal Recommendation Systems (MRSs) have been a research focus in the community of recommendation systems, where multi-modal contents such as product images, textual descriptions, and instructional videos are combined for recommendation [2, 3, 10, 10, 20, 21, 28, 34, 37, 38, 42]. *Multi-modal alignment* is a major issue that affects the performance of MRSs [18]. Because large semantic gaps exist among the embedding spaces of multiple modalities, properly aligning them helps the MRSs to correlate multi-modal content, generate more comprehensive item representations, and make more accurate predictions.

Bidirectional alignment has been widely adopted in multi-modal systems [14–16]. In bidirectional alignment, multi-modal contents are considered parallel and describe the same concepts. Thus, the textual representations are pulled closer to visual representations and vice versa for all samples via contrastive learning. Our previous work FETTLE [18] challenges this conventional bidirectional alignment paradigm and points out that MRSs should allow *one-way directional alignment*. FETTLE argues that multi-modal content may not serve equally for understanding user preferences. For example, a jacket's visual appearance is more important for making consumption decisions than textual descriptions. Thus, the core idea of FETTLE is to **self-adapt item-level alignment direction**. For each item, the "strength" of a modality for this item is estimated, e.g., if the visual embedding of an item can produce high preference scores for users who actually click on this item, then the visual modality is strong for this item. Then, a weaker modality is oriented towards a stronger modality to reduce the adverse effect of low-quality contents or irrelevant modalities.

This paper extends FETTLE and further explores two key questions.

Firstly, we ask *when is one-way directional alignment optimal*? FETTLE adopts one-way directional alignment and completely discards bidirectional alignment. Since modality-specific embeddings are extracted from pre-trained models and fine-tuned in the recommendation task, we monitor their abilities to predict user preferences throughout the training phase. We discover that modality-specific embeddings are insufficiently distinguishing and unstable in predicting user preferences at the early stages of training. This inspires us to investigate **dynamic directional alignment**. On the one hand, bidirectional alignment in early training stages allows each modality to benefit from others and avoids "lazy" modalities or excessive reliance on ID embeddings. On the other hand, one-way directional alignment in late training stages reduces the chance of producing predictions that conflict with actual user preferences.

Secondly, we ask *how to enhance multi-modal alignment with collaborative signals*? FETTLE is a pioneer work in utilizing collaborative signals at different levels: item-level alignment and cluster-level alignment. The item-level alignment maps embeddings from different modalities for one item, while the cluster-level alignment matches all modalities of items in the same cluster with the cluster prototype. The two different levels have their advantages and flaws. The direction of item-level alignment depends on user feedback, and it is less robust due to noisy user feedback. In contrast, cluster-level alignment does not involve user feedback, but it overly homogenizes different items and modalities. This motivates us to introduce **alignment at a medium granularity** to balance the effect of noisy user feedback and aggressive smoothing across items.

Our solutions to the above questions are integrated into CROSS (feedba**C**k-o**R**iented multi-m**O**dal alignment in recommendation **S**y**S**tem), which builds on FETTLE [18]. CROSS consists of three main components. The first component is `Dynamic Item-Level Alignment` (Section 4). `Dynamic Item-Level Alignment` improves the calculation of the "strength" of each modality for each item, i.e., the "strength" of modality is adjusted by the variance of estimated user feedback. Thus, in early training stages, a low estimation score is offset by a large uncertainty, preventing the weaker modality from being overshadowed. The second component is `Multi-Modal Alignment` (Section 5), which refines item representations and resolves conflicting directional signals between modalities. This component is consistent with FETTLE. The third component is `Multi-grained Collaborative Alignment` (Section 6), which implements alignment at different granularities. In addition to the item-level and

cluster-level alignment in FETTLE, CROSS identifies robust neighboring items that share similar user feedback profiles and align the modality-specific embeddings for these neighbors.

CROSS is a plug-and-play framework that works seamlessly with any existing item embeddings derived from a Recommendation System and any pre-trained multi-modal content features. We conduct extensive experiments on four real datasets of textual and visual modalities, including three datasets as in [18] and one new dataset. We apply CROSS to five conventional RSs and six MRSs. CROSS demonstrates significant performance improvements, average 21.52% − 70.78% improvements for conventional RSs and 8.70% − 20.73% for MRSs. Compared with FETTLE, CROSS achieves 3.82% − 5.24% improvements.

To summarize, this paper is a comprehensive extension of the previously proposed, non-invasive, and easily adaptable multi-modal alignment method FETTLE. This paper highlights two directions to improve FETTLE, i.e., adopt `Dynamic Item-Level Alignment` and incorporate `Multi-grained Collaborative Alignment` to further improve FETTLE's performance. The effectiveness of FETTLE and the proposed improvements are verified on more real datasets.

2 Related Works

Multi-modal Alignment. Most state-of-the-art Multi-Modal Models (MMMs) implement multi-modal alignment using contrastive learning techniques [4, 14–17, 39]. For instance, in a bidirectional alignment process, an instance in the visual modality can be treated as the anchor sample, with its corresponding instance in the textual modality serving as the positive sample, while all other instances in the textual modality are treated as negative samples. Conversely, when an instance in the textual modality is used as the anchor, its counterpart in the visual modality is considered the positive sample, and all other visual instances act as negative samples. This bidirectional alignment ensures that each modality is closely associated with its corresponding counterpart while being distinctly separated from unrelated instances.

Multi-modal Recommendation Systems. Existing Multi-modal Recommendation Systems (MRSs)[10, 26, 28, 31, 32, 38, 42] generally adhere to a standardized workflow. In the preprocessing stage, user ID embeddings and item ID embeddings are generated using feed-forward networks (FFNs), while multimodal embeddings are extracted from pre-trained models and aligned through a projector to ensure dimensional consistency. During the learning stage, multimodal embeddings and item ID embeddings are fine-tuned using either an FFN[10] or a graph neural network (GNN) constructed on a user-item bipartite graph [28, 31, 32] or an item-item graph [38, 42]. In the merging stage, multimodal embeddings and item ID embeddings are either concatenated [26] or combined through element-wise addition [26, 28, 31, 32, 38, 42], and jointly optimized with user ID embeddings using the Bayesian Personalized Ranking (BPR) loss [25].

Remarks. To the best of our knowledge, the recent study [43] explicitly incorporates multi-modal correspondence in MRS by minimizing the cosine distance between multi-modal embeddings and ID embeddings. However, this approach differs from our previous work, FETTLE [18]. BM3 adopts a bidirectional alignment strategy, while FETTLE advocates for a one-way directional approach. Building upon FETTLE, we propose CROSS, which introduces `Dynamic Item-Level Alignment` to make the one-way alignment mechanism for multi-modal user preferences more intelligent. Additionally, we introduce `Multi-grained Collaborative Alignment`, which aligns the stable neighbor items to bridge item-level and cluster-level alignment.

3 Framework Overview

We begin by briefly describing the workflow of a Recommendation System (RS). For simplicity, let the RS model accept three primary inputs: a user set $\mathcal{U}$, an item set $\mathcal{I}$, and a binary user feedback matrix $\mathcal{Y}$. Here, $\mathcal{Y}_{u,i} = 1$, for $u \in \mathcal{U}, i \in \mathcal{I}$ indicates that user u has interacted with item i. In the case of a Multi-modal Recommendation System (MRS), the input also includes multi-modal content features extracted by pre-trained models. In this paper,

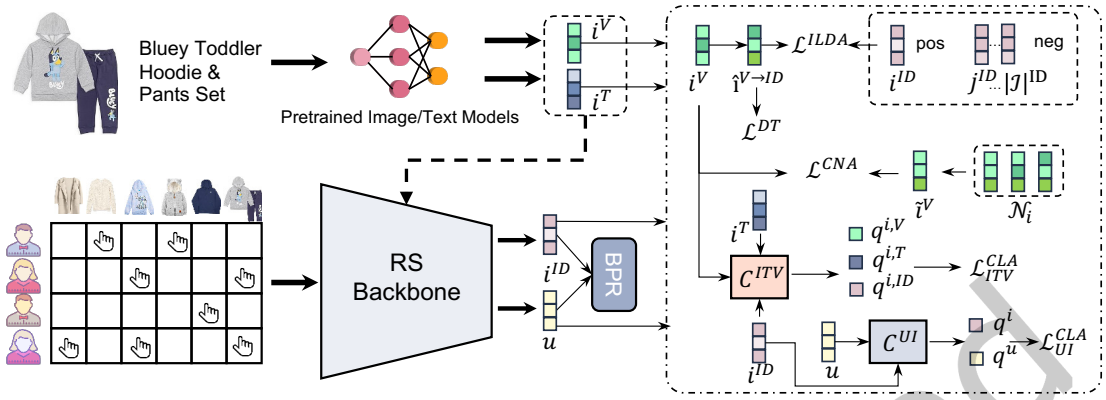


Fig. 1. The overall framework: CROSS works on an arbitrary RS's derived item representation and the pre-trained multi-modal content features; CROSS adds four loss terms to the RS's original BPR loss.

we focus on visual and textual content, specifically image and text embeddings for each item. The RS learns vector representations for both items and users, denoted by $\mathbf{i}^{ID} \in \mathbb{R}^L$ for item i and $\mathbf{u} \in \mathbb{R}^L$ for user u, where L is the embedding dimension. Typically, these user and item representations are optimized via the Bayesian Personalized Ranking (BPR) loss [25].

Similar as FETTLE, CROSS is designed as a plug-and-play framework that can be applied to any existing RS model, regardless of whether it is a multi-modal recommendation system (MRS) or a traditional collaborative filtering (CF) approach. As such, CROSS only operates on the input data (i.e., $\mathbf{i}^V, \mathbf{i}^T, \forall i \in \mathcal{I}, \mathcal{Y}$) and the output data (i.e., $\mathbf{i}^{ID}, \forall i \in \mathcal{I}, \mathbf{u}, \forall u \in \mathcal{U}$), without interfering with the computation of the BPR loss. In this regard, CROSS is positioned as an independent component stacked on top of existing RS models, as illustrated in Figure 1. In general, CROSS is comprised of two main components.

(1) **Pre-processing the input**. Since the dimensionality of image embeddings $\mathbf{i}^V$ and text embeddings $\mathbf{i}^T$ are different. CROSS first applies a projection layer to align them to a uniform embedding size. Concurrently, the existing RS model generates item representations referred to as the ID modality $\mathbf{i}^{ID}$ Thus, each item is associated with three distinct modalities, $\mathcal{M} = \{ID, V, T\}$. The vectors $\mathbf{i}^{ID}, \mathbf{i}^V, \mathbf{i}^T$ and $\mathbf{u}$ are all L-dimensional embeddings.

(2) **Regularizing the BPR loss**. To provide a non-invasive framework, CROSS introduces four regularization terms to the original RS's BPR loss: namely the Item-Level Dynamic Alignment loss (Section 4), the Multi-Modal Alignment loss (Section 5), and the Multi-grained Collaborative Alignment loss (Section 6).

4 Dynamic Item-Level Alignment

4.1 Motivation

Bidirectional alignment has been widely employed in Multi-Modal Models (MMMs)[4, 14, 15, 23], which brings different modalities closer in the embedding space. Suppose we are handling three modalities. The bidirectional alignment essentially assigns each item $i \in \mathcal{I}$ with three pairs of parallel loss terms, i.e., $\mathcal{L}_i^{V \to T}$, $\mathcal{L}_i^{T \to V}$, $\mathcal{L}_i^{V \to ID}$, $\mathcal{L}_i^{ID \to V}$, $\mathcal{L}_i^{T \to ID}$, $\mathcal{L}_i^{ID \to T}$, where V, T, ID are the visual, textual, and ID modalities.

Our previous work FETTLE [18] proposes *one-way directional alignment*, which assigns each item one loss term for a pair of modalities. Specifically, FETTLE computes the "strength" score s_i^m of modality m for item i for each

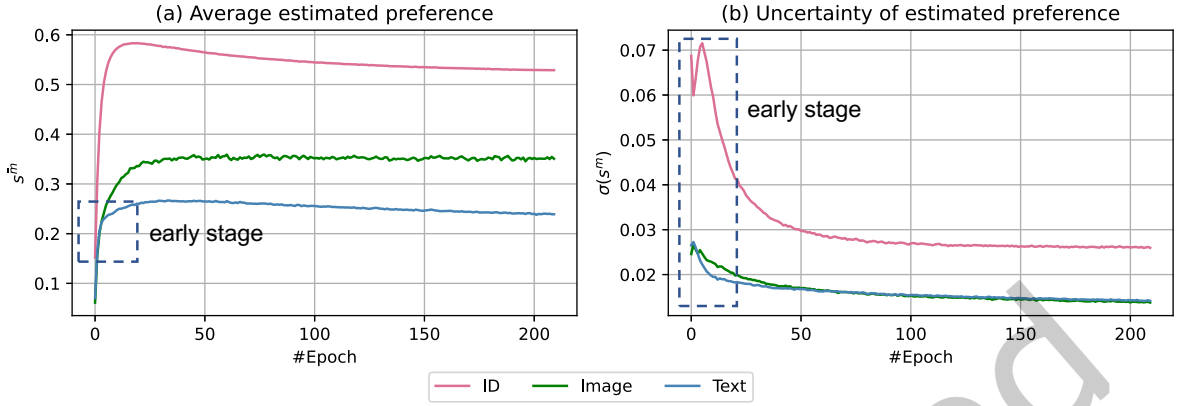


Fig. 2. Change of estimated user preference scores during the training

modality $m \in \mathcal{M}$:

$$s_i^m = \text{avg}_{\mathcal{Y}_{u,i}=1 \in \mathcal{B}}(\mathbf{u}^T \mathbf{i}^m), \tag{1}$$

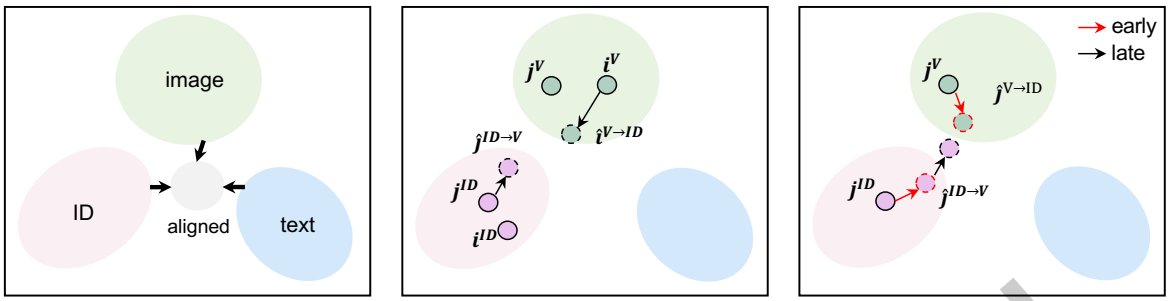
where $\mathbf{u}$ denotes the RS's derived user embedding, and $\mathbf{i}^m$ represents the modality-specific item embedding, $\mathcal{Y}_{u,i} = 1 \in \mathcal{B}$ refers to the interacted users within the training batch for item i.

Thus, s_i^m estimates the average user feedback. For any pair of modalities, e.g., $m, n \in \mathcal{M}$, if $s_i^m \neq s_i^n$, then for item i, the modality m should be aligned with the modality n, denoted as the loss term $\mathcal{L}_i^{m \to n}$. This strategy considers the quality of modality, e.g., which modality has higher quality and better relevance in predicting true user interactions. It avoids performance deterioration that is caused by aligning high-quality modality to low-quality modality.

Since FETTLE determines the alignment direction for each item i by estimating the user feedback score s_i^m, we are curious how s_i^m changes during the entire training phase. We calculate the average estimation across all items $s^{\bar{m}} = \text{avg}_{i \in \mathcal{I}} s_i^m$ and the variance of estimation $\sigma^2(s^m) =$ at each training epoch. As shown in Figure 2(a), in the early training stage, e.g., training epochs 1 to 10, the estimated user preference scores are low for all modalities, and the scores are close to each other. In the late training stage, the ID modality generates the highest estimated user feedback on truly interacted users. As illustrated in Figure 2(b), the variance of estimated user preference scores drastically declines as the training proceeds.

The findings in Figure 2 suggest that we adopt `Dynamic Item-Level Alignment` for several reasons. (1) The user embeddings in early training stages are inaccurate to determine the alignment direction. (2) The modality-specific embeddings are insufficiently fine-tuned in the early training stages. Forcing one-way directional alignment at this point may cause information loss. (3) Image and text modalities have a relative disadvantage compared to the ID modality, which can lead to 'lazy' evolvement, causing the MRS to degrade into a collaborative filtering (CF) model based solely on ID information.

We illustrate the difference among bidirectional alignment, FETTLE's one-way directional alignment, and `Dynamic Item-Level Alignment` in Figure 3. After bidirectional alignment, the three modality-specific embedding spaces, i.e., the embeddings of all items as a whole, are pulled toward a single centroid region (the gray area). Thus, bidirectional alignment compresses the three modalities into a unified space. After FETTLE's unidirectional alignment, the embeddings of various items are dragged along different directions. Thus, each item can maintain the distinctions between different modalities. In `Dynamic Item-Level Alignment`, different modality-specific embeddings are drawn closer and then tugged along a particular direction. Thus, the mutual movement can



(a) Bidirectional (MMM) (b) One-way directional (FETTLE) (c) Dynamic directional (CROSS)

Fig. 3. Illustration of three alignment strategies

speed up multi-modal fusion, and the unidirectional movement can preserve the information richness among items and modalities.

4.2 Implementation

To capture the dynamics of alignment direction, we first define a *compensation score*:

$$g_i^m = \sigma^2(s^m)\left(e^{1-s_i^m} - 1\right), \tag{2}$$

where $\sigma^2(s^m) = E\left[(s_i^m - s^{\bar{m}})^2\right]$ is the variance of s_i^m across all items $i \in \mathcal{I}$, $s^{\bar{m}} = \text{avg}_{i\in\mathcal{I}} s_i^m$ is the average of s_i^m across all items.

We now determine the alignment direction. We separately examine two criteria for any pair of modalities $m, n \in \mathcal{M}$. If $s_i^n < s_i^m$, then the modality n should be aligned with the modality m, denoted as $n \rightarrow m$; if $s_i^n + g_i^n > s_i^m$ for item i, then $m \rightarrow n$.

Here, the compensation score is an offset threshold to trigger bidirectional alignment dynamically. Recall our discussion in Figure 2. In the early training stages, the difference among modalities is small, the value of estimated user feedback s_i^m is small, and the variance σ^2 is large, meaning the strength difference between s_i^m, s_i^n is small, and the compensation score is large. The bidirectional alignment is more likely to appear for i.

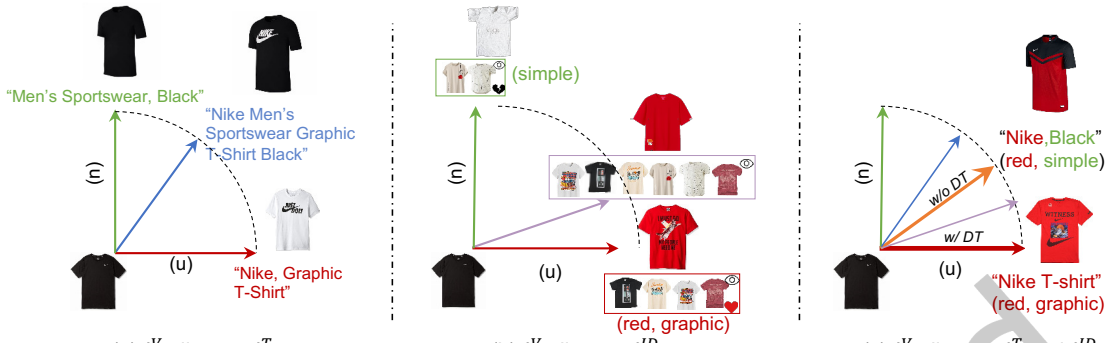
Consequently, we can derive the item-level dynamic directed alignment loss. For item i, given the alignment direction $m \rightarrow n$ determined as above, we aim to maximize the similarity between $\mathbf{i}^m$ and $\mathbf{i}^n$. To capture the specific information that needs to be emphasized in the alignment more effectively, we project $\mathbf{i}^m$ before the alignment using a Feed-Forward Network (FFN) with a residual structure. In contrast to existing approaches, such as ALBEF [16], which utilize only a standard FFN before alignment, the residual structure offers the advantage of explicitly representing the additional information required during alignment from m to n. Formally,

$$\hat{\mathbf{i}}^{m\rightarrow n} = \mathbf{i}^m + f^{m\rightarrow n}(\mathbf{i}^m), \tag{3}$$

where $f^{m\rightarrow n}(\mathbf{i}^m)$ is a FFN for aligning $m \rightarrow n$.

Thus, the item-level dynamic alignment loss is defined as:

$$\mathcal{L}^{ILDA} = -\sum_{i\in\mathcal{B}} \mathbf{avg}_{m\rightarrow n} \log \frac{\exp(\text{sim}(\hat{\mathbf{i}}^{m\rightarrow n}, \text{sg}(\mathbf{i}^n))/\lambda_f)}{\sum_{j\in\mathcal{B}} \exp(\text{sim}(\hat{\mathbf{i}}^{m\rightarrow n}, \text{sg}(\mathbf{j}^n))/\lambda_f)}, \tag{4}$$



(a) i^V aligns to i^T. (b) i^V aligns to i^{ID}. (c) i^V aligns to i^T and i^{ID}.

Fig. 4. Illustration of the inconsistent alignment direction

where sim(,) represents the cosine similarity, sg() is the stop gradient backward operation, λ_f is the temperature parameter, $\mathbf{avg}_{m\to n}$ means the ILDA loss is calculated over all item-level alignment directions $m \to n$, which will be further explained in Section 5.

5 Multi-Modal Alignment

When enumerating item-level alignment directions $m \to n$ in Equation 4 across $|\mathcal{M}| > 2$ modalities, two strategies can be considered. The *Topmost* approach aligns all remaining $|\mathcal{M}|-1$ modalities to the highest-scoring modality for each item. In contrast, the *Pairwise* approach constructs $|\mathcal{M}|(|\mathcal{M}|-1)/2$ modality pairs and aligns the lower-scoring modality to the higher-scoring modality within each pair. It is evident that the first approach may result in laziness in the modality [6]. Specifically, the ID modality, having been sufficiently optimized within the recommendation system (RS), is likely to be the highest-scored modality for most items. As a consequence, other modalities may become dominated by the ID modality, leading to their inactivity and reduced significance in the optimization process.

Based on the rationale presented above, we adopt the pairwise approach. Given three different modalities, there are six possible pairs: $V \to T$, $V \to ID$, $T \to V$, $T \to ID$, $ID \to V$, and $ID \to T$. We can partition the items into different subsets based on these alignment directions. Formally, $\mathcal{D}^{m\to n}$ denotes the set of items that support the alignment direction $m \to n$.

$$\mathcal{D}^{m\to n} = \{i \mid s_i^m < s_i^n\}. \tag{5}$$

Note that only three pairs can be constructed for any item, based on the values of $\mathcal{S}_i$. For example, as shown in Figure 1(c), if $s_i^V < s_i^T < s_i^{ID}$, we have $i \in \mathcal{D}^{V\to T}$, $i \in \mathcal{D}^{V\to ID}$, and $i \in \mathcal{D}^{T\to ID}$.

A limitation of pairwise alignment is that a single low-scoring modality may end up aligning to two different modalities simultaneously. For example, there exists an item $i \in \mathcal{I}$ such that $i \in (\mathcal{D}^{V\to T} \cap \mathcal{D}^{V\to ID})$. We refer to this issue as direction inconsistency, which can confuse the model and hinder the correct alignment.

To address this limitation, we propose an intuitive solution, as illustrated in Figure 4. Once alignment is performed, all modality-specific embeddings should lie within the same vector space spanned by two basis vectors: a user-preference-relevant vector (denoted as (u) in Figure 4) and a user-preference-irrelevant vector (denoted as (n) in Figure 4). Each modality-specific embedding of an item will then be decomposed into these two basis vectors. For instance, in Figure 4(a), the text "Nike Men's Sportswear Graphic T-shirt Black" can be split such that "Nike Graphic T-shirt" is user-preference-relevant, whereas "Men's Sportswear Black" is not. If the visual modality aligns to the textual modality, the resulting image still includes irrelevant attributes like

Fig. 5. Example from the Amazon Baby dataset

"Sportswear Black." Similarly, in Figure 4(b), the ID modality can be decomposed into user-preference-relevant parts such as "red graphic" and irrelevant parts like "white simple." If the visual modality aligns with the ID modality, it may inherit irrelevant elements like "simple."

We resolve this inconsistency by ensuring that alignment directions coincide with the user-preference-relevant component. Concretely, we adjust the alignment direction by maximizing the post-alignment user preference score of each modality embedding. As shown in Figure 4(c), combining two alignment directions might leave certain undesired features (e.g., "black simple") in the image embedding. After direction tuning, the embedding retains only the user-preferred elements, such as "Nike T-shirt red graphic." Formally, we define the direction tuning loss $\mathcal{L}^{DT}$ as

$$\mathcal{L}^{DT} = -\frac{1}{|\mathcal{M}|(|\mathcal{M}|-1)|\mathcal{I}|}\sum_{m,n}\sum_{i\in\mathcal{D}_{m\to n}} \text{avg}_{y_{u,i}=1\in\mathcal{B}}(\mathbf{u}^T\hat{\mathbf{i}}^{m\to n}). \tag{6}$$

6 Multi-grained Collaborative Alignment

The alignments discussed previously rely on user feedback, which includes noisy interactions. We present an illustrative example from the Amazon baby dataset. As shown in Figure 5(a), User 46 interacted with booster seats and rocking chairs, suggesting a preference for furniture [1]. However, the user also interacted with unrelated diaper accessories, possibly due to a misclick. This type of noisy behavior can adversely affect the ILDA strategy by incorporating the irrelevant preferences of User 46 when determining the modality "strength" for Item 7049. In contrast, Figure 5(b) displays the co-occurring neighbors of Item 7049. These neighbors, including night inserts, cloth diapers, and steam sterilization bags, are all essential accessories within the diaper category. Aligning the modalities of Item 7049 with those of its neighbors allows the model to capture the genuine interests of the user community. Moreover, this approach mitigates the impact of noisy interactions from unrelated users, such as User 46, thereby enhancing the robustness of the alignment.

[1]This case is extracted from the baby dataset, so all item categories are related to baby products

6.1 Item Neighbor-Level Alignment

We can directly utilize the concept of neighbor items for alignment. Specifically, we first calculate the similarity of user preferences between items to obtain an item co-occurrence matrix C, using the user feedback matrix $\mathcal{Y}$.

$$C_{i,j} = sim(\mathcal{Y}_{:,i}, \mathcal{Y}_{:,j}), \tag{7}$$

where $\mathcal{Y}u, j$ represents the interaction between item j and user u.

Each item's top-K similar items, as determined by C, are regarded as its neighbors. Formally, we define the neighbor item $k \in \mathcal{N}_i$ of an item i as the set of K items with the highest co-occurrence similarity $C_{i,k}$.

To yield a more robust alignment target from the neighborhood, we apply mean pooling on the neighbor items.

$$\widetilde{\mathbf{i}}^m = \text{meanpooling}(\frac{\mathbf{k}^m}{\|\mathbf{k}^m\|_2^2}, k \in \mathcal{N}_i), \tag{8}$$

where $\mathbf{k}^m$ represents neighbor k's modality-specific embedding on m.

We perform neighbor-level alignment by pulling the item closer to its neighborhood and pushing it away from other items. Accordingly, the neighbor-level alignment loss is formulated as follows:

$$\mathcal{L}^{CNA} = \sum_{i \in \mathcal{B}} \log \frac{\exp(\text{sim}(\mathbf{i}^m, \widetilde{\mathbf{i}}^m)/\lambda_n)}{\sum_{j \in \mathcal{B}} \exp(\text{sim}(\mathbf{i}^m, \widetilde{\mathbf{j}}^m)/\lambda_n)}, \tag{9}$$

where λ_n represents the temperature parameter.

6.2 Item Cluster-Level Alignment

We introduce a cluster-level alignment for items, where different modalities are aligned with the cluster center for a more coarse alignment.

To cluster items efficiently, we draw inspiration from SwAV [1]. Conventional clustering methods like KMeans [8] can be computationally expensive. By contrast, we learn a codebook in a dynamic manner, updating item cluster assignments on the fly without requiring a finalized codebook. This substantially reduces the time complexity typically associated with clustering.

We construct a codebook $\mathbf{C}^{\mathbf{ITV}}$ for different modalities $\{ID, T, V\}$, which records the vectorized representations of typical items in a cluster (i.e., cluster prototypes). Formally, a codebook in the vector space is defined as $\mathbf{C}^{\mathbf{ITV}} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_P\} \in \mathbb{R}^{L \times P}$, where P is the number of prototypes, and $\mathbf{c}_p \in \mathbb{R}^L$ (for $p \leq P$) is the learnable representation of the p-th cluster prototype. The codebook is initialized randomly.

For each item's modality-specific embedding $\mathbf{i}^m$, we can obtain its cluster assignment, referred to as the code $\mathbf{q}^{m,i} \in \mathbb{R}^P$. Ideally, the code is determined by matching the embedding $\mathbf{i}^m$ to the cluster prototypes using the softmax function, i.e., $\text{SoftMax}(\mathbf{i}^m\mathbf{C}^{\mathbf{ITV}}/\tau)$, where τ is a temperature parameter. In the context of recommendation systems, uniformity of item assignments is essential [27]. For instance, popularity bias arises when item distributions are non-uniform, i.e., when most items are assigned to a dominant cluster. To address this, we ensure that the code assignments follow a uniform distribution across clusters. To achieve this, we utilize the Sinkhorn [5] optimal transport algorithm, which generates codes that preserve information integrity while promoting relatively uniform spatial distributions.

$$\mathbf{q}^{m,i} = \lim_{n \to N} \text{diag}(\mathbf{r}^{(n)}) \cdot \text{SoftMax}(\frac{\mathbf{i}^m\mathbf{C}^{\mathbf{ITV}}}{\tau}) \cdot \text{diag}(\mathbf{s}^{(n)}), \tag{10}$$

where $\mathbf{q}^{m,i}, \mathbf{C}^{\mathbf{ITV}}$ are iteratively refined for N times, and $\mathbf{r}^n, \mathbf{s}^n$ represent the renormalization vectors at round n, and $\mathbf{r}^0, \mathbf{s}^0$ are initialized with matrices filled with all ones.

The item cluster-level alignment assumes that the cluster assignments for the different modalities of an item are consistent. To enforce this, we utilize the cross-entropy loss between the code $\mathbf{q}^{m,i}$ and the "ground-truth"

assignment, which is derived from $\text{SoftMax}(\mathbf{i}^m\mathbf{C}^{\text{ITV}}/\tau)$. Formally, we minimize the cluster-level alignment loss for items as follows:

$$\begin{aligned}\mathcal{L}_{ITV}^{CLA} &= -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{m,n\in\mathcal{M},m\neq n}[\mathbf{q}^{n,i}\log\left(\text{SoftMax}(\frac{\mathbf{i}^m\mathbf{C}^{\text{ITV}}}{\lambda_c}\right)) \\ &\quad + \mathbf{q}^{m,i}\log\left(\text{SoftMax}(\frac{\mathbf{i}^n\mathbf{C}^{\text{ITV}}}{\lambda_c}\right))], \\ &= -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{m,n\in\mathcal{M}}[\frac{1}{\lambda_c}(\mathbf{i}^{m,i}\mathbf{C}^{\text{ITV}}\mathbf{q}^{n,i} + \mathbf{i}^n\mathbf{C}^{\text{ITV}}\mathbf{q}^{m,i}) \\ &\quad - \log\sum_{p=1}^{P}\exp(\frac{\mathbf{i}^m\mathbf{c}_p}{\lambda_c}) - \log\sum_{p=1}^{P}\exp(\frac{\mathbf{i}^n\mathbf{c}_p}{\lambda_c})],\end{aligned} \tag{11}$$

where $\mathbf{q}^{m,i}, \mathbf{q}^{n,i}$ are obtained by Equation 10, $\text{SoftMax}(\frac{\mathbf{i}^m\mathbf{C}^{\text{ITV}}}{\lambda_c})$ is the calculated "ground-truth" cluster assignment, the softmax function ensures the calculated assignment is a correct probability distribution, $\frac{\mathbf{i}^m\mathbf{C}^{\text{ITV}}}{\lambda_c}$) computes the similarity between the modality-specific embedding and the codebook, λ_c is the temperature parameter.

The item cluster-level alignment process alternatively update the codebook $\mathbf{C}^{\text{ITV}}$ and the codes $\mathbf{q}^{m,i}, \forall i \in \mathcal{I}, \forall m \in \mathcal{M}$ by optimizing $\mathcal{L}_{ITV}^{CLA}$.

6.3 User Cluster-Level Alignment

. The previous alignments focus solely on items, neglecting user embeddings. However, directly aligning user embeddings with item embeddings presents challenges. User embeddings are dynamic during the training process, and even subtle fluctuations can lead to unstable results. To address this, we propose aligning users with items at the cluster level, ensuring more stability and robustness in the alignment process.

Similarly, we maintain a codebook $\mathbf{C}^{\text{UI}}$ for users and items to represent a set of preference cluster prototypes. For an interacted user-item pair $\mathcal{Y}_{u,i} = 1 \in \mathcal{B}$, the codes $\mathbf{q}^u$ and $\mathbf{q}^i$ should be similar. For instance, if a user prefers "comedy" and "horror" movies, their interacted movies are likely to belong to these genres. Therefore, similar to Equation 11, we define the cluster-level alignment loss for users as follows:

$$\begin{aligned}\mathcal{L}_{UI}^{CLA} &= -\frac{1}{|\mathcal{B}|}\sum_{\mathcal{Y}_{u,i}=1\in\mathcal{B}}[\frac{1}{\lambda_c}(\mathbf{u}\mathbf{C}^{\text{UI}}\mathbf{q}^i + \mathbf{i}^{ID}\mathbf{C}^{\text{UI}}\mathbf{q}^u) - \log\sum_{p=1}^{P}\exp(\frac{\mathbf{u}\mathbf{c}_p}{\lambda_c}) \\ &\quad - \log\sum_{p=1}^{P}\exp(\frac{\mathbf{i}^{ID}\mathbf{c}_p}{\lambda_c})], \\ \mathbf{q}^u &= \lim_{n\to N}\text{diag}(\mathbf{r}^{(n)})\cdot\text{SoftMax}(\frac{\mathbf{u}\mathbf{C}^{\text{UI}}}{\tau})\cdot\text{diag}(\mathbf{s}^{(n)}), \\ \mathbf{q}^i &= \lim_{n\to N}\text{diag}(\mathbf{r}^{(n)})\cdot\text{SoftMax}(\frac{\mathbf{i}^{ID}\mathbf{C}^{\text{UI}}}{\tau})\cdot\text{diag}(\mathbf{s}^{(n)}).\end{aligned} \tag{12}$$

Remarks. Unlike item-level alignment, cluster-level alignment is inherently bidirectional for two main reasons: (1) Cluster-level alignment captures more abstract user, item, and multi-modal features, making it more stable during the learning process. Consequently, updating the cluster-level alignment and the codebook will not degrade the quality of modality-specific embedding vectors. (2) Unlike item-level alignment, which is guided by user feedback, cluster-level alignment aims to capture the high-level categories of items and user communities. Since clustering is inherently reciprocal, it is unnecessary to restrict cluster-level alignment to a one-way direction.

Table 1. Statistics of the datasets.

Datasets	#Users	#Items	#Inter	#Sparsity	#Avg Image Sim	#Avg Text Sim
Baby	19,445	7,050	160,792	99.88%	0.2240	0.2627
Sports	33,598	18,357	296,337	99.95%	0.2085	0.2184
Clothing	39,387	23,033	278,677	99.97%	0.2239	0.3880
TikTok	9,308	6,710	68,722	99.89%	0.8556	0.7113

7 Experiments

We conduct experiments to answer the following questions.

RQ1 Can CROSS improve the performance of existing recommendation methods?
RQ2 What is the impact of each component on CROSS's performance?
RQ3 Is the dynamic directional alignment more suitable for MRSs?
RQ4 How does CROSS perform under scenarios involving missing/noisy multi-modal contents and noisy user feedback?
RQ5 Can CROSS address the issue of modality misalignment?
RQ6 Does Multi-Grained Collaborative Alignment help reduce popularity bias in MRS?
RQ7 How sensitive is CROSS to its hyper-parameters?

7.1 Experimental Setup

Dataset. Following previous multi-modal recommendation systems [38, 42, 43], we conduct experiments on three categories of Amazon review datasets [22]: Baby, Sports, and Clothing. Each item in these datasets is associated with a 4096-dimensional visual feature vector [9] extracted using a pre-trained Convolutional Neural Network, as well as a 384-dimensional textual feature vector obtained from a sentence transformer [24]. The statistics for each dataset are presented in Table 1. To verify the effectiveness of CROSS in various domains, we further conduct experiments on the TikTok dataset following [30].

The TikTok dataset, collected from the streaming media platform TikTok, encompasses visual, textual, and audio modalities, whereas the Amazon dataset is sourced from an e-commerce site. To maintain consistency with our previous experiments, our analysis focuses exclusively on the image and text modalities. We consider that the TikTok dataset is inherently noisier than the Amazon dataset, as bloggers are generally less motivated than merchants to produce high-quality media content. To evaluate this, we calculate the cosine similarity between each pair of items and subsequently compute the average visual and textual similarities. As presented in Table 1, the TikTok dataset exhibits significantly higher inter-item similarities in both visual and textual modalities, with average image and text similarities of 0.8556 and 0.7113, respectively, and 0.2188 and 0.2897 in the Amazon dataset. These findings indicate that the TikTok dataset poses substantial challenges for Multi-modal Recommendation Systems (MRSs), because it is difficult to discriminate user preferences from highly similar multi-modal contents.

Evaluation Protocols. We evaluate following the approach used in many prior works [35, 38, 40, 42, 43]. Specifically, we adopt an 80-10-10 split for training, validation, and testing. Two widely used evaluation metrics are employed: Recall@K (R@K) and NDCG@K (N@K). The reported results are the average values across all users in the test dataset, with $K = \{10, 20\}$.

Implementation. Building on previous work [11, 38, 42, 43], we set the embedding size for both users and items to 64 across all models. The model parameters are initialized using the Xavier method [7], and Adam [13] is used as the optimizer. CROSS is developed on the classical multi-modal recommendation platform, MMRec [41]. The platform includes numerous classic and state-of-the-art MRS backbone architectures. To ensure fairness,

CROSS directly adopts the platform's complete training configurations, including learning rates, embedding dimensions, and hyperparameter settings of all backbone models. A grid search is conducted to identify the optimal hyperparameters of CROSS for different backbone models. Specifically, for CF models, we search the values of β and α within {1, 10, 100}, while for multimodal models, we search within {0.0001, 0.001, 0.01}. For the value of γ, we keep the same value with α. For the regularization parameters λ_c, λ_f and λ_n, we search within {0.1, 0.2, 0.3}, {0.05, 0.1, 0.15} and {0.2, 2, 20}, respectively. The number of prototypes is set to $P = \{10240, 20480\}$ for $\mathbf{C}^{\mathbf{ITV}}$ and $\mathbf{C}^{\mathbf{UI}}$, respectively. The neighbor number is set to $K = \{3, 5, 10\}$. The number of iterations is set to $N = 3$. To accelerate convergence, we employ an early-stopping strategy, using Recall@20 (R@20) on the validation data as the criterion for early stopping. Following prior works [35, 38, 40, 42, 43], the training is performed separately on each dataset.

Backbones. CROSS is a plug-and-play framework that can be seamlessly integrated with various backbone recommendation systems (RS). In our experiments, we evaluate several widely used collaborative filtering models (**CF backbones**), which rely solely on interaction data. These include BPR [25], LightGCN [11], SGL [33], DirectAU [27], and NCL [19]. Additionally, we experiment with multi-modal recommendation models (**MRS backbones**), which incorporate multi-modal content features, such as text and image embeddings from the three datasets. These models include VBPR [10], GRCN [31], DualGNN [28], SLMRec [26], LATTICE [38], and FREEDOM [42].

Competitors. To the best of our knowledge, the most recent work, FETTLE [43], proposes a fine-grained multimodal alignment strategy in multimodal recommendation systems (MRS). In FETTLE, the multimodal alignment direction is based on the modal's user preference score. For each item, the low score modal aligns with the high score modal.

For all the backbone models, we use the open-source implementation available at [2]. Our code is publicly accessible at [3].

7.2 Comparative Study

To address **RQ1**, we integrate CROSS and FETTLE with various backbone recommendation models. Table 2 and 3 present the performance of these backbone models both before and after applying multi-modal alignment with FETTLE and CROSS. The following observations can be made.

(1) CROSS *demonstrates substantial improvements across all backbone models.* Specifically, for collaborative filtering (CF) backbones, CROSS achieves average enhancements of 39.71% in R@10, 36.55% in R@20, 37.32% in N@10, and 36.15% in N@20. For multi-modal recommendation system (MRS) backbones, the average improvements are 14.22%, 12.55%, 15.88%, and 14.71%, respectively. Even for the top-performing MRS backbone, FREEDOM, CROSS delivers average gains of 6.90%, 5.85%, 8.65%, and 7.74%. Notably, the improvement is generally smaller for MRS backbones than for CF backbones, likely because MRS models already integrate multi-modal information to some extent. Nevertheless, CROSS still offers a significant boost of approximately 14.34% on average, indicating that even models leveraging multi-modal content can fail to fully exploit its potential without proper multi-modal alignment. By addressing this gap, CROSS further enhances recommendation performance.

(2) CROSS *demonstrates consistent performance across all datasets.* Specifically, it achieves average improvements of 23.94%, 22.02%, 25.05%, and 23.28% (R@10, R@20, N@10, N@20) on the Baby dataset; 18.17%, 17.13%, 17.64%, and 17.17% on the Sports dataset; and 43.48%, 39.38%, 42.99%, and 42.24% on the Clothing dataset. On the TikTok dataset, CROSS attains average improvements of 17.65%, 15.29%, 16.82%, and 15.15%. Notably, CROSS exhibits the most substantial gains on the Clothing dataset, whereas the improvements on TikTok are comparatively modest. This discrepancy can be attributed to the intrinsic characteristics of the datasets: the Clothing dataset, which focuses on

[2]https://github.com/enoche/MMRec/

[3]https://github.com/XMUDM/FETTLE/tree/main/CROSS/

Table 2. Performance of CF/MRS models before and after applying FETTLE and CROSS in Baby and Sports. The best performance is highlighted in bold. Δ*Imp.* indicates improvements over vanilla models in percentage.

	Models	Baby				Sports			
		R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
CF	BPR	0.0382	0.0595	0.0207	0.0263	0.0417	0.0633	0.0232	0.0288
	+FETTLE	0.0500	0.0790	0.0272	0.0347	0.0579	0.0874	0.0310	0.0385
	+CROSS	**0.0562**	**0.0850**	**0.0307**	**0.0381**	**0.0611**	**0.0942**	**0.0340**	**0.0426**
	LightGCN	0.0465	0.0754	0.0250	0.0325	0.0561	0.0846	0.0308	0.0381
	+FETTLE	0.0576	0.0884	0.0317	0.0395	0.0645	0.0967	0.0351	0.0434
	+CROSS	**0.0592**	**0.0913**	**0.0326**	**0.0408**	**0.0693**	**0.1032**	**0.0384**	**0.0472**
	SGL	0.0532	0.0820	0.0289	0.0363	0.0620	0.0944	0.0339	0.0423
	+FETTLE	0.0585	0.0903	0.0325	0.0407	0.0706	0.1057	0.0386	0.0476
	+CROSS	**0.0594**	**0.0916**	**0.0325**	**0.0407**	**0.0720**	**0.1071**	**0.0396**	**0.0486**
	DirectAU	0.0231	0.0342	0.0128	0.0156	0.0391	0.0570	0.0218	0.0264
	+FETTLE	0.0400	0.0619	0.0215	0.0272	0.0553	0.0828	0.0298	0.0369
	+CROSS	**0.0404**	**0.0625**	**0.0218**	**0.0274**	**0.0565**	**0.0840**	**0.0306**	**0.0377**
	NCL	0.0463	0.0750	0.0249	0.0323	0.0560	0.0842	0.0308	0.0381
	+FETTLE	0.0552	0.0836	0.0298	0.0371	0.0643	0.0966	0.0354	0.0438
	+CROSS	**0.0575**	**0.0907**	**0.0314**	**0.0399**	**0.0655**	**0.0993**	**0.0355**	**0.0442**
	+FETTLE Avg Δ*Imp.*	31.14%	30.52%	31.66%	30.96%	24.79%	24.87%	22.62%	22.97%
	+CROSS Avg Δ*Imp.*	37.03%	35.87%	37.52%	36.34%	29.53%	29.91%	28.73%	29.10%
MRS	VBPR	0.0424	0.0662	0.0223	0.0284	0.0560	0.0857	0.0307	0.0384
	+FETTLE	0.0555	0.0842	0.0297	0.0372	0.0622	**0.0957**	0.0330	0.0417
	+CROSS	**0.0564**	**0.0862**	**0.0308**	**0.0384**	**0.0646**	0.0952	**0.0351**	**0.0430**
	DualGNN	0.0507	0.0808	0.0277	0.0354	0.0589	0.0902	0.0325	0.0405
	+FETTLE	0.0532	0.0830	0.0285	0.0362	0.0624	0.0910	0.0343	0.0417
	+CROSS	**0.0540**	**0.0841**	**0.0297**	**0.0375**	**0.0656**	**0.0938**	**0.0364**	**0.0438**
	GRCN	0.0520	0.0841	0.0284	0.0367	0.0603	0.0911	0.0327	0.0407
	+FETTLE	0.0578	0.0900	0.0311	0.0394	0.0632	0.0964	0.0341	0.0426
	+CROSS	**0.0584**	**0.0913**	**0.0316**	**0.0401**	**0.0642**	**0.0966**	**0.0348**	**0.0431**
	SLMRec	0.0535	0.0820	0.0293	0.0366	0.0660	0.0989	0.0365	0.0449
	+FETTLE	0.0555	0.0840	0.0299	0.0375	0.0681	0.1008	0.0373	0.0457
	+CROSS	**0.0574**	**0.0852**	**0.0319**	**0.039**	**0.0691**	**0.1034**	**0.0378**	**0.0466**
	LATTICE	0.0547	0.0843	0.0291	0.0367	0.0622	0.0953	0.0338	0.0423
	+FETTLE	0.0569	0.0915	0.0310	0.0398	0.0655	0.0986	0.0351	0.0436
	+CROSS	**0.0599**	**0.0927**	**0.0328**	**0.0402**	**0.0672**	**0.1020**	**0.0361**	**0.0451**
	FREEDOM	0.0626	0.0986	0.0327	0.0420	0.0719	0.1076	0.0385	0.0477
	+FETTLE	0.0672	0.1029	0.0355	0.0447	0.0745	0.1115	0.0397	0.0492
	+CROSS	**0.0686**	**0.1047**	**0.0359**	**0.0453**	**0.0764**	**0.1143**	**0.0413**	**0.0510**
	+FETTLE Avg Δ*Imp.*	10.35%	8.85%	10.45%	9.66%	5.66%	4.56%	4.41%	4.04%
	+CROSS Avg Δ*Imp.*	13.04%	10.49%	14.66%	12.39%	8.70%	6.49%	8.40%	7.22%
	v.s. FETTLE	**4.02%**	**3.32%**	**4.96%**	**3.94%**	**3.82%**	**3.34%**	**4.96%**	**4.53%**

Table 3. Performance of CF/MRS models before and after applying FETTLE and CROSS in Clothing and TikTok. The best performance is highlighted in bold. Δ*Imp.* indicates improvements over vanilla models in percentage.

Models		Clothing				Tiktok			
		R@10	R@20	N@10	N@20	R@10	R@20	N@10	N@20
CF	BPR	0.0200	0.0295	0.0111	0.0135	0.0355	0.0538	0.0191	0.0237
	+FETTLE	0.0451	0.0696	0.0248	0.0310	0.0430	0.0637	0.0214	0.0267
	+CROSS	**0.0502**	**0.0738**	**0.0274**	**0.0334**	**0.0473**	**0.0742**	**0.0235**	**0.0303**
	LightGCN	0.0341	0.0527	0.0189	0.0236	0.0584	0.0932	0.0312	0.0399
	+FETTLE	0.0473	0.0698	0.0253	0.031	0.0620	0.0952	0.0340	0.0423
	+CROSS	**0.0499**	**0.0735**	**0.0270**	**0.0330**	**0.0624**	**0.1031**	**0.0345**	**0.0447**
	SGL	0.0332	0.0586	0.0216	0.0266	0.0522	0.0916	0.0251	0.0348
	+FETTLE	0.0516	0.0765	0.0284	0.0347	0.0561	0.0959	0.0263	0.0362
	+CROSS	**0.0527**	**0.0783**	**0.0292**	**0.0357**	**0.0594**	**0.0962**	**0.0286**	**0.0377**
	DirectAU	0.0302	0.0455	0.0165	0.0204	0.0276	0.0515	0.0125	0.0186
	+FETTLE	0.0497	0.0731	0.0266	0.0326	0.0381	**0.0643**	0.0173	0.0239
	+CROSS	**0.0510**	**0.0740**	**0.0274**	**0.0333**	**0.0401**	0.0640	**0.0176**	**0.0236**
	NCL	0.0342	0.0499	0.0183	0.0224	0.0666	0.1028	0.0363	0.0454
	+FETTLE	0.0433	0.0643	0.0234	0.0287	0.0709	**0.1113**	0.0359	0.0461
	+CROSS	**0.0441**	**0.0649**	**0.0239**	**0.0292**	**0.0722**	**0.1113**	**0.0382**	**0.0480**
	+FETTLE Avg Δ*Imp.*	62.16%	57.69%	55.57%	55.87%	15.85%	11.67%	12.62%	10.55%
	+CROSS Avg Δ*Imp.*	70.78%	63.19%	64.31%	63.01%	21.52%	17.22%	18.72%	16.16%
MRS	VBPR	0.0282	0.0420	0.0156	0.0191	0.0286	0.0486	0.0139	0.0189
	+FETTLE	0.0454	0.0675	0.0242	0.0299	0.0410	0.0637	0.0217	0.0274
	+CROSS	**0.0463**	**0.0691**	**0.0248**	**0.0306**	**0.045**	**0.0696**	**0.0232**	**0.0294**
	DualGNN	0.0458	0.0689	0.0243	0.0301	0.0555	0.0867	0.0278	0.0356
	+FETTLE	0.0511	0.0739	0.0278	0.0336	0.0558	0.0854	0.0298	**0.0373**
	+CROSS	**0.0514**	**0.0747**	**0.0281**	**0.034**	**0.0584**	**0.0873**	**0.0299**	0.0371
	GRCN	0.0428	0.0659	0.0225	0.0284	0.0446	0.0719	0.0217	0.0286
	+FETTLE	0.0502	0.0750	0.0266	0.0329	0.0463	0.0749	0.0219	0.0291
	+CROSS	**0.0505**	**0.0761**	**0.0267**	**0.0332**	**0.0492**	**0.0801**	**0.0225**	**0.0302**
	SLMRec	0.0451	0.0670	0.0243	0.0299	0.0686	0.1047	0.0358	0.0448
	+FETTLE	0.0477	0.0697	0.0257	0.0313	0.0709	0.1044	**0.0370**	0.0454
	+CROSS	**0.0494**	**0.0747**	**0.0269**	**0.0334**	**0.0729**	**0.1090**	0.0367	**0.0457**
	LATTICE	0.0486	0.0717	0.0265	0.0324	0.0591	0.0877	0.0315	0.0388
	+FETTLE	0.0531	0.0783	0.0288	0.0351	0.0598	0.0923	0.0306	0.0388
	+CROSS	**0.0556**	**0.0810**	**0.0370**	**0.0458**	**0.0601**	**0.1018**	**0.0318**	**0.0422**
	FREEDOM	0.0627	0.0940	0.0336	0.0415	0.0581	0.0890	0.0316	0.0393
	+FETTLE	0.0658	0.0970	0.0356	0.0435	**0.0620**	0.0936	0.0333	0.0412
	+CROSS	**0.0665**	**0.0981**	**0.0362**	**0.0442**	0.0614	**0.0949**	**0.0347**	**0.0431**
	+FETTLE Avg Δ*Imp.*	18.30%	16.37%	18.02%	16.98%	9.83%	7.31%	11.68%	9.61%
	+CROSS Avg Δ*Imp.*	20.73%	19.54%	25.22%	24.94%	14.42%	13.69%	15.24%	14.30%
	v.s. FETTLE	**5.24%**	**4.23%**	**7.90%**	**7.59%**	**5.08%**	**6.00%**	**4.71%**	**5.11%**

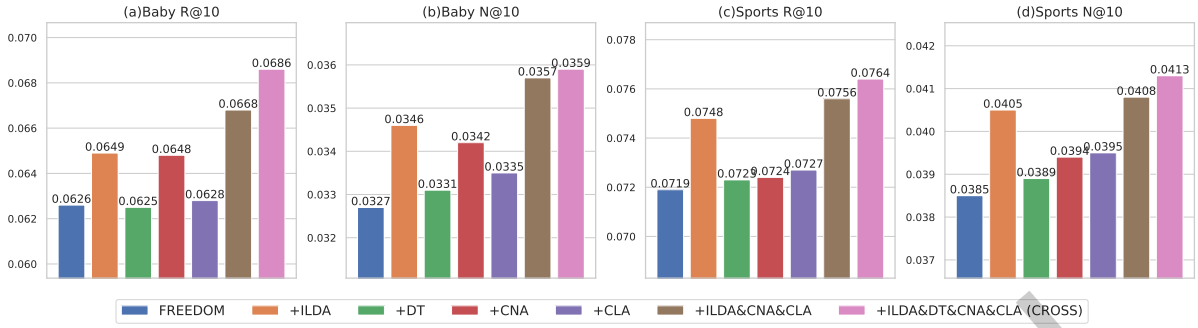


Fig. 6. Performance of each component on Baby and Sports datasets

fashion items (e.g., apparel and accessories), contains cleaner multi-modal data with clearly distinguishable image and text modalities. As both style-driven visuals and detailed textual descriptions play pivotal roles in fashion-related recommendations, aligning these modalities yields substantial benefits. In contrast, TikTok consists predominantly of short-form videos, placing a greater emphasis on the image modality. More importantly, the modals of items in the TikTok dataset exhibit higher noise and smaller inter-item differences, posing significant challenges for MRSs. Consequently, while CROSS excels at leveraging the dual-modal information in the Clothing dataset, it still provides notable improvements on TikTok, despite the lower multi-modal content quality.

(3) CROSS *consistently outperforms FETTLE.* While FETTLE demonstrates commendable performance, CROSS achieves even greater enhancements. Specifically, FETTLE attains average improvements of 21.27% in R@10, 19.23% in R@20, 19.99% in N@10, and 19.17% in N@20. CROSS achieves average improvements of 25.81% in R@10, 23.46% in R@20, 25.63% in N@10, and 24.46% in N@20. Across the majority of backbone models, CROSS surpasses FETTLE in performance metrics. This finding supports our hypothesis that One-way directional multi-modal alignment, which merely aligns low-score modalities to high-score ones, is not always optimal for Multi-modal Recommendation Systems (MRSs). Conversely, CROSS employs a dynamic alignment direction—combining both One-way directional and bidirectional strategies—based on user preferences and a compensation mechanism. This adaptive approach enables CROSS to achieve significant performance improvements, highlighting its superior efficacy in leveraging multi-modal information for enhanced recommendation accuracy. We will deeply discuss the influence of modality alignment ways for MRSs in Section 7.4.

7.3 Ablation Study

To address **RQ2**, we conduct a series of experiments to evaluate the contribution of each component in CROSS. For more robust results, we use the well-performed MRS, FREEDOM [42], as the backbone, progressively incorporate combinations of four components of CROSS, and report the R@10 and N@10 performance on Baby and Sports datasets. The four components include (1) `Dynamic Item-Level Alignment` (ILDA) determines the alignment direction based on user feedback and then adjusts the alignment scores using a compensation mechanism. This mechanism takes into account both the modality's initial score and its associated uncertainty. It utilizes the pairwise $\mathcal{L}^{ILDA}$ described in Section 4. (2) The multi-modal alignment with Direction Tuning loss (DT) addresses the issue of direction inconsistency that occurs when aligning multiple modalities simultaneously, using the $\mathcal{L}^{DT}$ described in Section 5. (3) `Multi-grained Collaborative Alignment` (CNA) identifies stable neighboring items that share similar user feedback profiles with the target item. By aligning modalities among these neighbors, CNA helps filter out noise from random interactions, as spurious or erroneous feedback usually lacks consistent neighbor-based support. This is achieved through the $\mathcal{L}^{CNA}$ described in Section 6. (4) Cluster-level Alignment

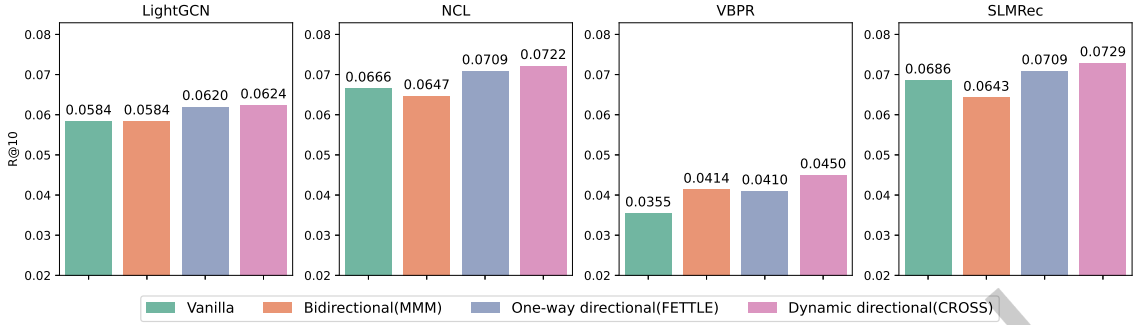


Fig. 7. Performance in different modality alignment direction approach on TikTok dataset

(CLA) aligns users and items, as well as the different modalities of items, at the cluster level using the $\mathcal{L}^{CLA}$ described in Section 6.

(1) *ILDA stands out as a particularly effective component of* CROSS. When applied to the backbone system FREEDOM in isolation, the ILDA strategy ("+ILDA") achieves a notable enhancement: it improves R@10 and N@10 on the Baby dataset by 3.67% and 5.81%, respectively, as shown in Figures 6(a) and 6(b), and on Sports dataset by 4.03% and 5.19%, in Figures 6(c) and 6(d). These gains are especially significant given that FREEDOM itself is already a sophisticated MRS, and making further performance boosts on FREEDOM is inherently challenging. The results confirm the importance of orienting item-level alignment based on user feedback. Moreover, a comparison with "+CLA" and "+CNA" reveals that "+ILDA" yields higher average improvements, underscoring the superior efficacy of ILDA over CLA or CNA when used alone.

(2) *Both CNA and CLA contribute to improving RS performance.* Compared with FREEDOM, "+CNA" achieves notable enhancements: R@10 and N@10 on Baby dataset, improve by 3.51% and 4.59%, respectively, in Figures 6(a) and 6(b), and on Sports dataset by 0.70% and 2.34% in Figures 6(c) and 6(d). Similarly, "+CLA" delivers improvements, with R@10 and N@10 on Baby dataset increasing by 0.32% and 2.45% in Figures 6(a) and 6(b), and on Sports dataset by 1.11% and 2.60% in Figures 6(c) and 6(d).

When combined with ILDA, both strategies significantly enhance RS performance. Specifically, "+ILDA&CNA&CLA" achieves substantial improvements: R@10 and N@10 on Baby dataset increase by 5.11% and 7.65%, respectively, in Figures 6(a) and 6(b), and on Sports dataset by 5.29% and 7.01% in Figures 6(c) and 6(d). These results indicate that CNA and CLA effectively mitigate the impact of noisy or inaccurate feedback in ILDA, refining the alignment process and achieving a denoising effect.

(3) *DT can address the directional inconsistency problem in ILDA.* As shown in Figure 6, applying "+DT" alone has little to no effect on RS performance. In fact, on the Baby dataset, R@10 even decreases slightly by -0.16%, as illustrated in Figure 6(a). However, when combined with "+ILDA," DT significantly improves RS performance. We present the performance of "+ILDA&CNA&CLA" without DT and with DT("+ILDA&DT&CNA&CLA") on the Baby and Sports datasets, as shown in Figure 6. The inclusion of DT yields notable improvements in R@10 and N@10: specifically, increases of 4.26% and 1.99% on the Baby dataset, and 0.92% and 0.24% on the Sports dataset. These results demonstrate that DT further enhances multi-modal alignment at item-level, neighbor-level, and cluster-level, effectively resolving alignment inconsistencies and improving overall RS performance.

7.4 Alignment Direction

To address **RQ3**, we conducted experiments on the TikTok dataset, which represents a more complex and realistic scenario, to compare three alignment strategies: Bidirectional, One-way directional, and Dynamic directional in

the context of Multi-modal Recommendation Systems (MRSs). Specifically, we selected two representative models each from Collaborative Filtering (CF) and MRS: LightGCN (a classic CF model) and NCL (the best-performing CF model on TikTok) for CF, and VBPR (a classic MRS model) and SLMRec (the best-performing MRS model on TikTok) for MRS. The results are shown in Figure 7, leading to the following observations:

(1) Dynamic directional alignment achieves the best overall performance. Across the four backbones, hybrid-directional alignment increases the average R@10 from 0.0573 to 0.0631, an improvement of 12.07%, outperforming the One-way directional alignment (7.87%) and bidirectional alignment (1.87%). This indicates that hybrid-directional alignment is particularly suitable for MRSs.

(2) Bidirectional alignment exhibits the weakest performance. Its average R@10 improvement across the four backbones is only 1.87%. Notably, on NCL and SLMRec, R@10 even decreases by 2.85% and 6.27%, respectively, suggesting that bidirectional alignment is largely unsuitable for MRSs.

(3) One-way directional alignment performs moderately well in most scenarios. compared with the vanilla backbones, it improves R@10 by 6.16% in LightGCN, 6.46% in NCL, and 3.35% in SLMRec. However, it underperforms in VBPR, achieving 1.13% less improvement than Bidirectional alignment. A plausible explanation is that VBPR's weaker representational capacity leads to lower-quality user embeddings, making it harder to estimate accurate modality scores and thus undermining the benefits of one-way directional alignment. Additionally, VBPR is more susceptible to noisy user feedback.

In summary, hybrid-directional alignment, which combines the ideas of both Bidirectional and One-way directional strategies, offers greater robustness and adaptability across various backbones, making it a more reliable choice for complex multi-modal recommendation scenarios.

7.5 Nosiy Content and Noisy User Feedback

To address **RQ4**, we examine the performance of the leading MRS backbone FREEDOM, FETTLE, and CROSS, under scenarios with missing modalities and noisy user feedback.

Modality Missing. To evaluate the performance of MRSs under scenarios with noisy/missing multi-modal content, we randomly remove a portion of the multi-modal data in the Amazon Baby dataset and assess the recommendation performance. Specifically, for each item's image and textual descriptions, we replace the original features with Gaussian noise at rates of 20%, 40%, 60%, and 80%. As discussed in [36], the Gaussian noise simulates missing modalities. Note that an item may lose both its image and text modalities simultaneously or only one modality at a time. With this setup, we construct a series of training sets featuring varying degrees of missing content and then train our recommendation system on these datasets to evaluate their robustness.

The approach(FETTLE's ILA and Ours ILDA) leverages user preferences to determine the direction of modality alignment. To evaluate its effectiveness, we conduct experiments on the Baby dataset. We compare the performance of the backbone model, FREEDOM, with its variant enhanced by FETTLE's ILA strategy and our proposed ILDA strategy. The evaluation metric used is R@10. As shown in Figure 8(a), two key observations can be made: (1)*Both ILA and* ILDA *enhance FREEDOM under various missing ratios.* FREEDOM achieves an average R@10 of only 0.0559 across different missing rates, whereas ILA and ILDA reach 0.0570 and 0.0576, respectively. This highlights that one-way directional alignment can effectively mitigate the impact of missing modalities. When a modality is replaced with noise, it naturally has a lower preference score. It is aligned toward the remaining "strong" modality, thus reducing the negative influence of corrupted content on the overall MRS. (2)ILDA *is more flexible and effective compared with ILA.* ILA employs a relatively rigid "low-score aligns to high-score" policy, leading to potential long-term disadvantages or even "laziness" for modalities with slightly lower scores. In contrast, ILDAincorporates a dynamic score compensation mechanism, alleviating this limitation and boosting the average R@10 from 0.0570 to 0.0576. This improvement demonstrates stronger robustness and adaptability in handling missing-modal scenarios.

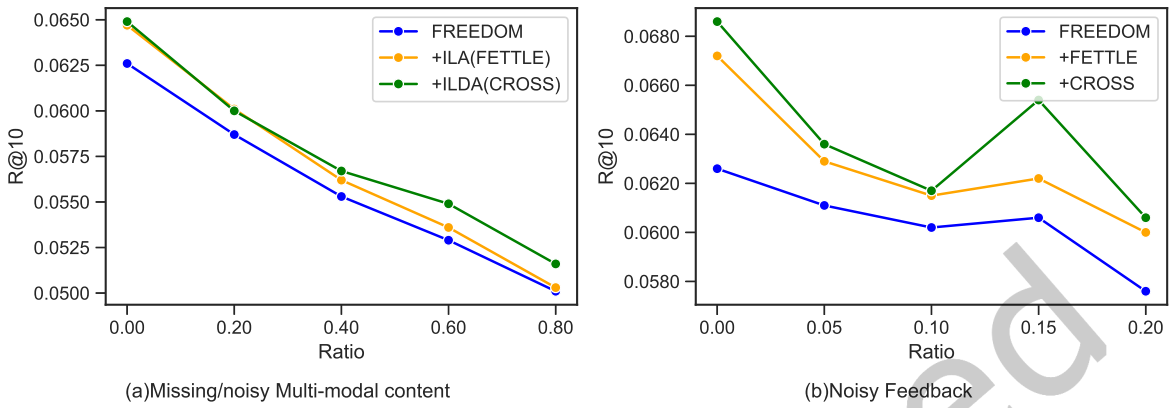


Fig. 8. Performance in scenarios with noisy/missing multi-modal content and noisy user feedback on Baby dataset

Noisy User Feedback. To evaluate performance under noisy user feedback, we artificially inserted 5%, 10%, 15%, and 20% of user-item interactions that did not actually occur into the training set of the Amazon Baby dataset, simulating user misclicks. We then compared the baseline model FREEDOM, FETTLE, and CROSS on different noise levels in terms of average R@10.

The results are shown in Figure 8(b), we draw two key conclusions:(1)*FETTLE and* CROSS *exhibit stronger robustness against noisy interactions.* Compared with FREEDOM, both FETTLE and CROSS better handle misclick noise, boosting the average R@10 to 0.0628 and 0.0640, respectively. It shows cluster-level alignment (CLA) can effectively assist the influence cause by the noisy user feedback. (2)*Neighbor-based modality alignment effectively filters out noise.* Compared with FETTLE, CROSS achieves 1.85% improvement in R@10. It shows that neighbor-level alignment strategies leverage stable feedback from similar items, reducing the impact of isolated misclicks and producing more accurate, robust user and item representations.

7.6 Visualization

To determine whether CROSS effectively alleviates the spatial misalignment among ID, image, and text embeddings in a multi-modal recommendation system(**RQ5**), we compare FREEDOM (baseline) and FREEDOM + CROSS on Baby dataset. We extract the ID, image, and text embeddings and apply T-SNE for dimensionality reduction. As illustrated in Figure 9(a), without CROSS, the embedding space exhibits clear segregation: ID and image embeddings occupy largely disjoint regions, and text embeddings also remain far from the other two. Once CROSS is integrated, we observe a pronounced overlap among all three modalities in the 2D space. Quantitatively, CROSS decreases the average Euclidean distance among modality embeddings by approximately 56.97%, compared with the baseline. These findings underscore CROSS's efficacy in harmonizing embeddings from distinct modalities, ensuring a more unified representation of the same item.

We further investigate whether improved multi-modal alignment yields tangible benefits for recommendation. Again, we conduct experiments on the Baby dataset with both FREEDOM and FREEDOM + CROSS, this time focusing on 500 user-item pairs with actual interaction records. In the T-SNE visualization shown in Figure 9(b), user embeddings (shown in blue) and item embeddings (red for text, green for images) are largely scattered in separate regions without CROSS. This distribution indicates a limited capacity to align users and their interacted items in a shared embedding space. In contrast, with CROSS enabled, all three embedding types exhibit extensive

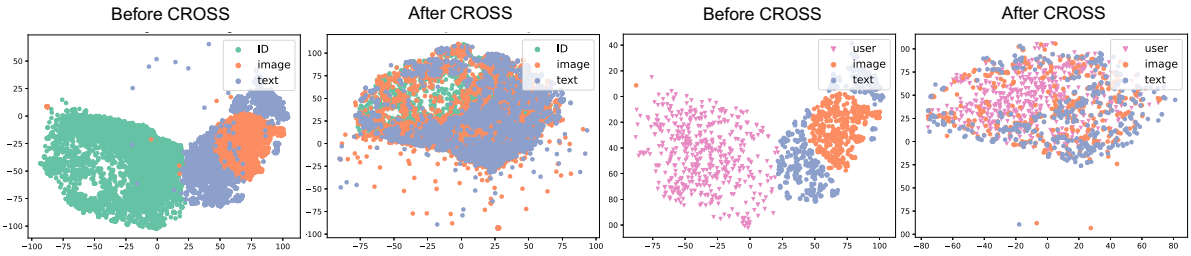


Fig. 9. (a) Visualization of Items ID, image, and text embeddings on Baby dataset. (b) Visualization of image and text embeddings of interacted users and items on Baby dataset.

overlap. A quantitative analysis reveals that the average cosine similarities between user embeddings and their interacted item's image embeddings and text embeddings increases by around 14.69% and 12.18% when CROSS is applied. Consequently, CROSS not only rectifies inter-modality mismatches at the item level but also enhances the model's ability to capture authentic user preferences, thereby improving overall recommendation accuracy and robustness.

7.7 Popularity Bias Mitigation

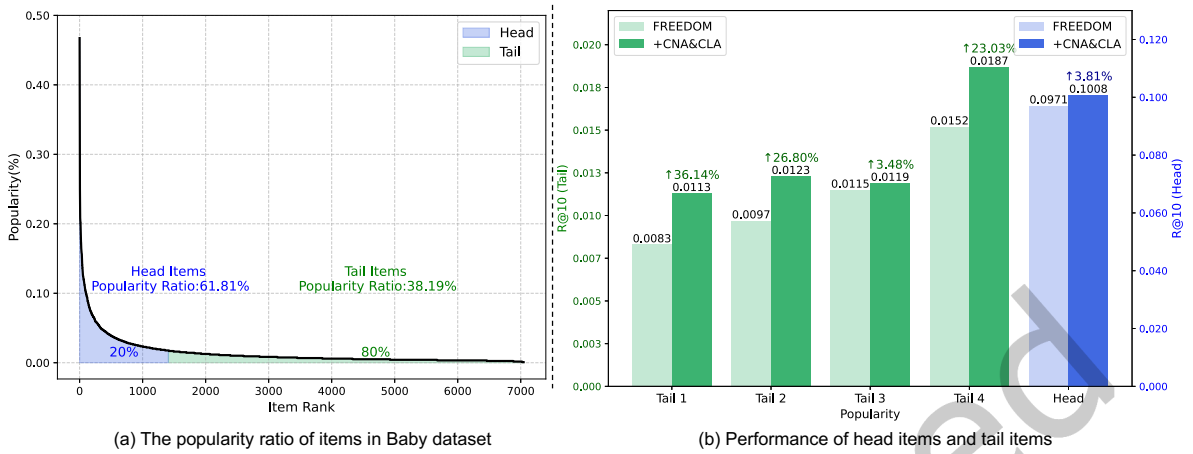
To validate that Multi-Grained Collaborative Alignment in CROSS can alleviate the popularity bias issue in recommendation systems (*RQ6*), we conducted experiments on Amazon Baby dataset, which exhibits a significant Matthew effect. As shown in Figure 10(a), following prior works [12, 29], we categorized items based on the number of interactions with users. The top 20% of items with the highest number of interactions were classified as head items (Head), while the remaining 80% were classified as tail items (Tail). We observed a power-law distribution in this dataset, with head items accounting for the majority of interactions (61.81%) and tail items only accounting for 38.19%. This long-tail phenomenon [12] is a key factor contributing to popularity bias.

To intuitively demonstrate the performance improvement of Multi-Grained Collaborative Alignment for tail items with different popularity levels, we further divided the tail items into four groups (Tail 1, Tail 2, Tail 3, and Tail 4) based on the number of interactions with users, from low to high. We then evaluated the recommendation performance for each group of items, using two experimental settings: (1) the original backbone (FREEDOM) and (2) the backbone with Multi-Grained Collaborative Alignment (+CNA&CLA). The experimental results, shown in Figure 10(b), lead to the following observations:

(1) Multi-Grained Collaborative Alignment improves all popularity levels. For Tail 1, Tail 2, Tail 3, Tail 4, and Head, the recommendation performance improved by at least 3.48%, with a maximum improvement of 36.14%. This indicates that Multi-Grained Collaborative Alignment provides positive gains for all item groups, enhancing the overall performance of the recommendation system.

(2) Multi-Grained Collaborative Alignment significantly improves performance for tail items. For head items, R@10 improved by 3.81%. For tail items, R@10 improved on average by 22.36%, indicating that Multi-Grained Collaborative Alignment effectively alleviates the influence of popularity bias in the recommendation system and enhances the recommendation ability for tail items.

(3) Multi-Grained Collaborative Alignment has the most significant improvement for cold-start items. Among the four popularity levels, Tail 1 consists of items with the fewest interactions with users, and their recommendation performance reflects the system's ability to recommend cold-start items. Due to the lack of user interactions, these



(a) The popularity ratio of items in Baby dataset

(b) Performance of head items and tail items

Fig. 10. (a) Distribution of item popularity and its ratio in the Baby dataset, (b) Comparison of R@10 (Tail and Head) performance for FREEDOM and +CNA&CLA(Multi-Grained Collaborative Alignment) across popularity levels

items often struggle to be recommended. However, we found that Multi-Grained Collaborative Alignment had the most significant improvement for Tail 1 items, with R@10 increasing from 0.0083 to 0.0113, a 36.14% improvement. This highlights the effectiveness of the strategy: *Firstly, Item Neighbor-Level Alignment establishes connections between niche items and other items in their neighborhood. Secondly, Item Cluster-Level Alignment, based on Sinkhorn optimal transport theory, normalizes the cluster distribution, preventing the recommendation system from overly focusing on clusters containing popular items. Finally, user clustering alignment helps the recommendation system learn the user's abstract interest distribution (e.g., 'baby furniture' rather than specific products). If the cold-start item aligns with the user's abstract interests, it could be recommended.*

7.8 Impact of Hyper-parameters

To examine the impact of hyperparameters (**RQ7**), we implement FREEDOM+CROSS on the Baby dataset with various hyperparameter configurations. We focus on three key sets of hyperparameters: the loss weights, α and β and γ, which balance the alignment losses, and the temperatures, λ_f and λ_n, which control the attention given to challenging alignment samples. Finally, we also conducted hyperparameter experiments on the K parameter used in the Top-K method for finding the most similar neighbors in the Collaborative-Neighbour Alignment section. Specifically, we explore different values for the loss weights α, β and γ from the set $\{1e{-}5, 1e{-}4, 1e{-}3, 1e{-}2, 1e{-}1\}$, as well as for the temperatures, varying λ_f within $\{0.05, 0.1, 0.15, 0.2, 0.25\}$ and λ_c within $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ and K within $\{1, 3, 5, 10, 15\}$. The R@10 results are presented in Figure 11. We make the following observations.

(1) CROSS *exhibits robustness to loss weights.* It is shown in Figure11. For values of α ranging from $1e-5$ to $1e-1$, the performance of CROSS in R@10 varies from 0.0639 to 0.0686. For the values of β from $1e-5$ to $1e-2$, the performance of CROSS in R@10 varies from 0.0657 to 0.0686. For the values of γ from $1e-5$ to $1e-2$, the performance of CROSS in R@10 varies from 0.0663 to 0.0686. Given that the performance of FREEDOM is 0.0626 in R@10, CROSS consistently improves the performance of the backbone model, FREEDOM, across different settings of β, α, and γ. There are only two exceptions occurring in extreme configurations, such as $\beta = 1e-1$ and $\gamma = 1e-1$, where CROSS's performance is impacted. Since γ controls item neighbor-level alignment and β controls cluster-level alignment, high values of β and γ reduce the influence of item-level dynamic directional

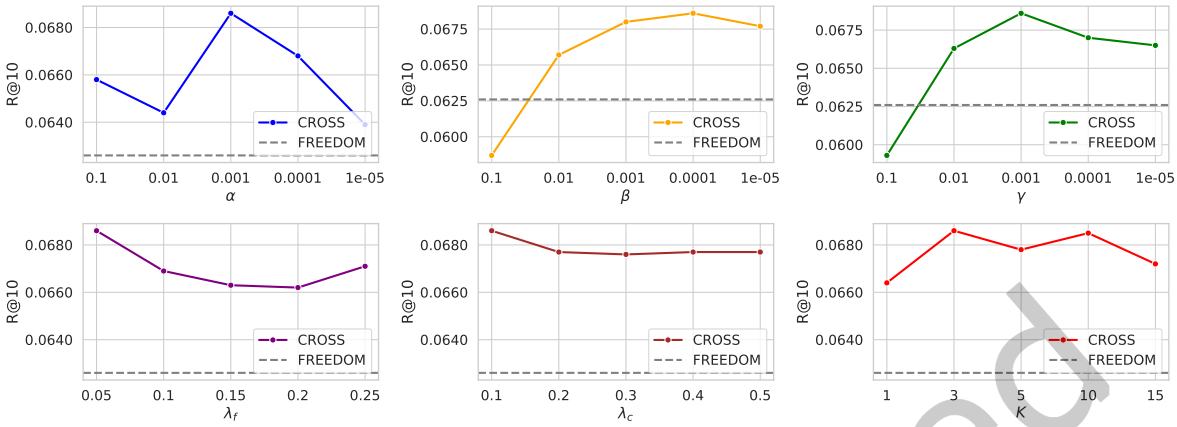


Fig. 11. Performance of CROSS under different hyper-parameters

alignment. This observation further highlights the importance of adaptive alignment at the item level. The optimal performance is achieved with $\alpha = 1e-3$, $\beta = 1e-4$, and $\gamma = 1e-3$.

(2) CROSS *is robust to temperature coefficients.* As demonstrated in Figure 11, CROSS shows insensitivity to the temperature coefficients λ_f and λ_c. For values of λ_f ranging from 0.05 to 0.25, the performance of CROSS in R@10 varies from 0.0662 to 0.0686. For the values of λ_c from 0.1 to 0.5, the performance of CROSS in R@10 varies from 0.0676 to 0.0686. Among all the tested temperature coefficient selections, CROSS brings at least a 5.75% higher recommendation performance compared with FREEDOM in R@10. The best performance is achieved with moderate values of $\lambda_f = 0.05$ and $\lambda_c = 0.1$.

(3) CROSS *is robust to Top-K parameter K.* As illustrated in Figure 11, under all tested values of K, the performance of CROSS changes from 0.0664 to 0.0686 in R@10, demonstrating a significant performance advantage over FREEDOM, with at least a 6.07% improvement. This observation also confirms the rationality and robustness of our Collaborative-Neighbour Alignment approach.

8 Conclusion

Aligning multi-modal content and ID embeddings is a pivotal challenge in multi-modal recommendation systems. Traditional solutions predominantly rely on bidirectional alignment paradigms. In contrast, our prior work, FETTLE, proposed a unidirectional item-level alignment, mitigating the adverse effects of low-quality modalities. This paper introduces CROSS (feedba**C**k-o**R**iented multi-m**O**dal alignment in recommendation **S**y**S**tem), a versatile plug-and-play framework that extends FETTLE and combines three components.

First, we propose `Dynamic Item-Level Alignment`, a variance-based compensation mechanism that dynamically calibrates the contribution of each modality, preventing dominant modalities from overshadowing weaker ones during early training stages. Second, we refine the item-level representations via `Multi-Modal Alignment` to resolve potentially conflicting directional signals among modalities and ensure consistent user-preference modeling. Third, we develop `Multi-grained Collaborative Alignment`, a medium-granularity alignment approach leveraging robust neighbor-level alignment, which incorporates items with similar user feedback profiles. This method balances the trade-off between noisy user interactions and over-smoothing across items.

This paper presents a comprehensive extension to FETTLE, achieving an additional performance gain of 3.82%–5.24% on four real datasets. There are several questions left unanswered. (1) The multi-modal alignment is

conducted on items, which is natural. How is the alignment problem interpreted from the user's perspective? (2) The multi-modal embeddings are extracted from pre-trained models, usually uni-modal and small-sized models. Can they be integrated with SOTA multi-modal language models? We will explore these directions in our future work.

9 Acknowledgments

Chen Lin is the corresponding author. Chen Lin is supported by the National Key R&D Program of China (No. 2022ZD0160501), the Natural Science Foundation of China (No.62372390,62432011).

References

[1] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems* 33 (2020), 9912–9924.

[2] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 335–344.

[3] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 765–774.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.

[5] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* 26 (2013).

[6] Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. 2023. On Uni-Modal Feature Learning in Supervised Multi-Modal Learning. In *Proceedings of the 40th International Conference on Machine Learning*. 8632–8656.

[7] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 249–256.

[8] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.

[9] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.

[10] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[11] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

[12] Michael A Hitt. 2007. The long tail: Why the future of business is selling less of more.

[13] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1412.6980

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.

[16] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[17] X Li, X Yin, and Oscar LI C. 2020. Object-Semantics Aligned Pre-training for Vision-Language Tasks [C]. In *European Conference on Computer Vision. Springer, Cham*. 121–137.

[18] Yang Li, Qi'Ao Zhao, Chen Lin, Jinsong Su, and Zhilin Zhang. 2024. Who To Align With: Feedback-Oriented Multi-Modal Alignment in Recommendation Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 667–676. doi:10.1145/3626772.3657701

[19] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. 2022. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In *Proceedings of the ACM Web Conference 2022*. 2320–2329.

[20] Fan Liu, Zhiyong Cheng, Changchang Sun, Yinglong Wang, Liqiang Nie, and Mohan Kankanhalli. 2019. User diverse preference modeling by multimodal attentive metric learning. In *Proceedings of the 27th ACM international conference on multimedia*. 1526–1534.
[21] Qiang Liu, Shu Wu, and Liang Wang. 2017. Deepstyle: Learning user preferences for visual recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*. 841–844.
[22] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.
[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
[24] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3982–3992.
[25] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*. 452–461.
[26] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia* (2022).
[27] Chenyang Wang, Yuanqing Yu, Weizhi Ma, Min Zhang, Chong Chen, Yiqun Liu, and Shaoping Ma. 2022. Towards representation alignment and uniformity in collaborative filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1816–1825.
[28] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia* (2021).
[29] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 1791–1800.
[30] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023*. 790–800.
[31] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*. 3541–3549.
[32] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1437–1445.
[33] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 726–735.
[34] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig MacDonald. 2022. Multi-modal Graph Contrastive Learning for Micro-video Recommendation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*. ACM, 1807–1811.
[35] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. 2023. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM international conference on multimedia*. 6576–6585.
[36] Jinghao Zhang, Guofan Liu, Qiang Liu, Shu Wu, and Liang Wang. 2024. Modality-Balanced Learning for Multimedia Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7551–7560.
[37] Jinghao Zhang, Qiang Liu, Shu Wu, and Liang Wang. 2023. Mining Stable Preferences: Adaptive Modality Decorrelation for Multimedia Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*. ACM, 443–452.
[38] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3872–3880.
[39] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Making Visual Representations Matter in Vision-Language Models. *CVPR 2021* (2021).
[40] Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. 2023. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. In *ECAI 2023*. IOS Press, 3123–3130.
[41] Xin Zhou. 2023. MMRec: Simplifying Multimodal Recommendation. *arXiv preprint arXiv:2302.03497* (2023).
[42] Xin Zhou and Zhiqi Shen. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 935–943.
[43] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*. 845–854.

Received 30 December 2024; revised 24 February 2025; accepted 22 April 2025