Who To Align With: Feedback-Oriented Multi-Modal Alignment in Recommendation Systems

Yang Li
goatxy@stu.xmu.edu.cn
Institute of Artificial Intelligence,
Xiamen University
Xiamen, China

Qi'Ao Zhao
Chen Lin*
Jinsong Su
chenlin@xmu.edu.cn
School of Informatics,
Xiamen University
Xiamen, China

Zhilin Zhang†
zzhilin@amazon.com
Amazon
Seattle, United States

ABSTRACT

Multi-modal Recommendation Systems (MRSs) utilize diverse modalities, such as image and text, to enrich item representations and enhance recommendation accuracy. Current MRSs overlook the large misalignment between multi-modal content features and ID embeddings. While bidirectional alignment between visual and textual modalities has been extensively studied in large multi-modal models, this study suggests that multi-modal alignment in MRSs should be in a one-way direction. A plug-and-play framework is presented, called **FE**edback-orien**T**ed mul**T**i-modal a**L**ignm**E**nt (FETTLE). FETTLE contains three novel solutions: (1) it automatically determines item-level alignment direction between each pair of modalities based on estimated user feedback; (2) it coordinates the alignment directions among multiple modalities; (3) it implements cluster-level alignment from both user and item perspectives for more stable alignments. Extensive experiments on three real datasets demonstrate that FETTLE significantly improves various backbone models. Conventional collaborative filtering models are improved by 24.79% – 62.79%, and recent MRSs are improved by 5.91% – 20.11%.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

multi-modal recommendation, alignment,recommender systems

ACM Reference Format:
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 (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626772.3657701

*Corresponding author supported by the Natural Science Foundation of China (No.62372390)

†The publication does not relate to the author's work at Amazon.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657701

1 INTRODUCTION

We have witnessed an increasing availability of multi-modal content, such as product images, textual descriptions, instruction videos, etc. Users can understand products of interest more comprehensively with the help of multi-modal content. Thus, Multi-modal Recommendation Systems (MRSs) that utilize multi-modal contents have emerged on various online platforms [10, 26, 31–33, 36].

For efficiency and effectiveness, MRSs generally do not directly operate on multi-modal contents. Instead, they take multi-modal content features as input, e.g., text and image embedding encoded by pre-trained language and visual models. Then, they extract ID embeddings from the user-item interaction matrix and merge the multi-modal content features and ID embeddings using different fusion [2, 3, 10, 18, 19] or graph-based methods [26, 27, 33, 36] to derive the final item representations.

One critical issue of previous MRSs is the large misalignment between the multi-modal content features and the ID embeddings. We next illustrate how the misalignment harms the recommendation performance using a state-of-the-art (SOTA) MRS (i.e., FREEDOM [36]) on the Amazon Baby dataset [20]. We project FREEDOM's ID, image, and text embeddings before the merge phase to the 2D space. As shown in Figure 1(a)-(d), three modalities lie in different regions (i.e., misalignment). Because FREEDOM derives the final item representation without aligning the three modalities, it produces false positives in determining similar items, which leads to inaccurate recommendations. For example, items 1049 and 1902 are dissimilar in each modality, but FREEDOM erroneously decides they are similar based on its derived item representation, and recommends item 1049 to an improper user 117 who likes 1092. On the contrary, adopting our alignment method FETTLE, FREEDOM can correctly identify these two items as dissimilar and reduce false positives. The above example is not an isolated case. We discover FREEDOM produces 90, 132 false positives, approximately 30% of the dissimilar item pairs, where items (i, j) are dissimilar if i is less similar to j than 90% of all items.

Although multi-modal alignment has been overlooked in previous MRSs, it has been widely adopted in general-purpose Multi-Modal Models (MMMs) [4, 13, 14, 16, 21]. Modern MMMs typically utilize *bidirectional* alignment via contrastive learning, i.e., pull textual representations closer to visual representations, and vice

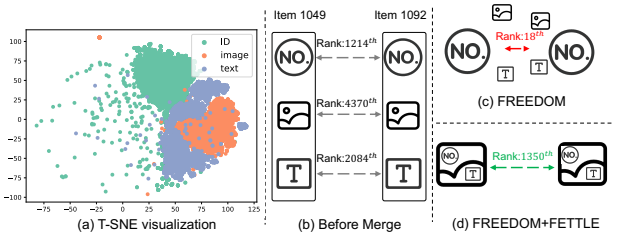

Figure 1: (a) Visualization of ID, image, and text embedding obtained by FREEDOM (before the merge step) in the Baby dataset (best shown in color). (b) Item 1049 has a low-ranked similarity with item 1902 in each modality. (c) The similarity between 1049 and 1902, based on FREEDOM's derived final item representations, is abnormally high. (d) The similarity between 1049 and 1902 based on FREEDOM+FETTLE's derived item representations remains low.

versa [13, 14, 21]. Bidirectional alignment is successful in MMMs because they aim to learn a unified multi-modal representation on a corpus where images and text describe the same item or concept in parallel. On the contrary, MRSs intend to understand user preferences, in which multi-modal content may not serve this purpose equally. We argue that the alignment in MRSs should only be allowed *in a one-way direction, depending on the user preference*. For example, if users prefer a jacket's visual elements (e.g., style, color) more than the jacket's other aspects (e.g., price, fabric) in the textual descriptions, then (a) the textual modality should be pulled closer to the visual modality for the jacket, and (b) the visual modality should not be brought closer to the textual modality, because it may adversely affect the prediction of user preference.

However, it is challenging to determine the one-way alignment direction. Specifically, three key challenges should be addressed.

C1: self-adapting item-level alignment direction. Naturally, the alignment direction varies for each item because of the item's inherent characteristics. For example, because consumers often buy fashion and jewelry items based on visual impressions, the textual modality is expected to be dragged toward the visual modality. An opposite example is smartphones. As most smartphones look similar, users often pay more attention to brands and functionalities depicted by textual content. The direction of the alignment should be from the visual to the textual modality. The problem is: How do we efficiently determine the alignment direction for each item?

C2: coordinating multiple modalities. MRSs have at least three modalities, i.e., visual, textual, and ID modalities, and there is more than one alignment direction to be decided. On the one hand, a *topmost* approach, i.e., pulling the other modalities toward one chosen modality, might lead to modality laziness [6], i.e., other modalities are not sufficiently refined during training. This problem is particularly pronounced in recommendation systems due to the domination of the ID modality, i.e., the ID modality will be chosen for most items, and the other modalities will not contribute to the RS. On the other hand, a *pair-wise* approach, i.e., a modality is chosen for every pair of modalities to ensure more than one modality is selected in the alignment process, raises an additional problem: If there is an inconsistency in the alignment directions, i.e., m is pulled to n, and m is also drawn to o, how do we alleviate the direction inconsistency?

C3: denoising alignment for items and users. Item-level alignment based on user feedback may not always be accurate because user feedback is inevitably noisy. For example, a user accidentally clicks an item, or clicks are missing because the item is not exposed to the target population. Moreover, the users have been neglected in the alignment. As user embeddings are dynamically updated in the learning process, directly aligning user embeddings and item embeddings is unstable. How can we derive a more robust multi-modal alignment for items and users?

In this paper, we propose FETTLE (**FE**edback-orien**T**ed mul**T**i-modal a**L**ignm**E**nt). FETTLE is a plug-and-play framework that operates on any Recommendation System's derived item representation and the pre-trained multi-modal content features. To address **C1**, FETTLE determines `Item-Level Alignment`, i.e., estimates the average user feedback for each modality based on the collected information in a training batch, and a modality with lower estimated score is oriented towards a higher-scored modality which indicates stronger user preference for this item. To address **C2**, FETTLE implements `Multi-Modal Alignment`, i.e., the item representations after pair-wise directional alignment are fine-tuned by maximizing their estimated feedback on interacted users to reduce irrelevant factors resulting from inconsistent alignment direction. To address **C3**, FETTLE proposes `Cluster-Level Alignment`, from both the item and user perspectives. Items are clustered by learning a multi-modal codebook for all modalities, and the cluster assignments are matched across each item's different modalities. As the item cluster-level alignment does not involve user feedback, it is more robust and unaffected by noisy feedback. Similarly, users and items are clustered w.r.t. the same interaction codebook, and interacted user-item pairs are matched via the more abstract and more stable cluster assignments.

We apply FETTLE to five conventional RSs and six MRSs and conduct experiments on three real datasets. FETTLE demonstrates significant performance improvements, average 24.79% – 62.79% improvements for conventional RSs and 5.91% – 20.11% for MRSs. FETTLE outperforms the state-of-the-art (SOTA) MRSs FREEDOM by 3.62% – 7.35% across all datasets.

To summarize, the main contribution of this work is three-fold.

(1) To our knowledge, we are the first to address the problem of multi-modal alignment direction in MRSs and propose to orient the item-level alignment direction for multiple modalities by user feedback.
(2) We advance beyond item-level alignment to cluster-level alignment and learn more robust user and item representations to alleviate noisy user feedback.
(3) We propose a non-invasive multi-modal alignment framework that can be easily plugged into numerous conventional or multi-modal recommendation systems and exhibit significant performance improvements on various datasets.

2 RELATED WORK

Multi-modal Alignment. Most state-of-the-art Multi-Modal Models(MMMs) implement multi-modal alignment with contrastive learning [4, 13–16, 21, 34]. For example, considering the visual

modality of an instance as the anchor sample, the same instance in the textual modality is a positive sample and other instances in the textual modality are negative samples. Similarly, given an instance in the textual model as an anchor sample, the same instance in the visual modality is the positive sample. In contrast, other instances in the visual modality are negative samples. Therefore, the alignment is bidirectional.

Multi-modal Recommendation Systems. Existing MRSs [10, 24, 26–28, 33, 33, 36] mostly follow a similar workflow. In the *pre-processing* stage, user ID embeddings and item ID embeddings are obtained with forward feedback networks(FFNs). Multi-modal embeddings are obtained with pre-trained models and a projector to make the dimensionality consistent. In the *learning* stage, multi-modal embeddings and item ID embeddings are fine-tuned by a FFN network [10] or a GNN network built on user-item bigraph [26–28] or item-item graph [33, 36]. In the *merge* stage, multi-modal embeddings and item ID embeddings are concatenated [24] or added [24, 26–28, 33, 36] and optimized with user ID embeddings by the Bayesian Personalized Ranking loss [23].

Remarks. To our best knowledge, only one recent work [37] explicitly performs multi-modal alignment in MRS by minimizing the cosine distance between multi-modal embeddings and ID embeddings. It is different from our proposed method FETTLE in two aspects. First, it falls into the bidirectional alignment category, while FETTLE advocates the one-way directional alignment. Second, it aligns multi-modality at the item level, while FETTLE involves instance-level and cluster-level alignment.

3 METHOD

We first briefly introduce the workflow of a Recommendation System (RS). Without loss of generality, let the input of the RS model be the set of users $\mathcal{U}$, the set of items $\mathcal{I}$, a binary user feedback matrix $\mathcal{Y}$, e.g., $\mathcal{Y}_{u,i} = 1, u \in \mathcal{U}, i \in \mathcal{I}$ indicates user u clicks item i. For a Multi-modal recommendation system (MRS), the input also includes multi-modal content features encoded by pre-trained models. This paper considers visual and textual content, i.e., image and text embeddings for each item. The RS eventually learns a vectorized representation for any item and user, denoted as the item representation $\mathbf{i}^{ID} \in \mathbb{R}^L, \forall i \in \mathcal{I}$ and the user representation $\mathbf{u} \in \mathbb{R}^L, \forall u \in \mathcal{U}$, where L is the embedding size. The user and item representations are usually optimized via a BPR loss [23].

FETTLE aims to support any existing RS models in a plug-and-play manner, whether a multi-modal recommendation approach (MRS) or a traditional collaborative filtering approach (CF). Therefore, FETTLE can only operate on the input (i.e., $\mathbf{i}^V, \mathbf{i}^T, \forall i \in \mathcal{I}, \mathcal{Y}$) and the output (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. Therefore, FETTLE is stacked on the existing RS models as a standalone component, as depicted in Figure 2(a). FETTLE consists of two major parts.

(1) **Pre-processing the input**. Since the input image and text embeddings often differ in dimension, FETTLE first feeds them through a projector to transform them to the same embedding size. We denote the derived item representations by existing RS as the ID modality $\mathbf{i}^{ID}$. Thus, we have a set of three different modalities for the item, i.e., $\mathcal{M} = \{ID, V, T\}$. Both modality-specific item embeddings $\mathbf{i}^V, \mathbf{i}^T, \mathbf{i}^{ID}$, and user embedding $\mathbf{u}$ are $L-$dimensional vectors.

(2) **Regularizing the BPR loss**. To provide a non-invasive framework, FETTLE adds three regularization terms to the original RS's BPR loss, namely the item-level alignment loss (Section 3.1), the multi-modal direction tuning loss (Section 3.2), and the cluster-level alignment loss (Section 3.3).

3.1 Item-Level Alignment

Bidirectional alignment loss has been widely adopted in Multi-Modal Models (MMMs) [4, 13, 14, 21], where each item's visual embeddings are pulled closer to its text embeddings, and vice versa.

$$\begin{aligned} \mathcal{L}^{MMM} &= \mathcal{L}^{V\to T} + \mathcal{L}^{T\to V}, \\ \mathcal{L}^{V\to T} &= -\sum_{i\in\mathcal{I}} \log \frac{\exp(\text{sim}(\mathbf{i}^V, \mathbf{i}^T)/\lambda}{\sum_{j\in\mathcal{I}} \exp(\text{sim}(\mathbf{i}^V, \mathbf{j}^T)/\lambda)}, \\ \mathcal{L}^{T\to V} &= -\sum_{i\in\mathcal{I}} \log \frac{\exp(\text{sim}(\mathbf{i}^T, \mathbf{i}^V)/\lambda}{\sum_{j\in\mathcal{I}} \exp(\text{sim}(\mathbf{i}^T, \mathbf{j}^V)/\lambda)}, \end{aligned} \tag{1}$$

where $i, j \in \mathcal{I}$ are two items, $\mathbf{i}^T, \mathbf{i}^V$ are the image and text embeddings, λ is the temperature parameter.

The aforementioned bidirectional alignment loss is easily influenced by noisy modalities, which exist widely in RSs. For example, some items contain many popular words unrelated to the item's content in the textual description [30]. The bidirectional alignment will orient high-quality images toward low-quality texts, degrading the image embeddings' quality.

Therefore, it is necessary to align modalities in a one-way direction. Intuitively, we can estimate the average user feedback for each modality on each item. If a modality $m \in \mathcal{M}$ has large estimated user feedback on item $i \in \mathcal{I}$, it implies that users generally prefer to make decisions related to item i based on its modality m. Considering m has higher quality and better relevance in decision-making, other modalities should be oriented toward it.

As shown in Figure 2(b), we traverse through the user-item pairs within a training batch to estimate the user feedback. Since the cost of computing the entire user-item interaction matrix is high, our calculation is only based on collected information in the batch to reduce computational resources. Formally, we obtain $\mathcal{S}_i$, the estimated user preference for item i on each modality, as

$$\mathcal{S}_i = \{s_i^V, s_i^T, s_i^{ID}\}, \tag{2}$$

where $s_i^m = \text{avg}_{\mathcal{Y}_{u,i}=1\in\mathcal{B}}(\mathbf{u}^T\mathbf{i}^m)$ for each $m \in \mathcal{M}$ (e.g. ID, textual, or visual modalities), with $\mathbf{u}$ being the RS's derived user embedding, $\mathbf{i}^m$ being the RS's derived modality-specific item embedding, $\mathcal{Y}_{u,i} = 1 \in \mathcal{B}$ being a pair of interacted user and item in the training batch. We only collect interacted user-item pairs as they indicate positive feedback. Non-interacted user-item pairs can be either due to data missing or negative feedback. It yields biased and uncertain results if they are included in estimating user feedback. Furthermore, avg means we calculate the average feedback over different users in the batch to minimize the effect of sample numbers.

The item-level alignment direction is determined based on $\mathcal{S}_i$, i.e, if $s_i^m < s_i^n, m \in \mathcal{M}, n \in \mathcal{M}, m \neq n$, then for the item i, the modality m should be pulled toward n, denoted as $m \to n$. We can rewrite Equation 1 to form the item-level directed alignment loss, i.e., for item i, given the alignment direction $m \to n$, we can

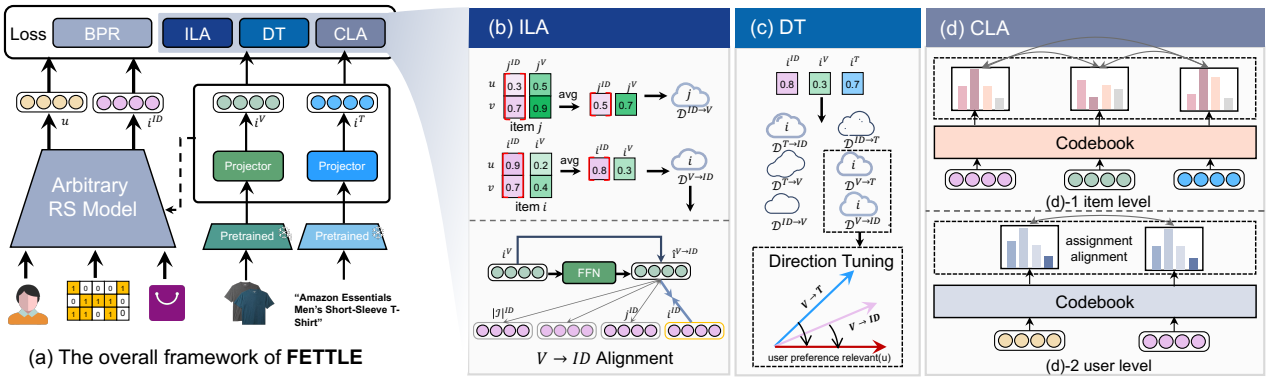

Figure 2: The overall framework: FETTLE works on an arbitrary RS's derived item representation and the pre-trained multi-modal content features; FETTLE adds three loss terms to the RS's original BPR loss.

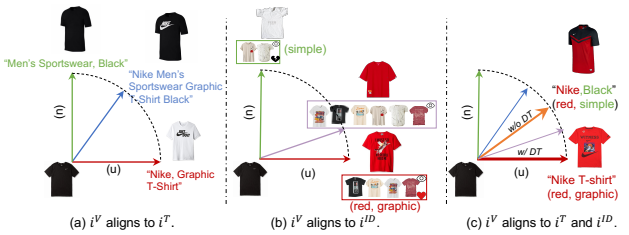

Figure 3: Illustration of the inconsistent alignment directions

maximize the similarity between $\mathbf{i}^m$ and $\mathbf{i}^n$. To more accurately capture what information needs to be emphasized in the alignment, we project $\mathbf{i}^m$ before the alignment using a Feed-Forward Network with a residual structure. Compared with existing studies such as ALBEF [15], which uses only a Feed-Forward Network (FFN) before alignment, the advantage of using a residual structure is to explicitly represent what information needs to be supplemented in the process of aligning from m to n. Formally,

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

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

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

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

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 ILA loss is calculated over all item-level alignment directions $m \to n$, which will be further explained in Section 3.2.

Remarks. The alignment is from low-scored to high-scored modality; thus, the alignment is non-cyclic (i.e., one-way direction).

3.2 Multi-Modal Alignment

To enumerate the item-level alignment directions $m \to n$ in Equation 4 across $|\mathcal{M}| > 2$ modalities, there are two approaches. *Topmost* approach orients all remaining $|\mathcal{M}| - 1$ modalities toward the topmost-scored modality for each item, and *pairwise* approach constructs $|\mathcal{M}|(|\mathcal{M}|-1)/2$ pairs of modalities, and within each pair orients a lower-scored modality toward a higher-scored modality. Clearly, the topmost method may lead to modality laziness [6]. The ID modality has been sufficiently optimized in the RS and is likely the topmost-scored modality for most items. Thus, other modalities will be dominated by the ID modality and left inactive and insignificant in the optimization objective.

Based on the rationale above, we adopt a pairwise approach. Since there are three different modalities, there are six possible pairs, i.e., $V \to T, V \to ID, T \to V, T \to ID, ID \to V$, and $ID \to T$. We can divide items into different subsets. Formally, $\mathcal{D}^{m\to n}$ is a set of items that support 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 2(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 $i \in \mathcal{D}^{T\to ID}$.

One limitation of the pairwise alignment approach is that the lowest-scored modality eventually aligns simultaneously with two other modalities. e.g., $\exists i \in \mathcal{I}, i \in (\mathcal{D}^{V\to T} \cap \mathcal{D}^{V\to ID})$. We call this direction inconsistency, and it is confusing for the model to produce a correct alignment.

Our intuition is illustrated in Figure 3 to overcome this limitation. After alignment, all modality-specific embeddings should be in the same vector space, spanned by two basis vectors, i.e., the user-preference-relevant vector (i.e., (u) in Figure 3) and the user-preference-irrelevant vector (i.e., (n) in Figure 3). Then, each item's modality-specific embeddings will be decomposed into the two basis vectors. For example, in Figure 3(a), the text description is "Nike Men's Sportswear Graphic T-shirt Black", the user preference relevant part is "Nike Graphic T-shirt", and the irrelevant part is "Men's Sportswear Black". Suppose the visual modality is aligned with the textual modality. In that case, the image after alignment will contain unpreferred elements such as "Sportswear Black." Similarly, in Figure 3(b), the ID modality can be decoupled into user preference relevant parts "red graphic" and irrelevant parts "white simple." If the visual modality is aligned with the ID modality, the image after alignment will contain unpreferred elements such as "simple."

Thus, to resolve modality inconsistency, we ensure that the direction is consistent with the user-preference-relevant basis vector. This means tuning the direction by maximizing the user preference score of the modality-specific embeddings after alignment. As shown in Figure 3(c), if we combine the two alignment directions, the image embeddings will contain undesired elements "black simple". After direction tuning, the image embeddings will contain only user-preferred elements "Nike T-shirt red graphic". Formally, we define the direction tuning loss $\mathcal{L}^{DT}$,

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

3.3 Cluster-Level Alignment

Item cluster-level alignment. The above alignments are guided by user feedback at the item level. Since user feedback may contain noise, e.g., the user accidentally clicked on an item due to a mistouch. The alignments based on this user feedback may sometimes be inaccurate. Next, we introduce a cluster-level alignment for items, which is to align different modalities with the cluster center. Because the item cluster-level alignment does not involve user feedback, it is more robust and stable.

Cluster-level alignment requires clustering the items. Traditional clustering methods such as KMeans [8] face high time costs. Inspired by SwAV [1], we can learn a codebook and maintain each item's cluster assignments on the fly without actually finishing the learning of the codebook.

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

We can obtain the cluster assignment for each item's modality-specific embedding $\mathbf{i}^m$, called the code $\mathbf{q}^{m,i} \in \mathrm{R}^P$. Ideally, the code can be determined by matching the embedding $\mathbf{i}^m$ to the cluster prototype using the softmax function, i.e., $\text{SoftMax}(\mathbf{i}^m\mathbf{C}^{\text{ITV}}/\tau)$. Furthermore, for recommendation systems, uniformity is crucial [25]. We utilize the Sinkhorn [5] optimal transport algorithm to generate codes that maintain information integrity while producing relatively uniform spatial distributions.

$$\mathbf{q}^{m,i} = \lim_{n\to N}\text{diag}(\mathbf{r}^{(n)})\cdot\text{SoftMax}(\frac{\mathbf{i}^m\mathbf{C}^{\text{ITV}}}{\tau})\cdot\text{diag}(\mathbf{s}^{(n)}), \quad (7)$$

where $\mathbf{q}^{m,i}, \mathbf{C}^{\text{ITV}}$ 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.

Item cluster-level alignment assumes the cluster assignments of an item's different modalities are similar. We utilize the cross entropy loss between the code $\mathbf{q}^{m,i}$ and the "ground-truth" assignment, which is calculated based on $\text{SoftMax}(\mathbf{i}^m\mathbf{C}^{\text{ITV}}/\tau)$. Formally, we minimize the cluster-level alignment loss for items,

$$\begin{aligned}\mathcal{L}^{CLA}_{ITV} = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{m,n\in\mathcal{M},m\neq n}[&\mathbf{q}^{n,i}\log(\text{SoftMax}(\frac{\mathbf{i}^m\mathbf{C}^{\text{ITV}}}{\lambda_c}))\\ &+\mathbf{q}^{m,i}\log(\text{SoftMax}(\frac{\mathbf{i}^n\mathbf{C}^{\text{ITV}}}{\lambda_c}))],\end{aligned} \quad (8)$$

where $\mathbf{q}^{m,i}, \mathbf{q}^{n,i}$ are obtained by Equation 7, $\text{SoftMax}(\frac{\mathbf{i}^m\mathbf{C}^{\text{ITV}}}{\lambda_c})$ 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 $\mathcal{L}^{CLA}_{ITV}$.

User cluster-level alignment. The above alignments focus on items, and users are ignored. However, aligning the user embeddings with item embeddings is problematic. User embeddings are dynamic during training, and subtle turbulence can lead to unstable results. Therefore, we are motivated to align users with items at the cluster level.

Similarly, we maintain a codebook $\mathbf{C}^{\text{UI}}$ for users and items to represent a set of preference cluster prototypes. For a user and item in an interacted user-item pair $\mathcal{Y}_{u,i} = 1 \in \mathcal{B}$, their code $\mathbf{q}^u, \mathbf{q}^i$ should be similar. For example, a user likes "comedy, horror movie", and his/her interacted movies are likely to fall in the genres "comedy, horror movie". Thus, similar to Equation 8, we define the cluster-level alignment loss for users:

$$\begin{aligned}\mathcal{L}^{CLA}_{UI} &= -\frac{1}{|\mathcal{B}|}\sum_{\mathcal{Y}_{u,i}=1\in\mathcal{B}}[\mathbf{q}^i\log(\text{SoftMax}(\frac{\mathbf{u}\mathbf{C}^{\text{UI}}}{\lambda_c})) + \mathbf{q}^u\log(\text{SoftMax}(\frac{\mathbf{i}^{ID}\mathbf{C}^{\text{UI}}}{\lambda_c}))],\\ \mathbf{q}^u &= \lim_{n\to N}\text{diag}(\mathbf{r}^{(n)})\cdot\text{SoftMax}(\frac{\mathbf{u}\mathbf{C}^{\text{UI}}}{\tau})\cdot\text{diag}(\mathbf{s}^{(n)}),\\ \mathbf{q}^i &= \lim_{n\to N}\text{diag}(\mathbf{r}^{(n)})\cdot\text{SoftMax}(\frac{\mathbf{i}^{ID}\mathbf{C}^{\text{UI}}}{\tau})\cdot\text{diag}(\mathbf{s}^{(n)}).\end{aligned} \quad (9)$$

Remarks. Unlike item-level alignment, the cluster-level alignment is bidirectional. There are two reasons for doing so. (1) The cluster-level alignment represents more abstract user/item/multi-modal features. It is more stable in the learning process. Thus, updating the cluster-level alignment and the codebook will not degrade the quality of a modality-specific embedding vector. (2) The cluster-level alignment is not guided by user feedback. Instead, it aims to capture the high-level category of items and communities of users. Naturally, the clustering is reciprocal. Therefore, it is unnecessary to restrict the cluster-level alignment to be in a one-way direction.

3.4 Optimization

The original loss function for recommendation systems is generally the BPR Loss $\mathcal{L}^{BPR}$, as shown below.

$$\mathcal{L}^{BPR} = -\sum_{(u,i,j)\in\mathcal{B}}\ln\sigma(\mathbf{u}\mathbf{i}^{ID} - \mathbf{u}\mathbf{j}^{ID}), \quad (10)$$

where $(u, i, j) \in \mathcal{B}$ is the sampled triple, where user u interacts with item i, $(\mathcal{Y}_{u,i} = 1)$, and item j is a negative sample $(\mathcal{Y}_{u,j} = 0)$. σ is the sigmoid function.

Finally, the overall loss includes the recommendation objective loss $\mathcal{L}^{BPR}$, the item-level alignment loss $\mathcal{L}^{ILA}$, the multi-modal direction tuning loss $\mathcal{L}^{DT}$, and the cluster-level alignment loss for users and items:

$$\mathcal{L} = \mathcal{L}^{BPR} + \alpha(\mathcal{L}^{CLA}_{UI} + \mathcal{L}^{CLA}_{ITV}) + \beta(\mathcal{L}^{ILA} + \mathcal{L}^{DT}), \quad (11)$$

where α, β are the weight coefficients.

4 EXPERIMENTS

We conduct experiments to answer the following questions.

RQ1: Does FETTLE enhance existing recommendation methods?
RQ2: How do different components affect FETTLE's performance?
RQ3: Is it necessary to perform directed alignment?
RQ4: Can FETTLE address the issue of modality misalignment?
RQ5: How sensitive is FETTLE to its hyper-parameters?

4.1 Experimental Setup

Datasets. Following existing multi-modal recommendation systems [33, 36, 37], we conduct experiments on three categories of Amazon review dataset [20]: Baby, Sports, and Clothing. Each item in the datasets is associated with a 4096-dimensional vector of visual featuress [9] obtained from a pre-trained Convolutional Neural Network and a 384-dimensional vector of textual features obtained from a sentence-transformer [22]. The statistics of each dataset are shown in Table 1. The raw data of each dataset is pre-processed with a 5-core setting on both users and items.

Evaluation Protocols. We use the 80-10-10 split for training, validating, and testing [33, 36, 37]. We use two widely-used evaluation metrics: Recall@K (R@K) and NDCG@K (N@K). We report the average value of all users in the test dataset under $K = 10, 20$.

Implementation. Following existing works [11, 33, 36, 37], we fix the users and items embedding size to 64 for all models and initialize their parameters with the Xavier method [7] and use Adam [12] as the optimizer. We develop FETTLE on a classical multi-modal recommendation platform, MMRec [35]. We perform a grid search to find the optimal settings for different backbone models. Specifically, for CF models, we search α and β within {1, 10, 100}, while for multimodal models, we search them within {0.0001, 0.001, 0.01}. As for λ_c and λ_f, we search them within {0.1, 0.2, 0.3} and {0.05, 0.1, 0.15}, respectively. The number of prototypes $P = 10240, 20480$ for $\mathbf{C}^{\mathbf{ITV}}$ and $\mathbf{C}^{\mathbf{UI}}$. The iteration number $N = 3$.

Backbones. FETTLE is a plug-and-play framework and can easily be applied to different backbone models. We experiment with the most widely used collaborative filtering models (**CF backbones**), which are based solely on interaction data, including BPR [23], LightGCN [11], SGL [29], DirectAU [25], and NCL [17]. We also experiment with multi-modal recommendation models (**MRS backbones**), which integrate the text and image embeddings in the datasets, including VBPR [10], GRCN [27], DualGNN [26], SLMRec [24], LATTICE [33], and FREEDOM [36]. For all the backbones, we use the open-source MMRec implementation. Our codes are available online at https://github.com/XMUDM/FETTLE.

Competitors. To our best knowledge, only one latest work BM3 [37] applies multi-modal alignment in MRS. The alignment component in BM3 aligns multi-modal content with IDs by maximizing the cosine similarity between text embeddings, image embeddings, and ID embeddings.

4.2 Comparative Study

To answer **RQ1**, we stack FETTLE and BM3 on different backbone RS models. Table 2 reports the performance of varying backbone models before and after multi-modal alignment using BM3 and FETTLE. We have the following observations.

Table 1: Statistics of the experimental datasets.

Dataset	# Users	# Items	# Interactions	Sparsity
Baby	19,445	7,050	160,792	99.88%
Sports	33,598	18,357	296,337	99.95%
Clothing	39,387	23,033	278,677	99.97%

(1) FETTLE *shows significant improvements on all backbone models.* For CF backbones, FETTLE achieved an average improvement of 39.45% in terms of R@10, 37.69% in R@20, 36.62% in N@10 and 36.61% in N@20. For MRS backbones, FETTLE achieved an average improvement of 12.12%, 10.52%, 11.77% and 10.99%. Even for the SOTA model FREEDOM, FETTLE achieved an average improvement of 5.30%, 3.73%, 5.88%, and 4.80%. We observed a smaller improvement of FETTLE on MRS backbones compared with CF backbones. This is because MRS models already incorporate multi-modal information, yet we still achieve a significant enhancement of approximately 11%. This indicates that MRS models, lacking multi-modal alignment, fail to exploit the potential of multi-modal information fully, and FETTLE alleviates this problem and further improves the recommendation performance.

(2) FETTLE *performs consistently well on all datasets.* For Baby dataset, FETTLE achieved an average improvement of 20.88% in R@10, 19.69% in R@20, 21.06% in N@10 and 20.31% in N@20. For Sports dataset, FETTLE achieved an average improvement of 15.22%, 14.72%, 13.51% and 13.50%. For Clothing dataset, FETTLE achieved an average improvement of 41.14%, 37.74%, 37.73% and 37.29%. We observed that FETTLE exhibits the most significant improvement on the Clothing dataset, while the improvement on the Sports dataset is relatively modest. This is attributed to the nature of the Clothing dataset, which typically includes fashionable items like clothing and pants. The images showcase the style of the products, and the text reflects the parameters of the products, making it highly dependent on both modalities. On the other hand, the Sports dataset mainly comprises sports-related products, showing a lower dependence on both modalities. FETTLE focuses on multi-modal alignment to better leverage multi-modal information. As a result, it achieves a better improvement on the Clothing dataset than the improvement on the Sports dataset. Nonetheless, FETTLE still achieves satisfying improvements even on the Sports dataset that is less multi-modal dependent.

(3) FETTLE *is steadily better than BM3.* BM3 performs poorly while FETTLE is significant regarding the average improvement. BM3 decreases the average R@10, R@20, N@10 and N@20. FETTLE achieved an average improvement of 25.75% in R@10, 24.05% in R@20, 24.10% in N@10, 23.70% in N@20. On most backbones, FETTLE is better than BM3. This verifies our assumption that bidirectional multi-modal alignment, neglecting the alignment direction at the item level, is not always beneficial in MRS. FETTLE can self-adapt the alignment direction based on user preferences and achieve significant improvements.

4.3 Ablation Study

To answer **RQ2**, we conduct extensive experiments to show the effectiveness of different components in FETTLE. To make the results more convincing, we use the SOTA MRS FREEDOM [36]

Table 2: Performance of CF/MRS models before and after applying BM3 and FETTLE. The best performance is highlighted in bold. Δ*Imp.* indicates improvements over vanilla models in percentage.

	Models	Baby				Sports				Clothing			
		R@10	R@20	N@10	N@20	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	0.0200	0.0295	0.0111	0.0135
	+BM3	0.0418	0.0649	0.0228	0.0287	0.0342	0.0554	0.0185	0.0239	0.0270	0.0425	0.0140	0.0180
	+FETTLE	**0.0500**	**0.0790**	**0.0272**	**0.0347**	**0.0579**	**0.0874**	**0.0310**	**0.0385**	**0.0451**	**0.0696**	**0.0248**	**0.0310**
	LightGCN	0.0465	0.0754	0.0250	0.0325	0.0561	0.0846	0.0308	0.0381	0.0341	0.0527	0.0189	0.0236
	+BM3	0.0293	0.0494	0.0160	0.0212	0.0416	0.0646	0.0224	0.0283	0.0086	0.0151	0.0045	0.0061
	+FETTLE	**0.0576**	**0.0884**	**0.0317**	**0.0395**	**0.0645**	**0.0967**	**0.0351**	**0.0434**	**0.0473**	**0.0698**	**0.0253**	**0.0310**
	SGL	0.0532	0.0820	0.0289	0.0363	0.0620	0.0944	0.0339	0.0423	0.0332	0.0586	0.0216	0.0266
	+BM3	0.0547	0.0867	0.0301	0.0383	0.0650	0.0996	0.0359	0.0449	0.0430	0.0647	0.0234	0.0290
	+FETTLE	**0.0585**	**0.0903**	**0.0325**	**0.0407**	**0.0706**	**0.1057**	**0.0386**	**0.0476**	**0.0516**	**0.0765**	**0.0284**	**0.0347**
	DirectAU	0.0231	0.0342	0.0128	0.0156	0.0391	0.0570	0.0218	0.0264	0.0302	0.0455	0.0165	0.0204
	+BM3	0.0242	0.0399	0.0127	0.0167	0.0340	0.0524	0.0180	0.0227	0.0348	0.0553	0.0181	0.0233
	+FETTLE	**0.0400**	**0.0619**	**0.0215**	**0.0272**	**0.0553**	**0.0828**	**0.0298**	**0.0369**	**0.0497**	**0.0731**	**0.0266**	**0.0326**
	NCL	0.0463	0.0750	0.0249	0.0323	0.0560	0.0842	0.0308	0.0381	0.0342	0.0499	0.0183	0.0224
	+BM3	0.0296	0.0493	0.0161	0.0212	0.0183	0.0300	0.0104	0.0134	0.0085	0.0149	0.0044	0.0061
	+FETTLE	**0.0552**	**0.0836**	**0.0298**	**0.0371**	**0.0643**	**0.0966**	**0.0354**	**0.0438**	**0.0433**	**0.0643**	**0.0234**	**0.0287**
	+BM3 Avg Δ*Imp.*	-11.21%	-7.46%	-11.57%	-9.49%	-23.87%	-20.61%	-25.06%	-23.09%	-14.04%	-13.09%	-21.60%	-18.07%
	+FETTLE Avg Δ*Imp.*	31.42%	30.52%	31.66%	30.96%	24.79%	24.87%	22.62%	22.97%	62.16%	57.69%	55.57%	55.87%
MRS	VBPR	0.0424	0.0662	0.0223	0.0284	0.0560	0.0857	0.0307	0.0384	0.0282	0.0420	0.0156	0.0191
	+BM3	0.0418	0.0661	0.0221	0.0284	0.0550	0.0834	0.0299	0.0373	0.0286	0.0418	0.0160	0.0193
	+FETTLE	**0.0555**	**0.0842**	**0.0297**	**0.0372**	**0.0622**	**0.0957**	**0.0330**	**0.0417**	**0.0454**	**0.0675**	**0.0242**	**0.0299**
	DualGNN	0.0507	0.0808	0.0277	0.0354	0.0589	0.0902	0.0325	0.0405	0.0458	0.0689	0.0243	0.0301
	+BM3	0.0525	**0.0844**	**0.0289**	**0.0371**	0.0576	0.0884	0.0311	0.0391	0.0441	0.0674	0.0237	0.0296
	+FETTLE	**0.0532**	0.0830	0.0285	0.0362	**0.0633**	**0.0929**	**0.0354**	**0.0431**	**0.0511**	**0.0739**	**0.0278**	**0.0336**
	GRCN	0.0520	0.0841	0.0284	0.0367	0.0603	0.0911	0.0327	0.0407	0.0428	0.0659	0.0225	0.0284
	+BM3	0.0515	0.0822	0.0279	0.0358	0.0600	0.0906	0.0329	0.0408	0.0442	0.0674	0.0233	0.0292
	+FETTLE	**0.0578**	**0.0900**	**0.0311**	**0.0394**	**0.0632**	**0.0964**	**0.0341**	**0.0426**	**0.0502**	**0.075**	**0.0266**	**0.0329**
	SLMRec	0.0535	0.0820	0.0293	0.0366	0.0660	0.0989	0.0365	0.0449	0.0451	0.0670	0.0243	0.0299
	+BM3	0.0497	0.0768	0.0269	0.0338	0.0580	0.0868	0.0319	0.0393	0.0374	0.0552	0.0201	0.0247
	+FETTLE	**0.0555**	**0.0847**	**0.0299**	**0.0375**	**0.0681**	**0.1008**	**0.0373**	**0.0457**	**0.0477**	**0.0697**	**0.0257**	**0.0313**
	LATTICE	0.0547	0.0843	0.0291	0.0367	0.0622	0.0953	0.0338	0.0423	0.0486	0.0717	0.0265	0.0324
	+BM3	0.0549	0.0853	0.0294	0.0373	0.0604	0.0940	0.0327	0.0413	0.0404	0.0593	0.0212	0.0260
	+FETTLE	**0.0569**	**0.0915**	**0.0310**	**0.0398**	**0.0655**	**0.0986**	**0.0351**	**0.0436**	**0.0531**	**0.0783**	**0.0288**	**0.0351**
	FREEDOM	0.0626	0.0986	0.0327	0.0420	0.0719	0.1076	0.0385	0.0477	0.0627	0.0940	0.0336	0.0415
	+BM3	0.0619	0.0985	0.0326	0.0420	0.0718	0.1081	0.0385	0.0479	0.0634	0.0936	0.0341	0.0418
	+FETTLE	**0.0672**	**0.1029**	**0.0355**	**0.0447**	**0.0745**	**0.1115**	**0.0397**	**0.0492**	**0.0658**	**0.0970**	**0.0356**	**0.0435**
	+BM3 Avg Δ*Imp.*	-1.11%	-0.54%	-0.97%	-0.61%	-3.27%	-3.06%	-3.69%	-3.42%	-5.31%	-5.95%	-5.36%	-5.70%
	+FETTLE Avg Δ*Imp.*	10.35%	8.85%	10.45%	9.66%	5.91%	4.91%	4.98%	4.61%	20.11%	17.80%	19.89%	18.70%

as the backbone, add different components of FETTLE, and report the R@10 performance. These components include (1) Item-Level Alignment(ILA), which determines the alignment direction based on user feedback and adopts the pairwise $\mathcal{L}^{ILA}$ in Section 3.1. (2) multi-modal alignment with the Direction Tuning loss (DT), which addresses the direction inconsistency that arises when aligning multiple modalities simultaneously by the $\mathcal{L}^{DT}$ in Section 3.2. (3) Cluster-level Alignment(CLA), which aligns the users and items, and different modalities of items, at the cluster level by the $\mathcal{L}^{CLA}$ in Section 3.3. First, we add the alignment components "+ILA" and "+CLA" to FREEDOM to observe the impact of the alignments. Then, we combine these two alignment methods(CLA&ILA) to observe whether CLA can mitigate the issue of inconsistent alignment directions in ILA caused by noisy user feedback. Finally, since DT should be combined with ILA, we perform (CLA&ILA&DT) to observe whether DT can solve the direction inconsistency. We have the following observations.

(1) *ILA is the most effective component in* FETTLE. Applying ILA alone can significantly enhance the backbone FREEDOM, "+ILA" achieved an average improvement of 2.48% in R@10, 2.20% in R@20, 3.76% in N@10 and 3.35% in N@20. This is a significant improvement because FREEDOM is already a sophisticated MRS, and it isn't easy to boost its performance. For example, applying the other multi-modal alignment approach, BM3 decreases FREEDOM's performance by -9.80% in R@10, -8.45% in R@20, -11.37% in N@10 and -10.06% in N@20. It shows that determining the direction of the item-level feedback-oriented alignment is necessary. Furthermore, applying ILA alone is more effective than applying CLA alone. Compared with "+CLA", "+ILA" achieved more average improvements.

(2) *CLA also improves the RS performance.* Compared with FREEDOM, "+CLA" achieved an average improvement of 1.49% in R@10, 1.01% in R@20, 3.07% in N@10 and 2.52% in N@20. Furthermore, CLA is the perfect companion to PA. Compared with "+ILA", "+ILA&CLA" achieved an average improvement of 1.68%, 1.03%, 1.28% and 1.10%. It shows that combining CLA with ILA effectively mitigates the

Table 3: Performance of different components of FETTLE

Datasets	Variants	R@10	R@20	N@10	N@20
Baby	FREEDOM	0.0626	0.0986	0.0327	0.0420
	+ILA	0.0647	0.1002	0.0343	0.0434
	+CLA	0.0628	0.0984	0.0335	0.0427
	+ILA&CLA	0.0664	0.1021	0.0352	0.0445
	+ILA&CLA&DT	**0.0672**	**0.1029**	**0.0355**	**0.0447**
Sports	FREEDOM	0.0719	0.1076	0.0385	0.0477
	+ILA	0.0730	0.1101	**0.0397**	**0.0493**
	+CLA	0.0727	0.1089	0.0395	0.0489
	+ILA&CLA	0.0742	0.1113	0.0395	0.0491
	+ILA&CLA&DT	**0.0745**	**0.1115**	**0.0397**	0.0492
Clothing	FREEDOM	0.0627	0.0940	0.0336	0.0415
	+ILA	0.0643	0.0965	0.0347	0.0429
	+CLA	0.0646	0.0959	0.0350	0.0429
	+ILA&CLA	0.0648	0.0966	0.0353	0.0434
	+ILA&CLA&DT	**0.0658**	**0.0970**	**0.0356**	**0.0435**

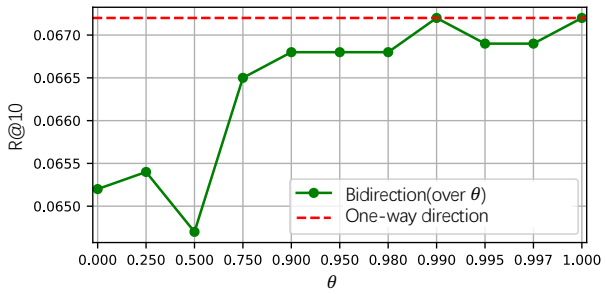

Figure 4: Performance of different alignment directions

issue of inaccurate alignment caused by noisy feedback, leading to further enhancement of RS performance.

(3) *DT can solve the direction inconsistency in ILA.* Compared with "+ILA&CLA", "+ILA&CLA&DT" achieved an average improvement of 1.05% in R@10, 0.46% in R@20, 0.74% in N@10 and 0.29% in N@20. Thus, DT has further enhanced the multi-modal item-level and cluster-level alignment.

4.4 Alignment Direction

Section 4.2 and Section 4.3 have already revealed that directed alignment is superior to bidirectional alignment by comparing ILA with BM3. In this section, we want to verify further that a one-way direction is necessary for multi-modal alignment in MRS (**RQ3**). As introduced in section 3.1, the item-level alignment direction is motivated by the assumption that bidirectional alignment might degrade the modality-specific embeddings by orienting a high-quality modality toward a low-quality modality. Thus, to investigate the effectiveness of one-way alignment direction, we compare it with bidirectional alignment based on the quality of modalities.

Specifically, we set a series of thresholds $\theta = 0.0 - 0.9$ with a step size 0.25. Focusing on high-quality modalities, we also test with smaller step sizes for $\theta \geq 0.95$. For each modality m and each item i, if the estimated feedback (Equation 2) is large, i.e., $s_i^m > \theta$, we employ bidirectional alignment. The underlying assumption is that

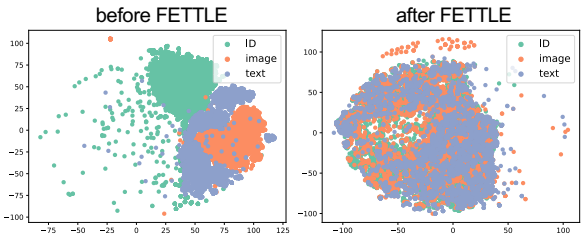

Figure 5: Visualization of ID, image, and text embeddings

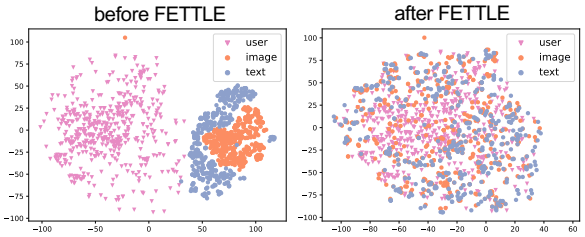

Figure 6: Visualization of interacted users and items

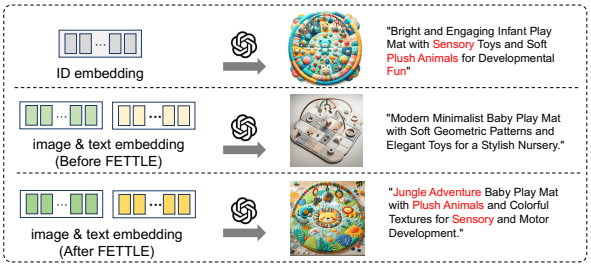

Figure 7: GPT -4's generated images and descriptions using item 42's ID, text, and image embeddings obtained before and after applying FETTLE

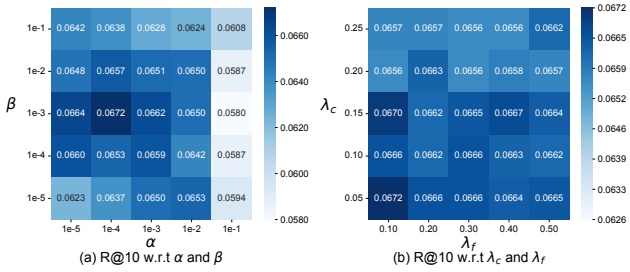

Figure 8: R@10 of FETTLE under different hyper-parameters

if $s_i^m > \theta$, then m is a high-quality modality. Thus, for any direction $m \rightarrow n$, and item $i \in \mathcal{D}^{m \rightarrow n}$, we also put it in $\mathcal{D}^{n \rightarrow m}$.

We evaluate the bidirectional and one-way directional alignment performance on the Baby dataset. As shown in Figure 4, the results

indicate that one-way directional alignment is always better than bidirectional alignment. A larger threshold suggests fewer items employ bidirectional alignment. Even at a threshold of 0.990, which means the items and modalities are prudently chosen, the R@10 for bidirectional alignment is only equivalent to directed alignment, which is 0.0672. The likely reason is that the multi-modal contents inevitably include parts irrelevant to user preferences. Bidirectional alignment can not distinguish the quality of multi-modal contents and spread the noise of certain modalities to different modalities, resulting in decreased recommendation performance.

4.5 Visualization

To demonstrate that FETTLE has reduced the large misalignment in MRS(**RQ4**), we implement FREEDOM with and without FETTLE on the Baby dataset and visualize the ID embeddings, image embeddings, and text embeddings by T-SNE. As shown in Figure 5, before applying FETTLE, the embedding space is well separated, especially since the ID and image embeddings are not in the same region. After applying FETTLE, the three types of embeddings are densely overlapped in the same region. Thus, we can conclude that FETTLE *effectively addresses the issue of multi-modal misalignment.*

Next, we aim to investigate whether multi-modal alignment benefits recommendation systems. We illustrate using backbone FREEDOM with and without FETTLE on the Baby dataset. We sampled 500 interactive user-item pairs and visualized the corresponding user embeddings and the multi-modal embeddings (i.e., image and text) of the items. Ideally, since the users and items have interacted, they must show a high degree of similarity, which means the user embeddings must correlate with the multi-modal embeddings of items in the embedding space. As shown in Figure 6, before applying FETTLE, the three types of embeddings are separated. Specifically, the user embeddings are far apart from the multi-modal item embeddings (including text and image embeddings). After applying FETTLE, the three types of embeddings reside in the same region, indicating that FETTLE can fully exploit the multi-modalities to predict users' preferences and successfully obtain high similarity for truly interacted user-item pairs. Thus, *the multi-modal alignment by* FETTLE *benefits the recommendation predictions.*

Furthermore, we aim to investigate whether FETTLE can improve the quality of multi-modal embeddings. We illustrate a case study using the backbone model FREEDOM with and without FETTLE on the Baby dataset. Since the datasets only provide the text and image embeddings without the raw data (i.e., without product description or product image), we use GPT-4 to generate images based on image embeddings for better demonstration purposes. Specifically, we select item 42, extract its ID embedding obtained by FREEDOM, and let GPT-4 generate an image based on the ID embedding. Although this image is not the actual product image, we can consider it the visual embodiment of the ID embedding from the GPT-4's understanding. As the ID embedding is optimized using the user-item interactions, and FREEDOM is a strong baseline, we consider its generated image close to the user's preference and denote them as the user-preferred image. We input the image embeddings before and after applying FETTLE to generate two images for comparison. We use GPT-4 to generate a description based on textual embeddings in the same way.

As shown in Figure 7, the image generated by the image embedding before applying FETTLE is dull and different from the user-preferred image. After applying FETTLE, the generated image is similar to the user-preferred image in color and style. The user-preferred description contains keywords like "sensory, plush animals, fun." While the description generated by the text embedding before applying FETTLE is also a "baby play mat", it does not contain these user-preferred keywords. The description generated by the text embedding after applying FETTLE correctly captures these keywords and has common elements with the user-preferred description. As these images (descriptions) are reconstructed from the embeddings, the above observations suggest that FETTLE *can improve the quality of multi-modal embeddings of items.*

4.6 Impact of Hyper-parameters

To investigate hyper-parameters impact (**RQ5**), we implement FREEDOM+FETTLE on the Baby dataset with different hyper-parameter settings. We focus on two sets of hyper-parameters, i.e., the loss weights β and α to balance the alignment losses, and the temperature λ_f and λ_c to refine the degree of attention dedicated to difficult alignment samples. Specifically, we vary the loss weight β and α in $\{1e-5, 1e-4, 1e-3, 1e-2, 1e-1\}$. Besides, we vary the temperature λ_f in $\{0.1, 0.2, 0.3, 0.4, 0.5\}$,and vary λ_c in $\{0.05, 0.10, 0.15, 0.20, 0.25\}$. The R@10 results are shown in Figure 8. We have the following observations.

(1) FETTLE *is robust to the loss weights.* For $\alpha = 1e-5$ ~$1e-2$ and $\beta = 1e-4$ ~$1e-2$, FETTLE's performance changes from 0.0642 to 0.0672 in R@10, as illustrated in Figure 8(a). As FREEDOM's performance is 0.0626 in R@10, FETTLE consistently enhances the performance of the backbone model FREEDOM with different values of α, β. Only in extreme settings, i.e., $\alpha = 1e-1$, ourmodel's performance is affected. Since α corresponds to the cluster-level alignment, a large α down-weighs the relative contribution of item-level directed alignment. This observation again verifies the importance of self-adapting alignment direction at the item level. The best performance is achieved at $\beta = 1e-4$ and $\beta = 1e-3$.

(2) FETTLE *is robust to temperature coefficients.* FETTLE is insensitive to the temperature λ_f,λ_c, as illustrated in Figure 8(b). All combinations of (λ_f,λ_c) can improve FREEDOM's recommendation performance by at least 4.79% in R@10. The best performance is achieved at $\lambda_f = 0.10$ and $\lambda_c = 0.05$, which is a moderate value.

5 CONCLUSION

Multi-modal alignment has been well-established in large Multi-Modal Models (MMMs) but has been largely overlooked in Multi-modal Recommendation Systems (MRSs). This paper argues that the multi-modal alignment in MRSs should not be bidirectional, as in most MMMs. We propose FETTLE (**FE**edback-orien**T**ed mul**T**i-modal a**L**ignm**E**nt), which contains three novel strategies for self-adapting item-level alignment direction, coordinating multiple modalities, and denoising alignment for items and users. FETTLE is a plug-and-play framework. Extensive experiments on three real-world datasets show that FETTLE significantly improves both traditional collaborative filtering and advanced multi-modal recommendation models and surpasses the SOTA recommendation systems.

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] 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

[13] 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).

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI*. 452–461.

[24] 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).

[25] 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.

[26] 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).

[27] 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.

[28] 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.

[29] 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.

[30] Guipeng Xv, Chen Lin, Wanxian Guan, Jinping Gou, Xubin Li, Hongbo Deng, Jian Xu, and Bo Zheng. 2023. E-commerce Search via Content Collaborative Graph Neural Network. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2885–2897.

[31] 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.

[32] 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.

[33] 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.

[34] 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).

[35] Xin Zhou. 2023. MMRec: Simplifying Multimodal Recommendation. *arXiv preprint arXiv:2302.03497* (2023).

[36] 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.

[37] 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.