Aligning and Balancing ID and Multimodal Representations for Recommendation

Binrui Wu*
Fudan University
Shanghai, China
23210180068@m.fudan.edu.cn

Shisong Tang*
KuaiShou Technology
Beijing, China
tangshisong@kuaishou.com

Fan Li
KuaiShou Technology
Beijing, China
lifan@kuaishou.com

Bing Han
KuaiShou Technology
Beijing, China
hanbing@kuaishou.com

Chang Meng
KuaiShou Technology
Beijing, China
mengchang@kuaishou.com

Jingyu Xiao
The Chinese University of Hong Kong
Hong Kong, China
jyxiao@link.cuhk.edu.hk

Jiechao Gao†
Center for SDGC, Stanford University
Palo Alto, CA, USA
jiechao@stanford.edu

Abstract

Large-scale recommendation systems mainly rely on sparse ID features, struggling with data sparsity. It's important to use multimodal information to assist ID learning for better performance. However, there exists two challenges: (1) **distribution discrepancy** between multimodal and ID makes direct integration prone to user-item mismatch; (2) slower convergence of multimodal representations compared to ID, causing **optimization imbalance** under a unified objective, which limits the potential of multimodal representations. In this paper, we comprehensively investigate the two problems and proposes a framework named AB-Rec to align and balance ID and multimodal representations learning for recommendation. We design three alignment tasks to fine-tune a pre-trained multimodal large language model (MLLM), which is then utilized to generate a unified multimodal representation for each item. AB-Rec aligns the distributions of ID and multimodal representations by minimizing the in-batch Wasserstein distance, and maximizes the distance between the two types of representations for the same item to avoid representation collapse. To solve the optimization imbalance, we propose a gradient modulation method that adaptively controls the optimization process by monitoring the contribution differences between ID and multimodal representations. Finally, we conduct extensive offline experiments on four datasets and an A/B test on an online video platform, demonstrating the effectiveness and scalability of our proposed method.

*Equal contributions.
†Corresponding Author.

KDD '25, August 3–7, 2025, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737275

CCS Concepts

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

Keywords

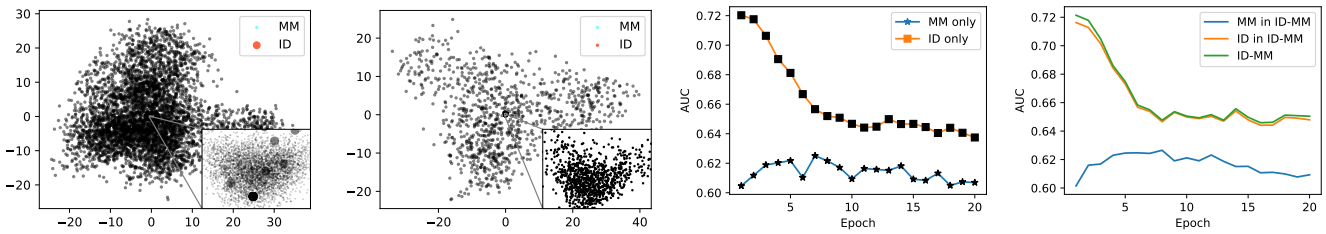
Multimodal recommendation, Multimodal LLM

ACM Reference Format:
Binrui Wu, Shisong Tang, Fan Li, Bing Han, Chang Meng, Jingyu Xiao, and Jiechao Gao. 2025. Aligning and Balancing ID and Multimodal Representations for Recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3711896.3737275

1 Introduction

Recommendation systems are widely employed across various online platforms [15, 19, 41, 42], including e-commerce (e.g., Taobao), social media (e.g., Instagram), and online video (e.g., TikTok and Kuaishou). Large-scale industrial recommendation systems primarily rely on sparse ID features, leveraging their strong fitting capacity to memorize patterns between users and items. However, ID features inherently lack the ability to capture the semantic content of items [39, 54], exhibit limited generalization, and thus are insufficient for handling long-tail items, data sparsity, and cold-start.

Existing studies have shown that integrating multimodal information (e.g., images and text) with categorical features (e.g., ID and category) can improve recommendation performance [9, 23, 46, 51, 53, 56, 58, 59]. Several methods [9, 46, 47, 53, 58] use multimodal information as auxiliary features to improve the learning of categorical features. BM3 [59] jointly trains content loss and recommendation objectives, while AlignRec [26] employs contrastive learning for alignment. Meanwhile, Multimodal Large Language Models (MLLMs) such as GPT-4V [1] and Gemini [43] have achieved remarkable advancements in multiple areas. These models are capable of processing cross-modal data and generating unified representations with strong semantic understanding. Therefore, leveraging



(a) item (b) user-item pair (c) ID vs. MM (d) ID & MM

Figure 1: Results of using DCN as recommendation model on Amazon dataset: (a) Visualization of item ID and multimodal representations; (b) Visualization of ID and multimodal top-representations (i.e., the input vectors to the classifier); (c) Validation-set AUC of the ID-only model and the multimodal-only model; (d) Validation-set AUC of the combined model (ID-MM), where "MM in ID-MM" indicates the validation-set AUC contributed by the multimodal component within the ID-MM model.

MLLM-generated multimodal representations to enhance recommendation performance has emerged as a critical direction.

However, effectively integrating multimodal information into recommendation systems still encounters two challenges:

- **Distribution discrepancy** (Fig. 1(a)-(b)): There is a significant difference in the distributions of multimodal and ID representations. Multimodal data typically contains rich semantic information, whereas ID representations mainly memory matching patterns. Directly combining the two may thus lead to user–item mismatches and hinder recommendation performance.
- **Optimization imbalance** (Fig. 1(c)-(d)): Multimodal representations generally converge more slowly than ID features. Under a unified training objective, this discrepancy in convergence speeds can lead to an imbalance in optimization, thereby limiting the full potential of multimodal features.

To address the aforementioned challenges, we propose a framework called AB-Rec, designed to align and balance the learning of ID features and multimodal representations, thereby enhancing recommendation performance. First, we employ a pre-trained multimodal large language model (MLLM) to generate unified item representations from textual and visual modalities. To ensure the robustness and consistency of these representations, we fine-tune the MLLM in three respects: modal alignment, meta-data processing, and multimodal robustness. To reduce the distribution gap between multimodal and ID features, AB-Rec aligns the distributions of both representations for each user–item pair by minimizing the in-batch Wasserstein distance [36], while maximizing the distance between the two types of representations for the same pair to prevent representation collapse. Furthermore, to address the optimization imbalance caused by the differing convergence speeds of ID and multimodal features, we introduce a gradient modulation method that adaptively adjusts the optimization strategy by monitoring their relative contributions during training, thus ensuring more effective parameter learning.

Our contributions can be summarized as follows:

- We systematically investigate the distribution discrepancy and optimization imbalance between multimodal and ID representations, and propose a framework called AB-Rec to align and balance two representations.
- We reduce the distribution discrepancy by minimizing the in-batch Wasserstein distance and maximizing the distance between the two types of representations for the same item, thereby avoiding representation collapse.
- To solve optimization imbalance, we propose a gradient modulation method to effectively coordinate the training process.
- We conduct extensive experiments on four real-world datasets and an online video platform, demonstrating the effectiveness and scalability of AB-Rec.

2 Related Work

ID-based Recommendation. Currently, ID-based Recommendation (IDRec), which relies on user/item IDs, forms the core of industrial recommendation systems [4, 57]. These models typically use user-item pairs as input and integrate them with additional features to predict the matching score between users and items. Early ID-based models include item-to-item collaborative filtering [15, 24] and shallow factorization models [16, 33]. With the advent of deep learning, ID-based recommendation models have generally adopted embedding and MLP architectures. The embedding module converts each discrete ID from the original input into low-dimensional vectors, while the MLP module aggregates the embedded vectors and performs the final prediction. Although ID-based models are widely used, they have significant limitations, such as the inability to capture semantic information [28] and the persistent challenge of the cold-start problem [38, 62].

Modality-based Recommendation. Modality-based Recommendation (MoRec) replaces the ID embedding used in IDRec with an item modality encoder (ME) [29, 50, 54], focusing on modeling the content features of items based on their modalities (including text [29, 49], image [27], and text-image multimodal pairs [50]). Various methods have been proposed in news recommendation systems, such as PVisRec [29], which utilized a denoising autoencoder to represent news articles and a GRU model to capture user interests from previously clicked articles; NRMS [49], which leveraged multi-head self-attention networks to analyze news based on titles; and MM-Rec [50], which detected ROIs (regions of interest) from

news images using object detection and encoded them with news text in a co-attention Transformer.

ID & Multimodal Recommendation. Another approach to utilizing modal information for improved recommendations is to integrate ID features with multimodal features. Early methods typically concatenated visual or text features with ID-based features [3, 9, 46, 47]. For instance, VBPR [9] enhances BPR-based recommendations [34] by incorporating visual features. Recently, some self-supervised learning methods, such as BM3 [59], have simultaneously optimized inter-content loss and recommendation-related user-item loss. Additionally, certain methods introduce auxiliary losses using multimodal information to derive better categorical embeddings through graph structures, such as FREEDOM [58]. AlignRec [26] decomposes the recommendation objective into three alignments. First, it pretrains the initial alignment to obtain unified multimodal features. Subsequently, these multimodal features are utilized as input to jointly train the remaining two alignments.

Multimodal Large Language Models. Recent years have seen substantial advancements [21, 22, 32] in MultiModal (MM) pretraining research, but these improvements come at the cost of increased computational demands. To tackle this, researchers are using pre-trained large language models (LLMs) [14] to create Multimodal Large Language Models (MLLMs) [52, 55] that integrate language with other modalities. Initial research primarily focuses on multimodal content understanding and text generation, including tasks such as image-text understanding, with examples like BLIP-2 [20], LLAVA [25]. Subsequently, the capabilities of MLLMs have been expanded to support specific modal outputs. This includes tasks with image-text outputs, such as Kosmos-2 [30]. Recently, MLLMs like GPT-4 [1] and Qwen2-VL [44] have demonstrated impressive multimodal understanding and generation capabilities.

3 Method

We first introduce a formal definition of multimodal recommendation along with the associated notation in Section 3.1. In Section 3.2, we detail the fine-tuning methodology for the MLLM and the process of generating multimodal representations. In Section 3.3, we describe how AB-Rec utilizes the Wasserstein distance to mitigate the significant distributional discrepancy between ID and multimodal representations. In Section 3.4, we theoretically examine the optimization imbalance phenomenon that arises during the joint learning of ID and multimodal representations. To address this issue, we propose a gradient modulation method in Section 3.5.

3.1 Problem definition

Let $\mathcal{U}$ represent the set of users, defined as $\mathcal{U} = \{u_1, u_2, \ldots, u_{|\mathcal{U}|}\}$, and let $\mathcal{I}$ denote the set of items, expressed as $\mathcal{I} = \{i_1, i_2, \ldots, i_{|\mathcal{I}|}\}$. Each sample is represented by the tuple $(u, i, \mathcal{M}_i, r_{u,i})$, where $u \in \mathcal{U}$, $i \in \mathcal{I}$, $\mathcal{M}_i = \{m_i^v, m_i^t\}$ represents the multimodal information (visual and textual) corresponding to item i and the label $r_{u,i}$ is a binary indicator that indicates whether there is a positive interaction between user u and item i. The problem of multimodal recommendation, therefore, is to learn a prediction function, aiming to predict the binary label $r_{u,i}$ based on the input features.

3.2 Multimodal Representation Generation

Inspired by recent advancements in MLLMs [1, 43, 44, 52, 55], we employ a pretrained MLLM to unify textual descriptions and visual information (e.g., images/videos) into a cohesive embedding representation. Although pretrained MLLMs exhibit robust comprehension and representation abilities, their original evaluation metrics are not specifically designed for recommendation tasks. Specifically, the alignment strategies of these models across modalities in multimodal recommendation scenarios remain under explored and are not carefully designed. To address this limitation and improve the extraction of multimodal features, we introduce three alignment objectives for fine-tuning the MLLM. These objectives are designed to improve the alignment between modalities and enhance the model's performance in recommendation tasks. Furthermore, we append a special token `[Item_cls]` to the end of each item description, enabling the model to condense extensive multimodal token sequences into compact and informative embeddings.

For item i, a prompt is initially employed to integrate its corresponding multimodal information $\mathcal{M}_i$ attributes into a unified description. This prompt can be phrased as follows: "Integrate the textual and visual information into an embedding representation. Textual input: `[Text]`, Visual input: `[Image/video]`." Subsequently, this description is transformed by a MLLM (denoted as MMenc) into a token sequence that concludes with the `[Item_cls]` token, resulting in: $\{t_1, t_2, \ldots, t_m, \texttt{[Item\_cls]}\}$. After encoding by the MLLM, the hidden state associated with the `[Item_cls]` token is extracted as the multimodal embedding for item i.

$$\mathbf{e}_i^{mm} = MMEnc(m_i^v, m_i^t)\,. \tag{1}$$

During the fine-tuning phase, we designed three specific alignment tasks for multimodal recommendation to enhance the adaptability of MMEnc in recommendation scenarios:

Multimodal content alignment. To improve the alignment between visual information and textual descriptions in recommendation scenarios, we propose a method inspired by BERT [6]. Specifically, for an item i we randomly replace 20% of the input text T_i with a special `[MASK]` token, resulting in the masked text $\hat{T}_i$, representing the masked portion of the text. The masked text and the item's visual information are then used as input, with the corresponding complete textual description serving as the target. This alignment encourages the model to learn the unique relationship between visual and textual context.

Metadata processing. Textual information in recommendation systems often encompasses both structured metadata (e.g., titles, prices, tags, and versions) and unstructured text descriptions. Enhancing metadata processing is crucial, as items are frequently associated with metadata, which effectively characterize the items. Enhancing metadata understanding can strengthen the encoding capability of MLLMs in recommendation tasks. Specifically, for a given item i, its metadata serves as input, while its detailed textual description acts as the target. This approach bridges structured metadata with unstructured text.

Robustness of Multimodal Features. To enhance representation and generalization capabilities through data augmentation, we apply a masking strategy to partial tokens of both textual and visual modalities for a given item i. The masked inputs are then fed into MLLM to generate the hidden state corresponding to the

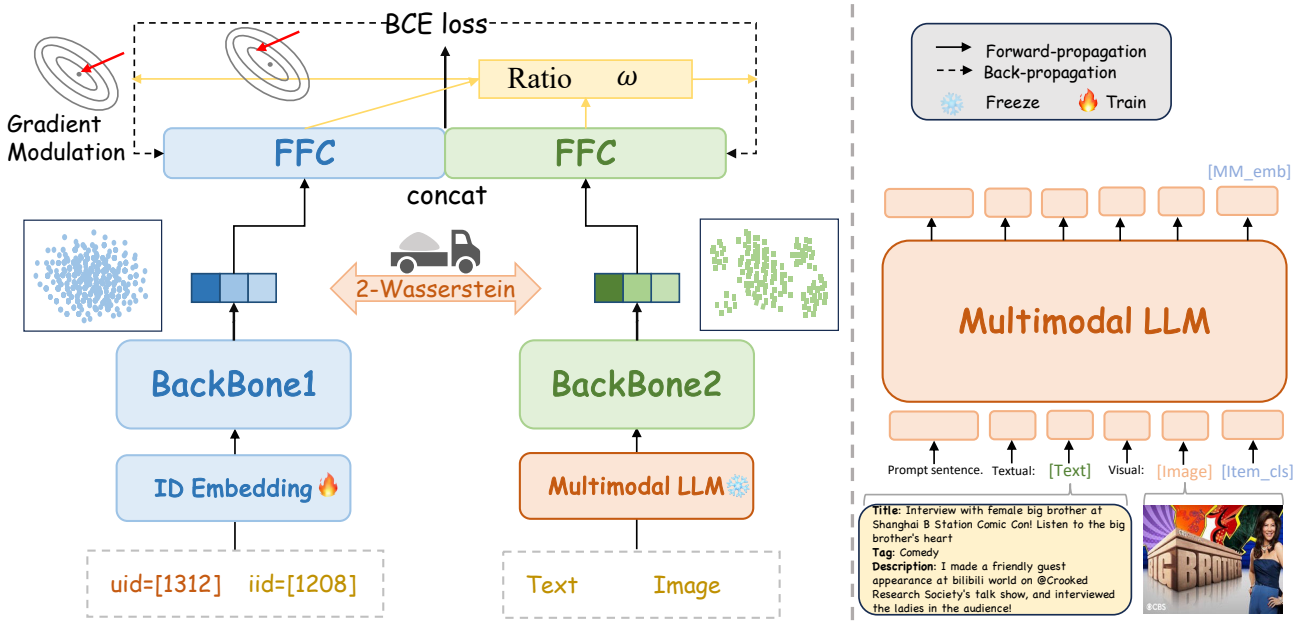


Figure 2: The architecture of our proposed method.

`[Item_cls]` token, denoted as emb_1. Simultaneously, we utilize the complete textual and visual tokens as inputs to generate the full multimodal features, denoted as emb_2, which serve as the fine-tuning target. This approach encourages the model to learn robust representations by reconstructing the complete multimodal features from partially obscured inputs, thus enhancing its generalization ability in recommendation scenarios.

3.3 Distribution Alignment

As described in Section 3.2, we obtain multimodal representations through the fine-tuned MLLM, which typically capture rich semantic and visual information. In contrast, ID representations are primarily utilized for memory-based matching, resulting in a significant distributional discrepancy between them. Directly combining these representations can lead to mismatches between users and items, thus degrading recommendation performance. This discrepancy can hinder the model's ability to effectively align user preferences with item features. To address this issue, AB-Rec balances the distributional differences by minimizing the in-batch Wasserstein distance between the two feature types. Additionally, to prevent representation collapse, it maximizes the distance between the two representations within the same pair.

First, we use trainable embedding layers to obtain the ID-based representations of items and users, denoted as $\mathbf{e}_i^{id}$ and $\mathbf{e}_u^{id}$, respectively. For item i, its multimodal representation $\mathbf{e}_i^{mm}$ is generated using the fine-tuned MLLM, as detailed in Section 3.2. For user u, the multimodal representation $\mathbf{e}_u^{mm}$ is computed by averaging the multimodal representations of the most recent k items interacted with by u , which are derived using the same approach. which are obtained using the same approach. The multimodal representation for user u is computed as follows:

$$\mathbf{e}_u^{mm} = \frac{1}{k} \sum_{j \in \mathcal{I}_u} \mathbf{e}_j^{mm} \tag{2}$$

where $\mathcal{I}_u = \{i_i^u, i_2^u, \ldots, i_k^u\}$ denotes the set of the k most recently interacted items by user u.

Second, we concatenate the two features to form $\mathbf{e}^{id}$, which acts as the input feature vector for the ID-based branch. Similarly, $\mathbf{e}^{mm}$ is obtained using the same approach. These features are then processed through two distinct neural networks to generate the ID representation and the multimodal representation, respectively. This process can be formally expressed as follows:

$$\mathbf{h}^{id} = f(\mathbf{e}^{id} = [\mathbf{e}_u^{id}, \mathbf{e}_i^{id}]; \theta^{id}) \tag{3}$$

$$\mathbf{h}^{mm} = g(\mathbf{e}^{mm} = [\mathbf{e}_u^{mm}, \mathbf{e}_i^{mm}]; \theta^{mm}) \tag{4}$$

For a batch of size n containing user-item pairs,we obtain the following representations through forward propagation:

$$H^{id} = \{\mathbf{h}_1^{id}, \mathbf{h}_2^{id}, \ldots, \mathbf{h}_n^{id}\}, \quad H^{mm} = \{\mathbf{h}_1^{mm}, \mathbf{h}_2^{mm}, \ldots, \mathbf{h}_n^{mm}\}, \tag{5}$$

where H^{id} and H^{mm} denote the ID-based and multimodal representations, respectively. To measure the distributional discrepancy between these two types of representations, we compute the Wasserstein distance [36] within the batch as follows:

$$W_2^2(H^{id}, H^{mm}) = \|\mu^{id} - \mu^{mm}\|^2 + \mathrm{Tr}\left(\Sigma^{id} + \Sigma^{mm} - 2(\Sigma^{id}\Sigma^{mm})^{1/2}\right) \tag{6}$$

where μ^{id} and Σ^{id} represent the mean and covariance matrix of the ID-based representations, respectively:

$$\mu^{id} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{h}_i^{id}, \quad \Sigma^{id} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{h}_i^{id} - \mu^{id})(\mathbf{h}_i^{id} - \mu^{id})^T \tag{7}$$

Similarly, μ^{mm} and Σ^{mm} are computed for the multimodal representations using the same procedure.

To ensure robust learning and prevent representation collapse, we introduce a regularization term to maximize the cosine distance between the ID and multimodal representations of the same user-item pair. The overall alignment loss $\mathcal{L}_{align}$ is formulated as:

$$\mathcal{L}_{align} = \alpha * \frac{1}{n_b} \sum_{k=1}^{n_b} 2/(1+e^{W_2^2(H_k^{id}, H_k^{mm})}) + \beta * \frac{1}{2N} \sum_{i=1}^{N} (1-\cos(\mathbf{h}_i^{id}, \mathbf{h}_i^{mm})) \tag{8}$$

where $n_b = N/n$, N represents the total size of the dataset, α, β are hyperparameters controlling the trade-off between the Wasserstein distance and the cosine regularization, and $\cos(\cdot,\cdot)$ denotes the cosine similarity. This formulation ensures distinct representations while aligning their distributions, thereby enhancing the model's robustness and generalization capability.

3.4 Optimization Imbalance Analysis

In the joint learning of ID and multimodal representations for recommendation, we observe an optimization imbalance phenomenon, where one representation dominates the learning process, causing the other to be under-optimized. We introduce the analysis of the optimization imbalance phenomenon for the model with concatenation as fusion method. In our recommendation model, the logits output is formulated as:

$$\varphi(x_i) = W[f(\mathbf{e}^{id}; \theta^{id}); g(\mathbf{e}^{mm}; \theta^{mm})] + b \tag{9}$$

To observe the optimization process of each component individually, W can be represented as the combination of two matrix: W^{id} and W^{mm}. The Equation 9 can be rewritten as:

$$\varphi(x_i) = W^{id} * f(\mathbf{e}^{id}; \theta^{id}) + W^{mm} * g(\mathbf{e}^{mm}; \theta^{mm}) + b \tag{10}$$

The model is trained using binary cross-entropy loss:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \log \hat{y}_i + (1-y_i) \log(1-\hat{y}_i) \right), \tag{11}$$

where the predicted probability is $\hat{y} = \sigma(\varphi(x_i))$. The gradient of the loss with respect to the logit is:

$$\frac{\partial L}{\partial \varphi(x_i)} = \sigma(\varphi(x_i)) - y_i. \tag{12}$$

Applying the chain rule, the gradients for the ID weights and backbone parameters of the ID representation are as follows:

$$\frac{\partial L}{\partial W^{id}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial L}{\partial \varphi(x_i)} * f(\mathbf{e}^{id}; \theta^{id}) \tag{13}$$

$$\frac{\partial L}{\partial \theta^{id}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial L}{\partial \varphi(x_i)} W^{id} \frac{\partial f(\mathbf{e}^{id}; \theta^{id})}{\partial \theta^{id}} \tag{14}$$

According to Equations 13 and 14, the optimization of W^{id} and $f(\cdot)$ is nearly independent of the optimization of multimodal parameters W^{mm} and $g(\cdot)$, except for the term associated with the training loss. As a result, the backbone have limited ability to make adjustments based on feedback from one another. The analysis of the feed-forward and back-propagation stages reveals that both the model predictions and gradients are governed by the sum of ID and multimodal (MM) components. Since ID features converge faster and contain stronger discriminative information, they dominate the model prediction $\varphi(x_i)$ and gradient $\frac{\partial L}{\partial \varphi(x_i)}$ through $W^{id} \cdot f(\mathbf{e}^{id}; \theta^{id})$. Even if the MM representations remain under-optimized and produce erroneous outputs during training, the more informative ID components can still "correct" these errors through summation, thereby influencing both the feed-forward and back-propagation processes. Consequently, according to Eq. (9) and Eq. (11), the MM, which has relatively lower confidence in the correct category, receives limited optimization, leading to its under-utilization. Based on this analysis, ID features play a dominant role in the optimization process. As the model approaches convergence, MM components may still require further training to compensate for their under-optimized features.

3.5 Gradient Modulation

Firstly, we follow the standard stochastic gradient descent (SGD) framework, where the parameters are updated as:

$$\theta_{t+1} = \theta_t - \kappa * \nabla_\theta \tilde{L}(\theta_t), \tag{15}$$

where θ_t denotes model parameters at iteration t, and κ is learning rate. $\tilde{L}(\theta_t) = \frac{1}{|\mathcal{B}_t|} \sum_{x_i \in \mathcal{B}_t} l(x_i; \theta_t)$ is the mini-batch loss, $\mathcal{B}_t$ is the t-th mini-batch.

As discussed, the optimization dynamics of ID and MM representations in recommendation model are often imbalanced, leading to under-optimization of the weaker representation. To mitigate this issue, we introduce the gradient modulation method, which adjusts the learning process by monitoring the relative contribution of each representation to the training objective.

We define a contribution score s that quantifies the influence of each representation in the training objective. Specifically, for a given sample x_i, the contribution score is formulated as:

$$s = y * p + (1-y) * (1-p). \tag{16}$$

where $p = \sigma(W * f(\mathbf{e}; \theta) + b/2)$ represents the predicted probability of the positive class, which serves as an approximation of the uni-modal prediction in the recommendation model. Here, we define the contribution difference ratio between ID and MM as follows:

$$\gamma_t^{id} = \frac{\sum_{x_i \in \mathcal{B}_t} s_i^{id}}{\sum_{x_i \in \mathcal{B}_t} s_i^{mm}}. \tag{17}$$

A higher value of γ_t^{id} indicates a stronger influence of ID representation in the training process. And γ_t^{mm} is accordingly defined as the reciprocal of γ_t^{id}. To mitigate the optimization imbalance, We can use the modulation coefficient ω_t to adaptively adjust the gradient updates:

$$\omega_t = \begin{cases} 1 - \tanh(\eta \cdot \gamma_t), & \gamma_t > 1 \\ 1, & \text{others} \end{cases} \tag{18}$$

where η is a hyper-parameter that controls the modulation strength. When the ID representation contributes significantly more than MM ($\gamma_t^{id} > 1$), the update for ID is scaled down to prevent excessive reliance on ID-based features, ensuring that the MM representation receives adequate optimization efforts. Integrating this modulation into the gradient-based optimization, the parameter update follows:

$$\theta_{t+1} = \theta_t - \kappa * \omega_t * \nabla_\theta \tilde{L}(\theta_t) \tag{19}$$

Table 1: Dataset Statistics.

Dataset	#Users	#Items	#Inters.	Sparsity
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%
Electronics	192,403	63,001	1,689,188	99.99%
Industrial	174267	610309	6005100	99.99%

By applying this gradient modulation, the optimization imbalance is effectively reduced, ensuring that both ID and MM representations contribute meaningfully to the recommendation model.

4 Experiment

4.1 Setup

4.1.1 Dataset. We adopt the offline experiments on industrial datasets and three categories of the Amazon dataset [27][1] for reproducibility. In these datasets, each review rating is considered a record of a favorable interaction between user and item. The statistical results of the datasets are shown in Table 1.

- The extensive industrial dataset(referred to as Industrial), has been gathered from a real-world streaming short-video platform. It includes approximately 6 billion impressions and over 50 million daily users. We collected interaction logs from a subset of users over a span of 14 days and utilized the data from the following day for evaluation purposes.
- We perform experiments on three categories of the Amazon dataset [27]: (a) *Baby*; (b) *Sports*; (c) *Electronics*. The raw data of each dataset are pre-processed with a 5-core setting on both items and users following [26, 59].

4.1.2 Baselines. Our baseline is divided into two parts. The first part presents a comparison between **AB-Rec** and other state-of-the-art multimodal recommendation methods, including **VBPR** [9], **BM3** [59], **FREEDOM** [58], **AlignRec** [26]. The second part demonstrates that **AB-Rec** is not dependent on the backbone architecture illustrated in Figure 2. To validate the generalization capability of our approach, we employ diverse backbone architectures, including **MLP** [17], **DCN** [45], **Fibinet** [12], **AutoInt** [40]. Notably, in our experimental setup, the backbone architectures for both the ID side and the multimodal side are kept consistent.

4.1.3 Evaluation Metrics. To ensure a fair comparison, we employ three widely-used evaluation metrics:

- AUC: AUC (Area Under the Curve) [11] refers to the area under the ROC curve. A higher AUC value indicates stronger classification capability.
- LogLoss: LogLoss (logarithmic loss) [8] is a loss function utilized to assess the performance of classification models. A smaller LogLoss value signifies more accurate predictions.
- *Recall@K*: denoted as *R@K* [31], measures the proportion of relevant items successfully retrieved within the top-N recommendations. It reflects the model's ability to identify and prioritize items that align with user preferences.

[1]The dataset can be accessed athttps://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html.

Table 2: Comparison of Zero-Shot Recommendation On Amazon dataset, measured by *R*@20 and *R*@50.

Generation	Baby		Sports		Electronics	
Methods	R@20	R@50	R@20	R@50	R@20	R@50
Amazon	0.0052	0.0150	0.0093	0.0135	0.0040	0.0072
MLLM	0.0140	0.0345	0.0202	0.0364	0.0053	0.0089
AB-Rec	**0.0276**	**0.0509**	**0.0293**	**0.0478**	**0.0175**	**0.0206**

4.1.4 Implementation Details. During the fine-tuning phase, we utilized Qwen2vl-2b[2] as the backbone model and conducted fine-tuning for 5 epochs on four datasets with a batch size of 128, employing 8 A100 GPUs. For AB-Rec, We conducted offline model evaluation on an RTX 4090 GPU using TensorFlow 2.9.0, with Adam as the optimizer. The hyper-parameters α, β and η were searched within $\{0.1, 0.2, ..., 1.0\}$, while batch size and learning rate were tuned over $\{256, 512, 1024, 2048\}$ and $\{1e-3, 1e-4, 1e-5\}$, respectively. The best model was selected based on the minimum validation loss, and early stopping was applied with a patience of 5 to prevent over-fitting.

4.2 Performance Comparison

4.2.1 ***Multimodal representation evaluation.*** We conduct zero-shot recommendation experiments [26] on the Amazon dataset to demonstrate the effectiveness of the multimodal representation generation method proposed in AB-Rec. Following AlignRec [26], for the zero-shot recommendation of user u, the most recently interacted item is treated as the target item, while the remaining items form the historical behavior sequence. The user's multimodal feature representation is generated by averaging the multimodal features of the items in the historical sequence. Then, the similarity between this representation and the multimodal features of each item in the candidate set is computed to check if the target item appears among the top-K most similar items.

We compared two approaches: one based on traditional methods [9, 27], where separate visual encoders (e.g., CNN) and text encoders (e.g., sequence Transformer) are used to extract visual and textual features, and another based on an unfinetuned MLLM. We used *R*@20 and *R*@50 as the evaluation metric. For a fair comparison, we concatenated the extracted visual and textual features to form a unified multimodal feature for the traditional methods. The experimental results are shown in Table 2. The findings reveal several key insights: (i) Compared to the traditional approach that separately uses CNN and Transformer to extract visual and textual features, AB-Rec employs a visual-text alignment strategy specifically designed for recommendation scenarios, demonstrating more effective utilization of multimodal information. (ii) Our fine-tuning strategy, tailored for recommendation tasks, fully leverages the pretrained MLLM's powerful capabilities in modality alignment and understanding, leading to higher-quality representations.

4.2.2 ***Compared to different backbones.*** Table 3 presents the AUC performance of our proposed AB-Rec framework evaluated on different backbone architectures We conducted experiments on four

[2]https://github.com/QwenLM/Qwen2-VL

Table 3: Performance comparison of AB-Rec and ID, MM, ID+MM models across different backbone architectures, measured by AUC.

Backbone	Model	Baby	Sports	Electronics	Industrial
MLP	ID	0.6748	0.6985	0.7180	0.8125
	MM	0.6222	0.6263	0.6572	0.7616
	ID+MM	0.6736	0.7084	0.7240	0.8267
	AB-Rec	**0.6871**	**0.7114**	**0.7284**	**0.8300**
DCN	ID	0.6704	0.6853	0.7181	0.8126
	MM	0.6208	0.6240	0.6513	0.7563
	ID+MM	0.6721	0.6816	0.7101	0.8247
	AB-Rec	**0.6839**	**0.6937**	**0.7203**	**0.8277**
Fibinet	ID	0.6790	0.7036	0.7206	0.8126
	MM	0.6150	0.6218	0.6542	0.7473
	ID+MM	0.6636	0.7109	0.7279	0.8232
	AB-Rec	**0.6805**	**0.7152**	**0.7305**	**0.8269**
AutoInt	ID	0.6745	0.6945	0.7099	0.8110
	MM	0.5830	0.6012	0.6373	0.7388
	ID+MM	0.6695	0.6830	0.7186	0.8237
	AB-Rec	**0.6835**	**0.7034**	**0.7216**	**0.8248**

Table 4: Comparison of AUC and LogLoss between AB-Rec and other multimodal recommendation methods.

Dataset	Metrics	VBPR	FREEDOM	BM3	AlignRec	AB-Rec
Baby	AUC	0.6729	0.6802	0.6715	0.6832	**0.6871**
	Logloss	0.6425	0.6382	0.6392	0.6358	**0.6335**
Sports	AUC	0.6985	0.7023	0.6932	0.7101	**0.7152**
	Logloss	0.6041	0.5978	0.6023	0.5925	**0.5898**
Electronics	AUC	0.7158	0.7221	0.7119	0.7274	**0.7305**
	Logloss	0.6012	0.5951	0.6053	0.598	**0.5935**
Industrial	AUC	0.8120	0.8284	0.8052	0.8253	**0.8300**
	Logloss	0.2513	0.2458	0.2601	0.2427	**0.2399**

model configurations: (1) ID, which uses only ID features for training; (2) MM, which relies solely on multimodal features; (3) ID+MM, which combines both ID and multimodal features but excludes the distribution alignment and Gradient Modulation modules present in AB-Rec; (4) AB-Rec.

The results reveal the following insights: (i) AB-Rec consistently achieves the best performance across all architectures, demonstrating its robustness and adaptability to various ID&MM recommendation frameworks, independent of the backbone choice. (ii) On smaller datasets (e.g., Baby, Sports), the performance of ID+MM models is inferior to that of ID models. This may be due to more significant distribution conflicts between the two feature types in smaller datasets, making their integration more challenging.

4.2.3 ***Compared to different baselines.*** Table 4 presents the AUC and LogLoss performance of our proposed AB-Rec framework compared to other multimodal recommendation methods on four datasets. From the experimental results, we find the following observations: (i) AB-Rec achieves superior performance over existing multimodal recommendation baselines across all four datasets. It is noteworthy that our framework primarily addresses the distributional conflicts between ID features and multimodal features, as well as the imbalance in their convergence rates, while employing relatively simple neural network architectures. Consequently, the observed performance improvements are largely attributed to our fine-tuning strategies for MLLM, the design of the distribution alignment loss function, and the gradient modulation mechanism. These strategies collectively validate the efficacy of our proposed framework. (ii) Unlike most previous works that employ separate visual and text encoders to extract visual and textual features, AB-Rec utilizes a fine-tuned MLLM to directly extract multimodal features, significantly improving the quality of embeddings. For instance, compared to VBPR, AB-Rec achieves at least a 1.5% improvement in AUC across all datasets. This enhancement is attributed to the powerful pretraining capabilities of MLLMs, which not only effectively align visual and textual representations but also seamlessly fuse information from both modalities.

Table 5: Ablation Studies on AB-Rec.

Model	Baby	Sports	Electronics	Industrial
AB-Rec	**0.6871**	**0.7152**	**0.7305**	**0.8300**
-DA	0.6804	0.7103	0.7288	0.8257
-CR	0.6842	0.7136	0.7290	0.8284
-GM	0.6741	0.7085	0.7213	0.8209

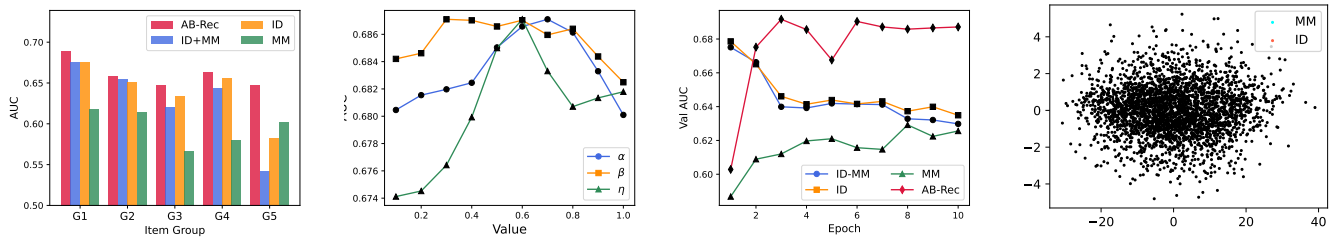
4.3 Ablation Study

In this part, we conduct ablation studies to investigate the contribution of each designed module in AB-Rec. To this end, we compare AB-Rec with three variants: (i) **w/o Distribution Alignment(-DA)**, where the distribution alignment loss function $\mathcal{L}_{align}$ is removed during training; (ii) **w/o Cosine Regularization(-CR)**, where the cosine regularization term, which prevents representation collapse, is removed from the Distribution Alignment module; and (iii) **w/o Gradient Modulation(-GM)**, where the Gradient Modulation module is omitted during the update process, thereby ignoring the issue of inconsistent convergence rates between the ID side and the multimodal side. These variants allow us to systematically evaluate the impact of each component on the overall performance of AB-Rec.

Table 5 presents the experimental results, which illustrate two key observations: (i) All designed modules are crucial for AB-Rec. The removal of any module leads to a noticeable decline in AUC performance. (ii) Gradient Modulation has the most significant impact on performance. This is because the discrepancy in convergence rates between the ID side and the multimodal side directly affects the contribution of multimodal features to the recommendation results. By dynamically balancing the convergence rates of both sides through Gradient Modulation, we ensure better compatibility between multimodal features and ID features, thereby enhancing the overall effectiveness of the model.

4.4 In-depth Analysis

4.4.1 ***Performance of Items with Different Frequencies.*** Figure 3(a) presents the results of a bucket experiment based on item frequency (where G1 represents the highest frequency and G5 the lowest), with AUC as the evaluation metric. The following observations can be made: (i) For the G5 bucket, which can be interpreted as addressing the cold-start problem, ID-based models struggle to effectively characterize item profiles due to the limited number



(a) Bucket Testing (b) Hyperparameter Analysis (c) Convergence Analysis (d) Representation Visualization

Figure 3: Comprehensive Analysis of using MLP as recommendation model on Amazon dataset: (a) AUC Results of Four type of Models Based on Item Interaction Frequency Buckets; (b) Sensitivity of AB-Rec's AUC Performance to Hyperparameters; (c) AUC Performance and Convergence Trends of Four Type Models Across Epochs; (d) Visualization of ID and Multimodal Top-Representations in AB-Rec (Input Vectors to the Classifier).

of interactions. In contrast, multimodal models excel in this scenario by leveraging visual and textual content to better represent items. (ii) Simply combining ID-based models with multimodal models does not always yield positive gains. This is primarily due to conflicts in feature distributions and significant disparities in convergence rates between the two types of models. (iii) Across all buckets, AB-Rec achieves superior performance, particularly in addressing the cold-start problem where interaction data is sparse. By employing distribution alignment and the Gradient Modulation strategy, AB-Rec effectively exploits multimodal features to enhance performance.

4.4.2 ***Hyper-parameter Study.*** In Figure 3(b), we analyze the hyperparameters α and β from Equation 8, as well as η from Equation 18. As illustrated in the figure, the optimal parameter configuration is determined to be $\alpha = 0.25, \beta = 0.7$, and $\eta = 0.6$. Further analysis reveals that variations in η have a more pronounced impact on the AUC compared to α and β. This is primarily because the value of η directly influences the modulation coefficient in Equation 18, whereas α and β are responsible for adjusting the weight of the distribution alignment strategy (i.e., $\mathcal{L}_{align}$). Consistent with the ablation studies in Section 4.3, these findings indicate that the Gradient Modulation module exerts a more significant influence on the AUC than the distribution alignment strategy.

4.4.3 ***Convergence Analysis.*** In Figure 3(c), We analyze the convergence imbalance between multimodal (MM) and ID representations and validate the effectiveness of AB-Rec in addressing this issue. As illustrated in the figure, significant convergence disparities exist between the ID and MM components, regardless of whether a single ID/MM architecture or a combined ID & MM architecture is employed. Specifically: (i) The performance of the MM architecture typically improves continuously as the number of training epochs increases, whereas the optimal performance of the ID architecture is usually achieved within the first epoch. (ii) The ID+MM architecture exhibits convergence behavior similar to that of the ID architecture, indicating that the ID side dominates during training, and multimodal features are underutilized. Our proposed AB-Rec addresses this issue through the Gradient Modulation module, which dynamically balances the convergence rates between the ID and MM sides, thereby enhancing the compatibility between ID and multimodal representations.

Table 6: Online A/B test of different methods.

Main	watch-time	app usage	long view	short view
+MM	+0.022%	+0.008%	+0.132%	-0.160%
AB-Rec	+0.175%	+0.124%	+0.678%	-0.235%

4.4.4 ***Representation Visualization.*** In Figure 3(d), we we present a visualization of the top representations of ID and multimodal features (i.e., the input vectors to the classifier) for the AB-Rec model with a DCN network backbone after training. Compared to Figure 1(b), which illustrates training without the Gradient Modulation and Distribution Alignment strategies, it is evident that AB-Rec effectively mitigates the distributional conflict between ID and multimodal representations.

4.5 Online A/B test

Table 6 summarizes the results of our online A/B test on a short-video platform, comparing the simple integration of multimodal features (+MM) with our proposed AB-Rec framework. Although incorporating multimodal features into the baseline model does yield some performance gains across a few key metrics, the overall improvement remains limited. This indicates that merely adding multimodal information does not fully capitalize on its potential and highlights the need for further analysis and optimization. By contrast, AB-Rec significantly enhances engagement and usage metrics, demonstrating that a more systematic alignment and balancing strategy is crucial for maximizing the benefits of multimodal information in an online production environment.

5 Conclusion

In this paper, we introduced AB-Rec, a framework designed to align and balance ID and multimodal representations for large-scale recommendation. By leveraging a pretrained multimodal large language model (MLLM), AB-Rec generates unified item representations, addresses the distribution discrepancy between ID and

multimodal features through in-batch Wasserstein distance minimization, and mitigates the optimization imbalance with a gradient modulation method. Experiments on four real-world datasets and an A/B test on an online video platform validate the effectiveness and scalability of our approach.

Acknowledgement

This research is partially sponsored by funding from Yonghua Foundation.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[3] 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.

[4] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.

[5] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, Pietro Piazzolla, and Massimo Quadrana. 2016. Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics* 5 (2016), 99–113.

[6] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Hao Ding, Yifei Ma, Anoop Deoras, Yuyang Wang, and Hao Wang. 2021. Zero-shot recommender systems. *arXiv preprint arXiv:2105.08318* (2021).

[8] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics* 28, 2 (2000), 337–407.

[9] 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.

[10] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[11] Jin Huang and Charles X Ling. 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering* 17, 3 (2005), 299–310.

[12] Tongwen Huang, Zhiqi Zhang, and Junlin Zhang. 2019. FiBiNET: combining feature importance and bilinear feature interaction for click-through rate prediction. In *Proceedings of the 13th ACM conference on recommender systems*. 169–177.

[13] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[14] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[15] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 426–434.

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

[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[18] Joonseok Lee and Sami Abu-El-Haija. 2017. Large-scale content-only video recommendation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 987–995.

[19] Fan Li, Xu Si, Shisong Tang, Dingmin Wang, Kunyan Han, Bing Han, Guorui Zhou, Yang Song, and Hechang Chen. 2024. Contextual Distillation Model for Diversified Recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5307–5316.

[20] 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. In *International conference on machine learning*. PMLR, 19730–19742.

[21] 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.

[22] 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.

[23] Meng Li and Haochen Sui. 2025. Causal Recommendation via Machine Unlearning with a Few Unbiased Data. In *AAAI 2025 Workshop on Artificial Intelligence with Causal Techniques*. https://openreview.net/forum?id=7btnaN4Std

[24] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 7, 1 (2003), 76–80.

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[26] Yifan Liu, Kangning Zhang, Xiangyuan Ren, Yanhua Huang, Jiarui Jin, Yingjie Qin, Ruilong Su, Ruiwen Xu, Yong Yu, and Weinan Zhang. 2024. AlignRec: Aligning and Training in Multimodal Recommendations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1503–1512.

[27] 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.

[28] Kaixiang Mo, Bo Liu, Lei Xiao, Yong Li, and Jie Jiang. 2015. Image feature learning for cold start problem in display advertising. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[29] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1933–1942.

[30] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824* (2023).

[31] David MW Powers. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020).

[32] 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.

[33] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*. IEEE, 995–1000.

[34] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[35] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Commun. ACM* 40, 3 (1997), 56–58.

[36] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40 (2000), 99–121.

[37] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.

[38] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 253–260.

[39] Anima Singh, Trung Vu, Nikhil Mehta, Raghunandan Keshavan, Maheswaran Sathiamoorthy, Yilin Zheng, Lichan Hong, Lukasz Heldt, Li Wei, Devansh Tandon, et al. 2024. Better generalization with semantic ids: A case study in ranking for recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1039–1044.

[40] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. 2019. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1161–1170.

[41] Shisong Tang, Qing Li, Xiaoteng Ma, Ci Gao, Dingmin Wang, Yong Jiang, Qian Ma, Aoyang Zhang, and Hechang Chen. 2022. Knowledge-based temporal fusion network for interpretable online video popularity prediction. In *Proceedings of the ACM Web Conference 2022*. 2879–2887.

[42] Shisong Tang, Qing Li, Dingmin Wang, Ci Gao, Wentao Xiao, Dan Zhao, Yong Jiang, Qian Ma, and Aoyang Zhang. 2023. Counterfactual video recommendation for duration debiasing. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4894–4903.

[43] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican,

et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[44] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).

[45] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*. 1–7.

[46] 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.

[47] 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.

[48] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2576–2584.

[49] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with multi-head self-attention. 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)*. 6389–6394.

[50] Chuhan Wu, Fangzhao Wu, Tao Qi, Chao Zhang, Yongfeng Huang, and Tong Xu. 2022. Mm-rec: Visiolinguistic model empowered multimodal news recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 2560–2564.

[51] Yuntian Wu, Yuntian Yang, Jiabao Sean Xiao, Chuan Zhou, Haochen Sui, and Haoxuan Li. 2024. Invariant Spatiotemporal Representation Learning for Cross-patient Seizure Classification. In *The First Workshop on NeuroAI @ NeurIPS2024*. https://openreview.net/forum?id=Ex6wAivo7G

[52] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* (2023).

[53] 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.

[54] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to go next for recommender systems? id- vs. modality-based recommender models revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2639–2649.

[55] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601* (2024).

[56] Kangning Zhang, Yingjie Qin, Ruilong Su, Yifan Liu, Jiarui Jin, Weinan Zhang, and Yong Yu. 2024. DRepMRec: A Dual Representation Learning Framework for Multimodal Recommendation. *arXiv preprint arXiv:2404.11119* (2024).

[57] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1059–1068.

[58] 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.

[59] 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.

[60] Yongchun Zhu, Kaikai Ge, Fuzhen Zhuang, Ruobing Xie, Dongbo Xi, Xu Zhang, Leyu Lin, and Qing He. 2021. Transfer-meta framework for cross-domain recommendation to cold-start users. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 1813–1817.

[61] Yongchun Zhu, Ruobing Xie, Fuzhen Zhuang, Kaikai Ge, Ying Sun, Xu Zhang, Leyu Lin, and Juan Cao. 2021. Learning to warm up cold item embeddings for cold-start recommendation with meta scaling and shifting networks. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1167–1176.

[62] Ziwei Zhu, Jingu Kim, Trung Nguyen, Aish Fenton, and James Caverlee. 2021. Fairness among new items in cold start recommender systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 767–776.