Multimodal Book Recommendation for Vietnamese E-commerce: Dataset, Evaluation, and Deployment

Binh Le Xuan[1,2] **Lam Thai Minh**[1,2] **Phat Ma Kim**[1,2]
[1]Faculty of Information Science and Engineering, University of Information Technology
Ho Chi Minh City, Vietnam
[2]Vietnam National University - Ho Chi Minh City, Vietnam
{22520131, 22520745, 22521071}@gm.uit.edu.vn

Abstract

The rapid growth of electronic commerce has substantially reshaped book purchasing behaviors, in which user preferences are increasingly influenced by diverse and dynamic factors. In this work, we present an empirical study on multimodal book recommendation for the Vietnamese online book market. First, we construct a Vietnamese book dataset by collecting data from a major domestic e-commerce platform and integrating user–item interaction data with rich multimodal features. Second, we conduct extensive experiments to evaluate and compare state-of-the-art multimodal recommendation models with traditional recommendation approaches. Experimental results demonstrate that multimodal models significantly outperform conventional methods. Finally, we deploy the best-performing model within a big data–oriented system architecture to enable real-time recommendation. The dataset and source code are publicly available at here.

1 Introduction

The rapid expansion of e-commerce has fundamentally transformed book purchasing behavior: readers today make decisions based on various factors such as cover images, content summaries, and online reviews (Zhang et al., 2019). While traditional recommender systems have demonstrated strong performance on user–item interaction data, they often struggle with sparse data and fail to fully exploit semantic information embedded in unstructured content (Su and Khoshgoftaar, 2009). Existing multimodal recommendation studies typically treat visual or textual data as isolated auxiliary features, lacking in-depth evaluations of the actual benefits of integrating these modalities (Deldjoo et al., 2020). Notably, the absence of a standardized Vietnamese book dataset that includes both textual and visual information presents a significant barrier to domestic research in this field.

This study aims to address these challenges through three main contributions:

- **Dataset construction:** We introduce a Vietnamese book dataset that integrates comprehensive multimodal features.
- **Empirical evaluation:** We implement and compare state-of-the-art recommendation models to assess the impact of multimodal integration relative to unimodal baselines.
- **System deployment:** We build a complete recommendation system optimized for scalability and integration with big data technologies, tailored to the practical demands of e-commerce platforms.

2 Related Work

Book Recommendation and Datasets: Standard benchmarks for book recommendation have progressed from the sparse Book-Crossing dataset (Ziegler et al., 2005) to large-scale resources like Amazon Books (McAuley et al., 2015) and Goodreads (Wan and McAuley, 2018). While these datasets provide extensive user–item interaction histories, they primarily focus on structured metadata. Prior works have adopted collaborative filtering, content-based filtering, and hybrid approaches to book recommendation, as explored in (Devika et al., 2021; Mathew et al., 2016; Rajpurkar et al., 2015; Kurmashov et al., 2015). Recently, the shift toward multimodal integration is exemplified by (Spillo et al., 2025), who augmented the DBbook dataset with textual and visual signals. However, these resources are almost exclusively English-based. To date, no standardized, large-scale multimodal dataset has been specifically curated for Vietnamese books—posing a major barrier to domestic research in this area.

Core Recommendation Paradigms: The literature broadly classifies recommendation strategies into three pillars: (1) Collaborative Filtering (CF) and its neural variants (NCF), which model latent user–item interactions (Su and Khoshgoftaar, 2009; He et al., 2017); (2) Content-Based Filtering (CBF), which leverages similarity in item metadata (Lops et al., 2010); and (3) Knowledge-aware Recommendation, which incorporates Knowledge Graphs to mitigate data sparsity issues (Guo et al., 2020). While effective, these paradigms often treat unstructured data (e.g., user reviews, book summaries) as secondary or auxiliary input, limiting their potential in domains where such data is rich and informative.

Multimodal Fusion and Domain Utility: Multimodal Recommender Systems (MRS) enhance conventional interaction modeling by incorporating raw textual and visual features (He and McAuley, 2015; Deldjoo et al., 2020). However, a critical gap remains in evaluating the relative contribution of each modality to recommendation quality within specific domains (Liu et al., 2024). Many models overlook that the predictive value of a modality—e.g., a book's content summary versus its cover image—can vary significantly based on the application context or dataset characteristics. Our work addresses this limitation by empirically analyzing the effectiveness of Vietnamese-specific visual and textual features within a unified multimodal framework.

3 Dataset

3.1 Data Collection

We constructed a new dataset by collecting book-related data from Tiki.vn[1], one of Vietnam's leading e-commerce platforms specializing in books. The data extraction process combined two techniques: we used BeautifulSoup[2] to quickly scrape metadata and user reviews, and Selenium[3] to capture dynamically loaded textual descriptions. The raw data includes book titles, cover images, textual descriptions, prices, and user review histories.

3.2 Preprocessing and Data Augmentation

Figure 1 illustrates our preprocessing and augmentation pipeline. The raw data underwent rigorous cleaning, including Unicode normalization, Vietnamese spelling correction, and removal of irrelevant noise such as hashtags, HTML tags, and emojis.

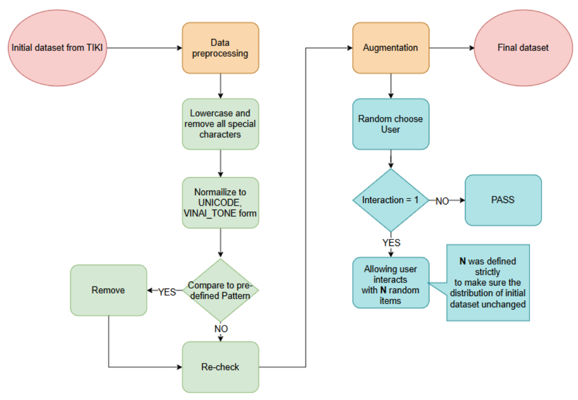

Figure 1: Preprocessing and data augmentation strategy.

To address sparsity and maintain data diversity during k-core filtering prior to model training, we performed targeted data augmentation: for items with fewer than five reviews, we generated synthetic reviews and assigned them to hypothetical user IDs under two constraints—(1) each pseudo-user had exactly one original interaction, and (2) the augmentation did not significantly alter the original rating distribution. This strategy preserved long-tail items while maintaining the structural density of the interaction matrix.

3.3 Data Analysis and Statistics

After cleaning, the dataset comprises 1,115 books across four main genres (Figure 2) and 26,050 user ratings. A full list of attributes is provided in Appendix A.

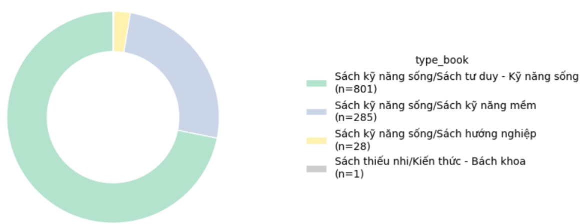

Figure 2: Distribution of book genres.

We conducted several analyses to understand the underlying distributions in the data, summarized in Figure 4.

The distribution of average ratings reveals a strong positive skew, with a significant concentration between 4.0 and 5.0. This suggests a generally favorable user sentiment, implying that recommendation models must be capable of fine-grained dis-

[1]TIKI
[2]beautifulsoup4
[3]Selenium

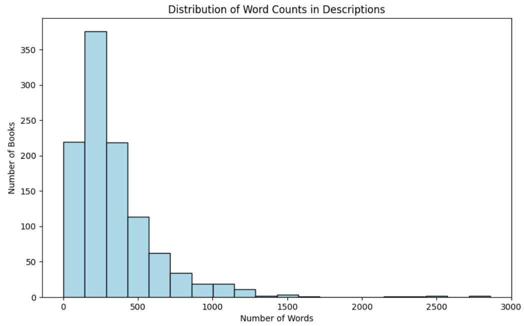

Figure 3: Distribution of book abstract lengths.

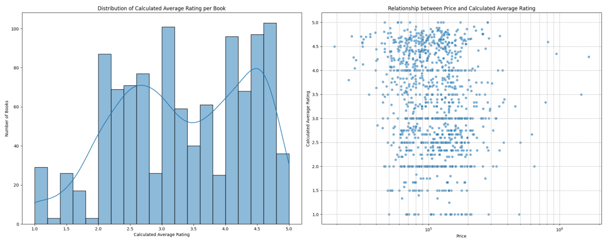

Figure 4: Distribution of average ratings (Left). Relationship between price and rating (Right).

crimination among highly rated items.

Additionally, we examined the relationship between book price (log-scaled) and average rating, and found no clear correlation. This suggests that price is not a major determinant of user satisfaction in this dataset, thereby reinforcing our decision to focus on content features, cover images, and user preferences rather than pricing.

To further characterize the textual content of book descriptions, we plotted the distribution of summary lengths as a histogram (Figure 3). The word count distribution indicates high variability: while most descriptions fall in the 100–400 word range, the peak occurs between 150 and 300 words. A few outliers exceed 1,000 words.

4 Experimental Setup

To ensure reliable model performance, we applied the following data filtering criteria:

- **Completeness filter:** We excluded all book records that were missing either cover images or textual descriptions.
- **k-core filtering:** We applied a $k = 3$ core filter to ensure that each remaining user and item had at least three associated interactions.

The train–test split strategy varied depending on the type of recommendation approach.

For collaborative filtering (CF) and content-based filtering (CBF) methods, we adopted the leave-one-out evaluation protocol. Specifically, user–item interaction histories were sorted chronologically, and the last interacted item for each user was held out as the test instance, while all preceding interactions were used for training.

For multimodal approaches, we extended the leave-one-out strategy to include an additional development (validation) set to facilitate hyperparameter tuning for deep models. In this setup, we extracted the last two items from each user's interaction history: the most recent item was assigned to the test set, the penultimate item to the development set, and the remaining interactions to the training set. This partitioning scheme allows for both objective model tuning and fair evaluation on unseen data.

To extract feature representations for each modality, we employed pretrained encoders listed in Table 1.

Table 1: Pretrained encoders used for text and image modalities

Text Encoder	ViSoBERT (Nguyen et al., 2023)
Image Encoders	ViT (Dosovitskiy et al., 2021), ResNet-152 (He et al., 2015)

We adopted MMRec (Zhou, 2023) as the experimental framework for training and evaluating multimodal recommender models. The overall pipeline included data preprocessing and packaging, configuration setup, model training, and performance evaluation.

5 Evaluation Analysis

5.1 Metrics

To assess the performance of the recommendation models, we employ four standard metrics at $k \in \{5, 10\}$. Let R_u be the set of top-k recommended items for user u, and T_u be the set of ground-truth items (items the user has actually interacted with).

Precision@k and Recall@k Precision measures the proportion of recommended items that are relevant, while Recall measures the proportion of relevant items that were successfully recommended:

$$Precision@k = \frac{|R_u \cap T_u|}{k} \quad (1)$$

$$Recall@k = \frac{|R_u \cap T_u|}{|T_u|} \quad (2)$$

Mean Average Precision (mAP@k) mAP accounts for the rank of relevant items. We first calculate the Average Precision (AP) for a user, where $rel(i)$ is a binary indicator of relevance for the item at rank i:

$$AP@k = \frac{1}{\min(|T_u|, k)} \sum_{i=1}^{k} (Precision@i \times rel(i)) \tag{3}$$

The mAP is then the average of $AP@k$ across all users U in the test set.

Normalized Discounted Cumulative Gain (NDCG@k) NDCG evaluates ranking quality by penalizing relevant items placed lower in the list using a logarithmic discount. The Discounted Cumulative Gain (DCG) is defined as:

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel(i)} - 1}{\log_2(i+1)} \tag{4}$$

NDCG is obtained by normalizing DCG by the Ideal DCG (IDCG), which is the maximum possible DCG achieved by a perfect ranking:

$$NDCG@k = \frac{DCG@k}{IDCG@k} \tag{5}$$

5.2 Experimental Results

Tables 2 and 3 summarize the performance of all evaluated recommendation models across the defined evaluation metrics.

All multimodal architectures—including VBPR (He and McAuley, 2015), LATTICE (Chiappa et al., 2023), FREEDOM (Zhou and Shen, 2023), MMGCN (Wei et al., 2019), and SLMRec (Tao et al., 2022)—significantly outperform the baseline BPR model (Rendle et al., 2012), which relies solely on implicit user–item interactions. These results underscore the value of incorporating visual and textual signals to alleviate data sparsity and enhance recommendation accuracy. Among the evaluated models, MMGCN consistently achieves the best performance on most metrics. Specifically, MMGCN with textual input from category information (Text-Type) reaches a Recall@5 of 0.0311 and NDCG@10 of 0.0243—approximately 3× higher than the BPR baseline. This highlights the strength of graph-based architectures in propagating multimodal signals over user–item interaction graphs.

Further analysis reveals that the choice of encoder plays a critical role in overall performance, particularly with respect to the **Recall@10** metric. The following insights were derived:

- **Textual Encoder:** ViSoBERT demonstrates strong and stable performance for Vietnamese text. Notably, category-based descriptions (Text-Type) outperform full product descriptions (Text-Desc), likely due to their conciseness and discriminative nature. These short, high-signal inputs help models better identify salient content features without distraction from irrelevant details. In multimodal fusion scenarios, combining textual category information with visual features produces a synergistic effect. This is especially evident in SLMRec, where the inclusion of Text-Type consistently yields better results than Text-Desc, regardless of the visual encoder used.

- **Visual Encoder:** The effectiveness of visual features is closely tied to the model architecture and its fusion strategy. For structure-aware models like LATTICE and FREEDOM, ViT provides superior global and semantic representations, leading to improved retrieval performance. Conversely, for factorization-based or graph-based models such as VBPR and MMGCN, ResNet-152 tends to perform better—suggesting that CNN-derived features offer greater stability and representational richness in these settings than the attention-based ViT features.

To objectively benchmark the effectiveness of multimodal methods, we also compare them against traditional baselines, including user-based collaborative filtering (User-CF) and content-based filtering using either description text or category information. In general, traditional baselines perform poorly: User-based (Cosine) achieves Recall@5 = 0.0057 and NDCG@10 = 0.0065, while Typebook-based (Cosine) performs slightly better with Recall@5 = 0.0141 and NDCG@10 = 0.0088. In contrast, the multimodal MMGCN (Text-Type) model reaches Recall@5 = 0.0311 and NDCG@10 = 0.0243, demonstrating substantial improvements over both traditional and pure interaction-based methods.

Compared to the best traditional baseline, MMGCN achieves approximately a 2.2× improvement in Recall@5 and a 2.8× gain in NDCG@10. When compared to the weakest baseline (User-based), the improvements rise to 5.5× and 3.7×, respectively.

Finally, while combining both text and image modalities (Text + Image) sometimes enhances per-

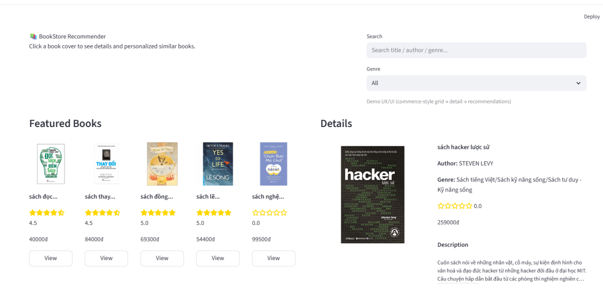

Figure 5: User interface of the multimodal book recommendation system.

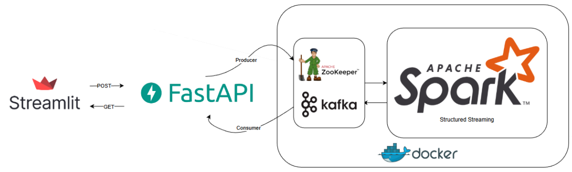

Figure 6: System architecture: real-time recommendation pipeline integrating multimodal model.

formance, it does not universally outperform strong unimodal baselines—particularly those using high-quality category text—on all metrics.

6 System Deployment

Following the experimental phase, we selected the best-performing model for deployment within a recommendation system.

The system was designed under the assumption of a Big Data infrastructure, as illustrated in Figures 5 and 6. The architecture employs Streamlit for the user interface, which communicates asynchronously with a FastAPI[4] backend. Data flow is managed by Apache Kafka[5], enabling high throughput and fault tolerance. The core processing layer is powered by Apache Spark[6] (Structured Streaming), which integrates a pretrained multimodal recommendation model to support real-time inference. For portability and deployment across various environments, the entire system is containerized using Docker[7], ensuring consistency and seamless deployment.

The user interface of the deployed system is shown in Figure 5.

[4] FastAPI
[5] Kafka
[6] Spark
[7] Docker

7 Limitations

Despite the promising results, the dataset used in this study remains relatively limited in size compared to real-world applications. This leads to a sparse user–item interaction matrix, which can hinder the ability of interaction-based models to learn stable and generalizable representations. In such sparse settings, the model may become biased toward popular items or frequent users, reducing recommendation accuracy for users with limited histories and long-tail items.

Furthermore, the quality and completeness of content data are not uniform. Book descriptions are sometimes short, uninformative, or overly promotional. This lowers the effectiveness of content-based signals and can introduce noise in multimodal fusion.

In addition, the current multimodal setup primarily focuses on cover images and textual descriptions. Other semantically rich attributes—such as publisher, publication year, language, detailed topic, content excerpts, or structured author metadata—are not yet utilized. The absence of these fields limits the model's ability to distinguish between books with similar covers or surface-level descriptions but different content.

8 Conclusion

Multimodal recommendation models effectively address the data sparsity problem inherent in traditional methods such as collaborative filtering and content-based filtering. Although we incorporated powerful pretrained embeddings (e.g., ViSoBERT) into traditional approaches, the dedicated backbones of multimodal architectures leverage multimodal signals more effectively, leading to superior results.

Traditional models are limited by their reliance on local similarity computations between two entities. In contrast, graph-based architectures such as MMGCN and LATTICE allow information to propagate across multiple layers of the user–item graph, enabling the discovery of latent relations that traditional methods cannot capture.

We also successfully deployed the advanced recommendation model within a simulated big data environment, utilizing modern technologies to demonstrate a scalable deployment pipeline suitable for real-world e-commerce platforms.

Finally, we emphasize that the overall performance of multimodal recommender systems is

Table 2: Experimental result of CF and CB methods.

	Pre@5	**Rec@5**	**mAP@5**	**NDCG@5**	**Pre@10**	**Rec@10**	**mAP@10**	**NDCG@10**
User-based (Cosine)	0.0011	0.0057	0.0027	0.0034	0.0016	0.0156	0.0039	0.0065
User-based (Pearson)	0.0006	0.0028	0.0008	0.0013	0.0004	0.0042	0.0011	0.0018
Content-based (Cosine)	0.0011	0.0057	0.0026	0.0034	0.0017	0.0170	0.0041	0.0070
Content-based (Pearson)	0.0011	0.0057	0.0026	0.0034	0.0017	0.0170	0.0041	0.0070
Typebook-based (Cosine)	**0.0028**	0.0141	**0.0052**	**0.0074**	**0.0018**	**0.0183**	**0.0058**	**0.0088**
Typebook-based (Pearson)	0.0019	**0.0099**	0.0022	0.0039	0.0014	0.0141	0.0027	0.0053

Table 3: Experimental result of multimodal models using different encoders.

Model	**Modality**	**Encoder**	**P@5**	**R@5**	**mAP@5**	**N@5**	**P@10**	**R@10**	**mAP@10**	**N@10**
BPR	-	-	0.0023	0.0113	0.0042	0.006	0.0021	0.0212	0.0054	0.0091
VBPR	Text (Desc)	ViSoBERT	0.0011	0.0057	0.0025	0.0033	0.0013	0.0127	0.0034	0.0055
	Text (Type)	ViSoBERT	0.0011	0.005	0.0028	0.0035	0.0014	0.0141	0.0039	0.0062
	Image	ViT (CLS)	0.0006	0.0028	0.001	0.0014	0.0006	0.0057	0.0014	0.0024
	Image	ResNet-152	0.0025	0.0127	0.0078	0.009	0.002	0.0198	0.0085	0.0111
	Text (Type) + Image	ViSoBERT + ResNet-152	0.0025	0.0127	0.0066	0.0081	0.0016	0.0156	0.0071	0.0091
LATTICE	Text (Desc)	ViSoBERT	0.0028	0.0141	0.0073	0.009	0.0033	0.0325	0.0097	0.0148
	Text (Type)	VisoBERT	0.0011	0.0057	0.005	0.0051	0.0024	0.024	0.0075	0.0112
	Image	ViT (CLS)	0.004	0.0198	0.0081	0.011	0.0031	0.0311	0.0097	0.0147
	Image	ResNet-152	0.002	0.0099	0.0023	0.0041	0.0018	0.0184	0.0035	0.0069
	Text (Type) + Image	ViSoBERT + ViT (CLS)	0.0042	0.0212	0.0095	0.0124	0.0033	0.0325	0.0111	0.0161
FREEDOM	Text (Desc)	ViSoBERT	0.0037	0.0184	**0.013**	0.0143	0.0027	0.0269	0.014	0.0169
	Text (Type)	VisoBERT	0.002	0.0099	0.0033	0.0049	0.0021	0.0212	0.0047	0.0084
	Image	ViT (CLS)	0.0023	0.0113	0.0045	0.0061	0.0025	0.0255	0.0064	0.0107
	Image	ResNet-152	0.0006	0.0028	0.0012	0.0016	0.0014	0.0141	0.0027	0.0053
	Text (Desc) + Image	ViSoBERT + ViT (CLS)	0.004	0.0198	0.0117	0.0137	0.0028	0.0283	0.0128	0.0165
MMGCN	Text (Desc)	ViSoBERT	0.0023	0.0113	0.0044	0.0061	0.0051	0.0509	0.0101	0.0193
	Text (Type)	VisoBERT	0.0023	0.0113	0.0041	0.0059	0.0048	0.0481	0.0089	0.0177
	Image	ViT (CLS)	**0.0062**	**0.0311**	0.0108	0.0157	0.0047	0.0467	0.0126	0.0204
	Image	ResNet-152	0.0054	0.0269	0.0101	0.0143	0.0052	0.0523	0.013	0.022
	Text (Desc) + Image	ViSoBERT + ResNet-152	**0.0062**	**0.0311**	0.0112	**0.0161**	**0.0057**	**0.0566**	**0.0145**	**0.0243**
SLMRec	Text (Desc) + Image	ViSoBERT + ViT (CLS)	0.0017	0.0085	0.0027	0.0041	0.003	0.0297	0.0051	0.0106
	Text (Type) + Image	ViSoBERT + ViT (CLS)	0.0048	0.024	0.0065	0.0107	0.0041	0.041	0.0086	0.016
	Text (Desc) + Image	ViSoBERT + ResNet-152	0.0014	0.0071	0.0024	0.0035	0.0033	0.0325	0.0054	0.0114
	Text (Type) + Image	ViSoBERT + ResNet-152	0.0042	0.0212	0.0075	0.0109	0.0035	0.0354	0.0095	0.0156

highly dependent on the choice of encoders. Depending on the specific recommendation task, one may opt for pretrained or fine-tuned encoders tailored to the target domain.

Acknowledgments

This study was conducted as part of the DS300 course project on Recommender Systems at the University of Information Technology – VNU-HCM, under the instruction and supervision of MSc. Huynh Van Tin.

References

Alberto Silvio Chiappa, Alessandro Marin Vargas, Ann Zixiang Huang, and Alexander Mathis. 2023. Latent exploration for reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender systems leveraging multimedia content. 53(5).

PV Devika, K Jyothisree, PV Rahul, S Arjun, and Jayasree Narayanan. 2021. Book recommendation system. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3549–3568.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *Preprint*, arXiv:1512.03385.

Ruining He and Julian McAuley. 2015. Vbpr: Visual bayesian personalized ranking from implicit feedback. *Preprint*, arXiv:1510.01784.

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*, pages 173–182.

Nursultan Kurmashov, Konstantin Latuta, and Abay Nussipbekov. 2015. Online book recommendation system. In *2015 Twelve International Conference on Electronics Computer and Computation (ICECCO)*, pages 1–4. IEEE.

Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2024. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 57(2):1–17.

Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2010. Content-based recommender systems: State of the art and trends. pages 73–105. Springer.

Praveena Mathew, Bincy Kuriakose, and Vinayak Hegde. 2016. Book recommendation system through content based and collaborative filtering method. In *2016 International conference on data mining and advanced computing (SAPIENCE)*, pages 47–52. IEEE.

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*, pages 43–52.

Quoc-Nam Nguyen, Thang Chau Phan, Duc-Vu Nguyen, and Kiet Van Nguyen. 2023. Visobert: A pre-trained language model for vietnamese social media text processing. *Preprint*, arXiv:2310.11166.

Sushama Rajpurkar, Darshana Bhatt, Pooja Malhotra, MSS Rajpurkar, and MDR Bhatt. 2015. Book recommendation system. *International Journal for Innovative Research in Science & Technology*, 1(11):314–316.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.

Giuseppe Spillo, Elio Musacchio, Cataldo Musto, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2025. See the movie, hear the song, read the book: Extending movielens-1m, last. fm-2k, and dbbook with multimodal data. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 847–856.

Xiaoyuan Su and Taghi M. Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Adv. Artif. Intell.*, 2009:421425:1–421425:19.

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*, 25:5107–5116.

Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM conference on recommender systems*, pages 86–94.

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*, pages 1437–1445.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, 52(1):1–38.

Xin Zhou. 2023. Mmrec: Simplifying multimodal recommendation. pages 1–2.

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*, pages 935–943.

Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32.

A Appendix

This appendix provides detailed descriptions of the features used in our dataset, including both book metadata and user interaction records.

Table 4: Book metadata schema

Field	Description
product_id	Unique identifier for the book
product_name	Title of the book
authors	Author names
price	Listed price of the book
seller_id	ID of the seller
seller_type	Type of seller (OFFICIAL_STORE, TRUSTED_STORE)
rating_average	Average rating received
review_count	Number of user reviews
order_count	Number of copies sold
url	Product page URL
image	Cover image URL
description	Textual summary of the book
type_book	Book category
product_index	Sequential index of the book

Table 5: User–book interaction schema

Field	Description
customer_id	Unique identifier for the user
product_id	ID of the interacted book
rating	Rating score given by the user
content	User's review text
customer_index	Sequential index of the user
product_index	Sequential index of the book