Appendix

Computational Complexity Analysis of DHMRec

The computational complexity of DHMRec can be decomposed into three principal components: the multimodal feature disentanglement, the graph representation learning, and the multimodal feature fusion. (1) The multimodal disentanglement module first transforms original features with time complexity $O(Bd_vd_k)$ for visual and $O(Bd_td_k)$ for textual modalities, where B denotes batch size, d_v and d_t represent original multimodal embedding dimensions, and d_k is the aligned dimension. Subsequent both the disentangled encoder and the gated denoising network require $O(Bd_k^2)$. Considering that d_v is typically larger than both d_t and d_k, the overall time complexity of this module is $O(Bd_vd_k)$. (2) In graph representation learning, LightGCN-based convolution on user-item and item-item graphs cost $O(L|\mathcal{E}|d_k)$ and $O(LK|\mathcal{I}|d_k)$ respectively, where L is the number of layers in LightGCN, $|\mathcal{E}|$ denotes the number of interaction edges, $|\mathcal{I}|$ denotes the number of items and K represents the number of the neighbors for each item. Note that each item-item graph is constructed before training, and its structure remains static during training. Thus, the time complexity of this part is $O\left((K|\mathcal{I}|+|\mathcal{E}|)Ld_k\right)$. (3) In the multimodal fusion module, the time complexity of the positive and negative attention module is $O(Bd_k^2)$. In addition, we compute the complexity of the loss computation of DHMRec. DHMRec includes $\mathcal{L}_{BPR}$, $\mathcal{L}_{JSD}$, $\mathcal{L}_{orth}$ (all of which cost $O(Bd_k)$), $\mathcal{L}_{KL}$ (which costs $O(B|\mathcal{I}|d_k)$). Therefore, the time complexity of DHMRec is $O\left(Bd_vd_k + (K|\mathcal{I}|+|\mathcal{E}|)\,Ld_k + B|\mathcal{I}|d_k\right)$.

Datasets

We conduct experiments on three widely used Amazon dataset (He and McAuley 2016a): (a) Baby, (b) Sports and Outdoors (Sports, in short), and (c) Clothing, Shoes and Jewelry (Clothing, in short). The statistics of these multimodal datasets are presented in Table 1. The raw data of each dataset are pre-processed with a 5-core setting on both users and items. Following (Zhang et al. 2021), We use 4096-dimensional visual features pre-extracted by convolutional neural networks (He and McAuley 2016a) and 384-dimensional textual features obtained by sentence-transformers (Reimers and Gurevych 2019).

Dataset	# Users	# Items	# Interactions	Sparsity
Baby	19,445	7,050	160,792	99.88%
Sports	35,598	18,357	296,337	99.95%
Clothing	39,387	23,033	278,677	99.97%

Table 1: Statistics of the experimental datasets.

Baselines

General CF Models

- BPR (Rendle et al. 2012): This method leverages matrix factorization with the integration of Bayesian Personalized Ranking (BPR) loss to enhance recommendation performance.
- LightGCN (He et al. 2020): This method simplifies the design of GCN by preserving only the essential neighborhood aggregation mechanism.
- LayerGCN (Zhou et al. 2023a): This method improves layer-wise representation updates and enhances recommendation performance by pruning graph edges using degree-sensitive probabilities to reduce noise from irrelevant nodes.

Multimodal models

- VBPR (He and McAuley 2016b): This method builds on the BPR method by combining the visual features and ID embeddings of the item to mitigate the cold start phenomenon.
- MMGCN (Wei et al. 2019): This method leverages message passing on modality-specific user-item bipartite graphs to capture user preferences by enriching node representations with neighbor features and graph structures.
- GRCN (Wei et al. 2020): This method adaptively refines the interaction graph during training by identifying and softly pruning potential false-positive edges, enabling more accurate user preference extraction.
- DualGNN (Wang et al. 2021): This method utilizes user-item and user co-occurrence graphs to collaboratively learn personalized fusion patterns.
- LATTICE (Zhang et al. 2021): This method learns modality-specific item-item structures, aggregates them into latent item graphs, and applies graph convolutions to enhance item representations.
- MICRO (Zhang et al. 2022): This method learns latent item-item structures for each modality and employs a contrastive fusion framework to capture both shared and modality-specific information.
- BM3 (Zhou et al. 2023b): This method boosts user-item representation learning by generating contrastive views through dropout and optimizing multi-modal objectives to reconstruct interactions and align modality features.
- FREEDOM (Zhou and Shen 2023): This method freezes the item-item graph and denoises the user-item interaction graph using degree-sensitive edge pruning to improve recommendation performance.
- LGMRec (Guo et al. 2024): This method learn collaborative-related and modality-related embeddings through local graph learning and capture robust global user interests using a hypergraph embedding module.
- DA-MRS (Xv et al. 2024): This method addresses triple denoising via cross-modal consistent item graphs, denoised BPR loss, and dual alignment strategies for robust multimodal recommendations.
- DiffMM (Jiang et al. 2024): This method integrates modality-aware graph diffusion with cross-modal contrastive learning to generate denoised user-item graphs, enhancing multimodal representation learning and addressing noise in modality-interaction alignment.

- TMLP (Huang et al. 2025): This method replaces GCNs with multi-layer perceptrons to model item-item relationships, enhancing performance through topological pruning, denoising, and integrating higher-order modality correlations.
- CMDL (Lin et al. 2025): This method disentangles multimodal representations into modality-invariant and modality-specific components, which aligns partially with the disentangled strategy proposed in our work. However, it adopts a fundamentally different approach by leveraging contrastive regularization to further enhance recommendation performance.

References

Guo, Z.; Li, J.; Li, G.; Wang, C.; Shi, S.; and Ruan, B. 2024. LGMRec: Local and Global Graph Learning for Multimodal Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8454–8462.

He, R.; and McAuley, J. 2016a. 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.

He, R.; and McAuley, J. 2016b. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 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.

Huang, J.; Qin, J.; Yu, Y.; and Zhang, W. 2025. Beyond Graph Convolution: Multimodal Recommendation with Topology-aware MLPs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11808–11816.

Jiang, Y.; Xia, L.; Wei, W.; Luo, D.; Lin, K.; and Huang, C. 2024. Diffmm: Multi-modal diffusion model for recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7591–7599.

Lin, X.; Liu, R.; Cao, Y.; Zou, L.; Li, Q.; Wu, Y.; Liu, Y.; Yin, D.; and Xu, G. 2025. Contrastive Modality-Disentangled Learning for Multimodal Recommendation. *ACM Transactions on Information Systems*.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.

Wang, Q.; Wei, Y.; Yin, J.; Wu, J.; Song, X.; and Nie, L. 2021. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25: 1074–1084.

Wei, Y.; Wang, X.; Nie, L.; He, X.; and Chua, T.-S. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*, 3541–3549.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 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.

Xv, G.; Li, X.; Xie, R.; Lin, C.; Liu, C.; Xia, F.; Kang, Z.; and Lin, L. 2024. Improving multi-modal recommender systems by denoising and aligning multi-modal content and user feedback. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3645–3656.

Zhang, J.; Zhu, Y.; Liu, Q.; Wu, S.; Wang, S.; and Wang, L. 2021. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, 3872–3880.

Zhang, J.; Zhu, Y.; Liu, Q.; Zhang, M.; Wu, S.; and Wang, L. 2022. Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(9): 9154–9167.

Zhou, X.; Lin, D.; Liu, Y.; and Miao, C. 2023a. Layer-refined graph convolutional networks for recommendation. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 1247–1259. IEEE.

Zhou, X.; and Shen, Z. 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.

Zhou, X.; Zhou, H.; Liu, Y.; Zeng, Z.; Miao, C.; Wang, P.; You, Y.; and Jiang, F. 2023b. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, 845–854.