Beyond Graph Convolution: Multimodal Recommendation with Topology-aware MLPs

Junjie Huang, Jiarui Qin, Yong Yu, Weinan Zhang*

Shanghai Jiao Tong Univeristy, Shanghai, China
{huangjunjie2019, wnzhang}@sjtu.edu.cn

Abstract

Given the large volume of side information from different modalities, multimodal recommender systems have become increasingly vital, as they exploit richer semantic information beyond user-item interactions. Recent works highlight that leveraging Graph Convolutional Networks (GCNs) to explicitly model multimodal item-item relations can significantly enhance recommendation performance. However, due to the inherent over-smoothing issue of GCNs, existing models benefit only from shallow GCNs with limited representation power. This drawback is especially pronounced when facing complex and high-dimensional patterns such as multimodal data, as it requires large-capacity models to accommodate complicated correlations. To this end, in this paper, we investigate bypassing GCNs when modeling multimodal item-item relationship. More specifically, we propose a Topology-aware Multi-Layer Perceptron (TMLP), which uses MLPs instead of GCNs to model the relationships between items. TMLP enhances MLPs with topological pruning to denoise item-item relations and intra (inter)-modality learning to integrate higher-order modality correlations. Extensive experiments on three real-world datasets verify TMLP's superiority over nine baselines. We also find that by discarding the internal message passing in GCNs, which is sensitive to node connections, TMLP achieves significant improvements in both training efficiency and robustness against existing models.

1 Introduction

Recommender Systems (RS) are indispensable in modern web services, ranging from video streaming platforms to online shopping services (Lin et al. 2023; Liu et al. 2024; Xi et al. 2024). Traditional RS (Rendle et al. 2012; He et al. 2020) primarily rely on user-item interactions to learn latent representations. However, abundant multimodal information, such as images, videos, and text descriptions related to user and item attributes, is often left unused. This gap motivates the research of Multimodal Recommender Systems (MMRS), which leverage rich semantic multimodal data to improve recommendations (Wei et al. 2019).

The main challenge in MMRS lies in how to effectively incorporate multimodal features within the collaborative filtering (CF) framework (Huang et al. 2024). Traditional methods (He and McAuley 2016b; Liu, Wu, and Wang 2017) enhance item representations by combining multimodal features with item ID embeddings. Recently, the growing interest in graph-based recommendation methods (Wang et al. 2019; He et al. 2020) has led to research utilizing Graph Neural Networks (GNNs) to capture higher-order topological structures from multimodal features. These studies emphasize that explicitly modeling item-item relationships can significantly improve recommendation performance (Zhang et al. 2021a; Yu et al. 2023). Existing models typically rely on a single-layer Graph Convolutional Network (GCN) to capture these relations. Though achieving remarkable progress, a crucial limitation is its scalability (Yang et al. 2022). More specifically, **due to the *over-smoothing* issue (Chen et al. 2020; Rusch, Bronstein, and Mishra 2023), GCNs are inherently shallow and unable to abide by the scalability law in most deep learning models (Kaplan et al. 2020; Rosenfeld et al. 2019).** This limits their capacity to handle complex multimodal data patterns. In Figure 1, we compare the performance of two state-of-the-art (SOTA) MMRS, FREEDOM (Zhou and Shen 2023) and MGCN (Yu et al. 2023), on the Amazon Sports dataset. As GCN layers increase, time

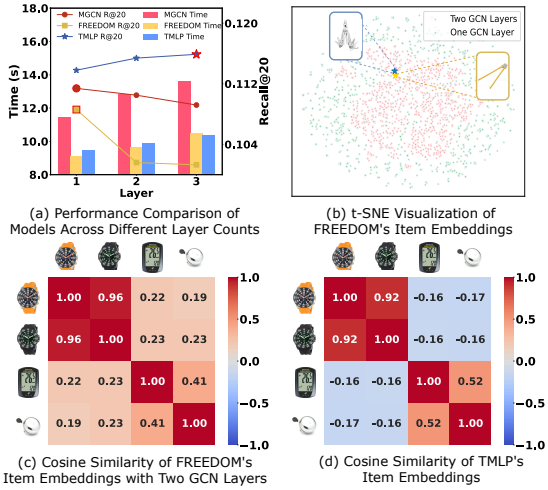

Figure 1: Scalability issues in existing MMRS and motivation for TMLP. Models are well-trained on Amazon Sports.

*Weinan Zhang is the corresponding author.

per epoch rises, but performance consistently declines. More GCN layers also lead to more concentrated item embeddings, reducing the model's discriminative ability. For instance, in FREEDOM with two GCN layers, unrelated items such as multitools and slingshots power bands (Figure 1(b)), as well as watches and bicycle accessories like bells and computers (Figure 1(c)), are drawn closer together. This implies that, unlike common architectures used in computer vision and natural language lrocessing tasks, given sufficient training data ($n > p$), increasing model capacity does not enhance representation power in GCNs. This bottleneck is a significant limitation for current MMRS frameworks.

Consequently, we argue that the complexity of multimodal data and the intricacies of cross-modality correlations may be too challenging for models with limited representation power. In this paper, we investigate the possibility of **capturing topological structure without graph convolution** by introducing Topology-aware MLP (TMLP). TMLP replaces GCNs' message passing with MLPs, which are better suited for handling high-dimensional, complex multimodal data. By sidestepping the message-passing mechanism, TMLP also gains higher efficiency. To the best of our knowledge, this is the first attempt in a similar line of work to model multimodal correlations without relying on GCNs. However, migrating from GCNs to MLPs in MMRS is not straightforward due to the following challenges:

- **(C1)** *How to train i.i.d model for non-i.i.d data?* Vanilla MLPs are designed for *identically and independently distributed* (i.i.d.) data, like images, and are topology-agnostic, making them blind to the rich correlations in non-i.i.d. data, such as graph-structured data. The key challenge is designing training objectives that enable MLPs to effectively incorporate topological information.
- **(C2)** *How to enhance TMLP's robustness with noisy connections in graph-structured data?* TMLP learns topological dependencies from injected item relations, but noisy connections can mislead the learning process and degrade performance. Previous efforts (Zhang et al. 2021a; Yu et al. 2023) derive item relations from modality-specific pre-trained models using a simple weighted sum. However, these representations may suffer from modality incompatibility issues, making the item relations noisy.

TMLP addresses these challenges with two key modules. For **Challenge 1**, TMLP employs *Intra (Inter)-Modality Learning* (IML) to learn topology-aware item representations by maximizing mutual information between an item's hidden representation and its neighborhood (Du et al. 2022, 2024). This is efficiently computed using Mutual Information Neural Estimation (MINE) (Belghazi et al. 2018), which transforms mutual information maximization into minimizing neighborhood alignment loss. This enables TMLP to learn the proximity of items within the same modality (intra-modality correlations) and the representation of similar items across different modalities (inter-modality correlations). For **Challenge 2**, TMLP uses *Topological Pruning Strategy* (TPS) to filter valuable information from pre-trained multimodal features, enhancing robustness against potential mismatches or misalignments between modalities. As shown in Figure 1(d), TMLP improves with increased network depth, maintaining the proximity of similar items while keeping unrelated items apart, preserving strong discriminative ability. In summary, the contributions of this paper are as follows:

- We propose TMLP, a scalable, efficient, robust, and high-performing MMRS method that pioneers altering the existing framework for item relationship modeling.
- We develop TPS, which effectively denoises and extracts valuable information from pre-trained features, enhancing robustness in both original and noise-added scenarios.
- We propose a unified learning procedure that enables modeling intra (inter)-modality correlations altogether, surpassing the explicit message-passing of GCN-based MMRS, thus free from their over-smoothing limitations.
- Extensive experiments on three real-world datasets show that TMLP significantly outperforms existing SOTA models, achieving over a 7% performance gain on Amazon Baby, highlighting the empirical superiority of TMLP.

2 Related Work

2.1 Graph Collaborative Filtering

User-item interactions can naturally be represented as a bipartite graph, prompting researchers to use GNNs for extracting user behavior features. Early GNN-based models, grounded in graph spectral theories, are computationally expensive (Monti, Bronstein, and Bresson 2017). More recent efforts have shifted focus to the spatial domain (Wu, Liu, and Yang 2018; Wang et al. 2019). Specifically, NGCF (Wang et al. 2019) captures user behavior features by iteratively aggregating neighbor data from the user-item view. LightGCN (He et al. 2020) simplifies the traditional graph convolution module, making it more suitable for recommendation scenarios. However, omitting modality information limits their precision in capturing user preferences.

2.2 Multimodal Recommendation

Multimodal recommender systems (MMRS) have evolved by integrating diverse information, such as visual and textual inputs, to alleviate data sparsity and cold-start issues in traditional CF (Su and Khoshgoftaar 2009). Early models like VBPR (He and McAuley 2016b) combines visual features from CNNs with ID embeddings, while VECF (Chen et al. 2019) uses VGG (Simonyan and Zisserman 2014) to model user attention on images and reviews. Later models like MMGCN (Wei et al. 2019) and GRCN (Wei et al. 2020) enrich representations using modality-specific graphs, and DualGNN (Wang et al. 2021) incorporates attention mechanisms to gauge user preferences across modalities and constructs a user co-occurrence graph. Recently, there is a rising trend in utilizing self-supervised learning techniques within MMRS. SLMRec (Tao et al. 2022) integrates self-supervised tasks like feature dropping into GNNs, while BM3 (Zhou et al. 2023) employs a latent embedding dropout mechanism and builds text and vision alignment loss.

Moreover, explicitly modeling item semantics has also been explored. LATTICE (Zhang et al. 2021a) merges

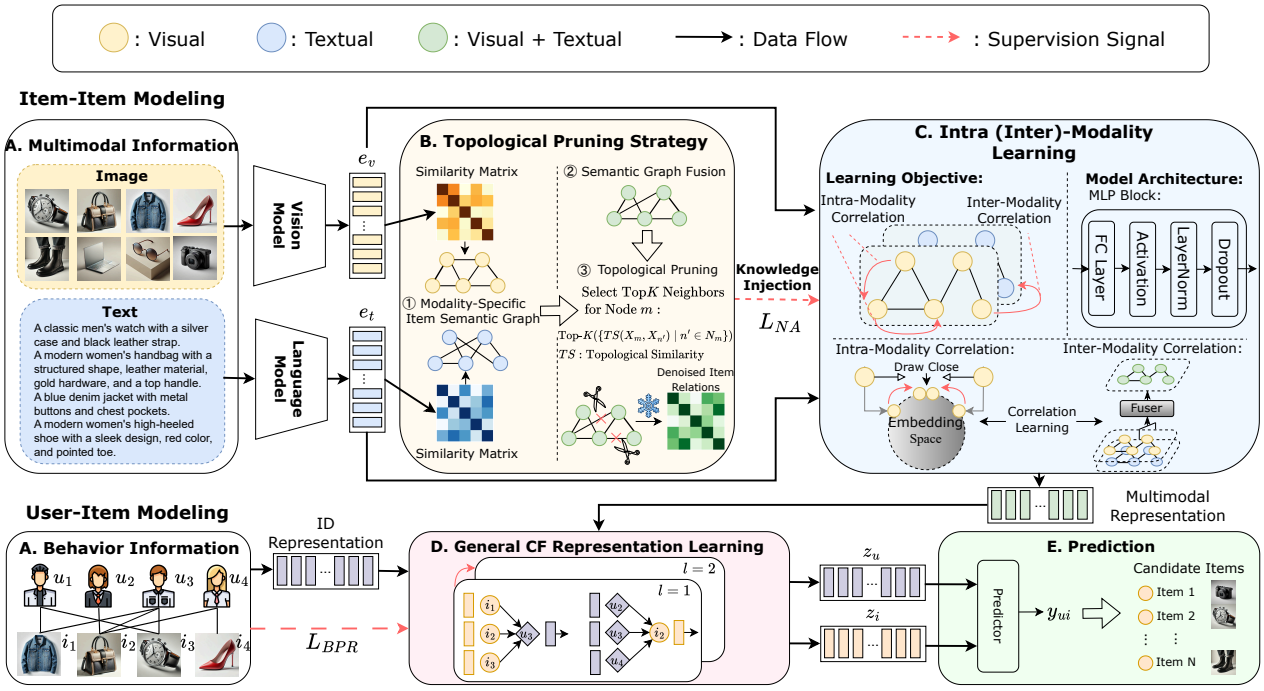

Figure 2: The framework of TMLP. It utilizes topological pruning strategy and intra (inter)-modality learning to capture modality correlations with MLPs (instead of GCNs in previous works). General CF representation learning refers to previous works that do not incorporate multimodal information, such as LightGCN.

modality-specific item graphs into an integrated graph, and FREEDOM (Zhou and Shen 2023) enhances efficiency by discretizing and freezing the item-item graph. MGCN (Yu et al. 2023) further utilizes attention mechanisms to identify modality significance. However, these approaches struggle with accurately depicting item relations due to GCN's over-smoothing issue and noise generated from directly merging modality graphs created by pre-trained models.

2.3 Connection between MLP and GNN

The connection between MLPs and GNNs is a fascinating area in deep learning. While MLPs process input features independently, GNNs use graph structures for message passing. Recent studies have aimed to enhance MLPs to match GNNs through techniques such as label propagation (Huang et al. 2020), contrastive learning (Hu et al. 2021; Dong et al. 2022), knowledge distillation (Zhang et al. 2021b; Tian et al. 2022), and additional regularization (Zhang et al. 2023). They seek to inject relational information into MLP, which is naturally agnostic to such information.

3 Problem Formulation

Denote user as $u \in \mathcal{U}$ and item as $i \in \mathcal{I}$, with $r_{u,i}$ indicating whether user u has interacted with item i. We construct the user rating matrix $\mathbf{R} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$ and a user-item bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ encompasses all users and items as nodes, and $\mathcal{E}$ depicts interactions as edges. Multimodal recommendations aim to predict user preferences using the interaction graph $\mathcal{G}$ and various multimodal features. Given user and item representations z_u and z_i, we optimize with Bayesian Personalized Ranking (BPR) (Rendle et al. 2012) loss, maximizing the difference in predicted scores for interacted and non-interacted user-item pairs:

$$\mathcal{L}_{\text{BPR}} = \sum_{(u,i,j)} -\ln \sigma(y_{ui} - y_{uj}), \quad y_{ui} = z_u^T z_i, \qquad (1)$$

where, u, i is a user-item pair, j is a randomly sampled non-interacted item, and σ is the sigmoid function. y_{ui} is the predicted score, typically an inner product of z_u and z_i.

4 Methodology

4.1 Framework Overview

Figure 2 provides an overview of TMLP, highlighting three main components: (i) Topological Pruning Strategy (TPS), which integrates and denoises item relations by leveraging the topological information of modality-specific graphs; (ii) Intra (Inter)-Modality Learning (IML), which uses MLPs to accurately model intra- and inter-modality correlations, producing comprehensive multimodal item representations by maximizing mutual information between node representations and their neighbors. TPS-derived item relations serve as additional supervision for MLP training; (iii) General Collaborative Filtering (CF) Representation Learning, which further refines user and item representations by combining multimodal and user-item interaction data, ultimately generating top-N recommendations through a predictor.

4.2 Topological Pruning Strategy

To ensure TMLP's robustness and avoid misguidance from noisy item relations, we propose the Topological Pruning Strategy (TPS). TPS aims to extract accurate information from pre-trained multimodal features, enhancing robustness against modality mismatches. Following prior work (Zhang et al. 2021a), we construct modality-specific item similarity matrices, A_v for vision and A_t for text, by computing feature similarity and retaining the top K' edges. The latent item-item graph A is then formed by a weighted sum of A_v and A_t, with β_m representing the importance of the visual modality. We set $K' = 10$ and $\beta_m = 0.1$ for consistency.

$$A = \beta_m \cdot A_v + (1 - \beta_m) \cdot A_t. \tag{2}$$

This approach is insufficient because representations from pre-trained models are specific to each modality and may suffer from modality incompatibility, leading to coarse and noisy item relations that can mislead subsequent representation learning. Inspired by Dong et al. (2022), TPS refines item relations by assessing topological similarity between nodes. Intuitively, if randomly selected nodes frequently fall within the neighborhoods of both nodes m and n, these nodes are likely to be topologically similar. We denote the topological information of node m as x_m, representing either $\mathcal{N}_m$ (its neighborhood) or $\overline{\mathcal{N}}_m = \mathcal{V} - \mathcal{N}_m$ (its complement). The probability of selecting a point within its neighborhood is given by Equation (3), where $a = 1$ corresponds to $\mathcal{N}_m$ and $a = 0$ corresponds to its complement. The topological information between nodes m and n is defined by the joint probability of selecting a point within both neighborhoods, as shown in Equation (4):

$$p\left(x_m = \mathcal{N}_m^a\right) = \frac{|\mathcal{N}_m^a|}{|\mathcal{V}|}, \tag{3}$$

$$p\left(x_m = \mathcal{N}_m^a, x_n = \mathcal{N}_n^b\right) = \frac{|\mathcal{N}_m^a \cap \mathcal{N}_n^b|}{|\mathcal{V}|}. \tag{4}$$

We define topological similarity between neighboring nodes m, n in Equation (5). Higher values indicate greater similarity. For each node m, we compute topological similarity with its neighbors and select the top-K to represent denoised item relations. The sampling size K is a tunable hyperparameter. As TPS relies solely on the topology of A, it can be performed before training without additional computational cost. The process is summarized in Equation (6), where $\mathbb{1}$ represents the indicator function and $\bar{A}$ stands for the denoised topological dependices between items:

$$TS(X_m; X_n) = \sum_{x_m} \sum_{x_n} p(x_m, x_n) \cdot \log \frac{p(x_m, x_n)}{p(x_m) \cdot p(x_n)}, \tag{5}$$

$$\bar{A}_{mn} = A_{mn} \cdot \mathbb{1}_{\{n \in \text{Top-}K(\{TS(X_m, X_{n'})\} | n' \in \mathcal{N}_m)\}}. \tag{6}$$

4.3 Intra-Modality Learning

With high-quality item relations, the next challenge is making the topology-agnostic MLP aware of topological dependencies. We achieve this by maximizing mutual information (Dong et al. 2022) between node representations and their contextual neighborhoods to capture intra-modality correlations. Mutual information, a core concept in information theory, quantifies dependency between variables. By maximizing the mutual information between the representations of adjacent items, we align them to be semantically similar, which mirrors the underlying mechanics of GCNs. We denote the Probability Density Function (PDF) of the node representation z_m as $p(Z(x))$ and the PDF of the contextual neighborhood representation c_m as $p(C(x))$. The joint PDF is $p(C(x), Z(x))$. The mutual information between node and neighborhood representations is defined as:

$$I(C(x); Z(x)) = \int p(C(x), Z(x)) \cdot \log \frac{p(C(x), Z(x))}{p(C(x)) \cdot p(Z(x))} dx. \tag{7}$$

However, mutual information is challenging to compute, so we use Mutual Information Neural Estimation (MINE) (Belghazi et al. 2018) to efficiently convert this into minimizing neighborhood alignment (NA) loss. NA loss estimates mutual information through node sampling, treating r-hop neighbors as contextual neighbors and other nodes as negative samples, as shown in Equation (8). The sim() function denotes cosine similarity, exp() is the exponential function, τ is the temperature parameter, and $\mathbb{1}$ is the indicator function. NA loss minimizes the feature distance between a node and its contextual neighbors while maximizing the distance from negative samples, aligning node features by their connectivity for a more accurate representation space:

$$\mathcal{L}_{\text{NA}} = -\mathbb{E}_{v_m \in \mathcal{V}} \left[\log \frac{\sum_{n=1}^{B} \mathbb{1}_{n \neq m} (\bar{A}_{mn})^r \exp(\text{sim}(z_m, z_n)/\tau)}{\sum_{k=1}^{B} \mathbb{1}_{k \neq m} \exp(\text{sim}(z_m, z_k)/\tau)} \right]. \tag{8}$$

4.4 Inter-Modality Learning

We now obtain multimodal item representations, as summarized in Equation (9). Our MLP structure includes linear layers followed by activation, Layernorm (Ba, Kiros, and Hinton 2016) for stability, and dropout. e_v and e_t represent the multimodal features from pre-trained models. Fuser combines these features to capture inter-modality correlations, using either a transformation matrix or simple concatenation. After obtaining h_i, we aggregate multimodal and ID-based information with graph $\mathcal{G}$ for user-item modeling, as in Equation (10), with h_u randomly initialized. The aggregator can be any CF model; we use LightGCN (He et al. 2020) for consistency with prior work.

$$h_v = \text{MLP}_v(e_v), \quad h_t = \text{MLP}_t(e_t), \quad h_i = \text{Fuser}(h_v, h_t). \tag{9}$$

$$z_u, z_i = \text{Aggregator}(h_u, h_i), \tag{10}$$

user and item representations at the $(l+1)^{th}$ GCN layer are given by Equation (11), with N_u and N_i as the one-hop neighbors in graph $\mathcal{G}$. The final representations z_u and z_i are obtained by summing the GCN layer outputs in Equation (12). The prediction score, computed as the inner product of these representations in Equation (1), is used to rank candidate items, with top-N items recommended to users:

$$h_u^{(l+1)} = \sum_{i \in N_u} \frac{1}{\sqrt{|N_u|}\sqrt{|N_i|}} h_i^{(l)}, \tag{11}$$

$$z_u = \sum_{l=0}^{L} h_u^{(l)}, \quad z_i = \sum_{l=0}^{L} h_i^{(l)}. \tag{12}$$

Table 1: Statistics of the experimental datasets.

Dataset	# Users	# Items	# Interactions	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%

4.5 Optimization Process

Incorporating the key components mentioned above, TMLP can now capture nuanced and complex modality correlations without explicit message passing. This section outlines TMLP's optimization process, including its objective function and procedures for training and inference.

Objctive Function To train the recommendation task, we use the BPR loss in Equation (1). Additionally, to guide TMLP with topological information, we employ an extra loss function, as shown in Equation (8). TMLP's joint loss function is expressed in Equation (13):

$$\mathcal{L} = \mathcal{L}_{\text{BPR}} + \alpha\mathcal{L}_{\text{NA}} + \lambda\|\Theta\|_2, \tag{13}$$

Here, λ is the L_2 regularization weight, Θ represents the model parameters, and α is a tunable parameter greater than zero, used to balance the losses. We train TMLP by minimizing the joint loss function on the training dataset.

Training and Inference TPS, based on the topological structure from pre-trained modality relations, is completed before training and does not affect efficiency. During forward propagation, TMLP learns topological dependices with neighborhood alignment loss. At inference, TMLP requires only node features without graph topological information.

5 Experiments

In this section, we present the details of our experiments. We are interested in the following research questions:
RQ1: How does TMLP perform against the SOTA methods?
RQ2: How does each component affect the performance?
RQ3: What is the influence of crucial hyperparameters?
RQ4: What is the training efficiency of TMLP?
RQ5: How does TMLP handle noisy connections?

5.1 Experiment Setup

Datasets. We conduct experiments on three categories from the Amazon review dataset (He and McAuley 2016a; McAuley et al. 2015): *Baby*, *Sports*, and *Electronics*. The dataset contains product descriptions and images as textual and visual features, with each review rating treated as a positive interaction. We retain 5-core users and items, using 4096-dimensional visual features and 384-dimensional textual features from prior research (Zhou et al. 2023). Table 1 summarizes the statistics for the datasets.

Baselines and Evaluation Metrics. We compare TMLP with several SOTA models, which can be divided into two groups: (1) General CF models: **BPR** (Rendle et al. 2012) and **LightGCN** (He et al. 2020). (2) Multimodal models: **VBPR** (He and McAuley 2016b), **MMGCN** (Wei et al. 2019), **GRCN** (Wei et al. 2020), **DualGNN** (Wang et al. 2021), **BM3** (Zhou et al. 2023), **MGCN** (Yu et al. 2023) and **FREEDOM** (Zhou and Shen 2023).

To ensure fair comparison, we use the same evaluation settings and 8:1:1 data split as in (Zhou et al. 2023) on the filtered 5-core data. We evaluate with NDCG@N and Recall@N (denoted as N@N and R@N) for top-N recommendation, averaging metrics across all test users for N=10 and N=20. During testing, we compute metrics using the all-ranking protocol based on user scores for all items.

Implementation Details. We implement TMLP[1] with PyTorch using MMRec (Zhou 2023), a comprehensive repository for multimodal recommendation models. Following He et al. (2020); Zhang et al. (2021a), we set the embedding size to 64, initialized with Xavier (Glorot and Bengio 2010). Optimal hyperparameters are found via grid search, with learning rates in {1e-4, 5e-4, 1e-3, 5e-3}, MLP layers in {2, 3, 4}, NA loss weight α in Equation (13) from 0 to 2 (interval 0.1), and sampling size K from 3 to 10. The regularization weight λ is set to 0. We fix the hidden dimension of MLP as 512, activation function as 'tanh' and dropout rate as 0. The ratio of visual features β_m in Equation (2) is set to 0.1, with $r = 1$ in Equation (8). Training is capped at 1000 epochs with early stopping at 20 consecutive epochs. We select the best models based on $Recall@20$ on the validation set and then report metrics on the test set.

5.2 Performance Comparison(RQ1)

Table 2 reports the recommendation performance of different models in terms of Recall and NDCG on three datasets, from which we have the following observations:

- TMLP outperforms existing multimodal baselines across all three datasets. Specifically, TMLP improves upon the second-best model by 7.53%, 5.49%, and 5.69% in $Recall@10$. This improvement is due to our use of MLPs instead of shallow GCNs to better capture high-dimensional multimodal information and avoid oversmoothing. Additionally, topological pruning strategy (TPS) denoises item relations from pre-trained visual and textual models, leading to more robust representations.
- Some baselines like FREEDOM, perform well on *Baby* and *Sports*, but underperform in *Electronics*, which has a larger user and item base. This is because FREEDOM constructs the item-item graph using multimodal information but only considers ID embeddings in recommendation process, neglecting multimodal integration. In contrast, TMLP integrates visual, textual, and ID information comprehensively, and bridges modality correlations by intra (inter)-modality learning. Additionally, TMLP's adjustable MLP architecture adapts to both small and large datasets, achieving better performance.
- The results show that models leveraging multimodal information consistently outperform matrix factorization (e.g., BPR) and graph-based models (e.g., LightGCN) that rely solely on ID features. This underscores the importance of incorporating multimodal knowledge and advanced correlation modeling in recommendations.

[1]Our code is available at https://github.com/jessicahuang0163/TMLP.

Table 2: Overall performance of various methods. Best results are in **boldface** and the second best is underlined. Improv. ↑ stands for the relative improvement of TMLP over the best baseline. We conduct experiments across 5 different seeds and * indicates that the improvements are statistically significant compared of the best baseline with $p < 0.01$ in t-test.

Dataset	Metric	General CF model		Multimodal model								
		BPR	LightGCN	VBPR	MMGCN	GRCN	DualGNN	BM3	MGCN	FREEDOM	TMLP	Improv. ↑
Baby	R@10	0.0357	0.0479	0.0423	0.0378	0.0532	0.0448	0.0534	0.0620	0.0624	**0.0671***	7.53%
	R@20	0.0575	0.0754	0.0663	0.0615	0.0824	0.0716	0.0845	0.0964	0.0985	**0.1016***	3.15%
	N@10	0.0192	0.0257	0.0223	0.0200	0.0282	0.0240	0.0285	0.0339	0.0324	**0.0360***	6.19%
	N@20	0.0249	0.0328	0.0284	0.0261	0.0358	0.0309	0.0365	0.0427	0.0416	**0.0449***	5.15%
Sports	R@10	0.0432	0.0569	0.0558	0.0370	0.0559	0.0568	0.0625	0.0729	0.0710	**0.0769***	5.49%
	R@20	0.0653	0.0864	0.0856	0.0605	0.0877	0.0859	0.0956	0.1106	0.1077	**0.1152***	4.16%
	N@10	0.0241	0.0311	0.0307	0.0193	0.0306	0.0310	0.0342	0.0397	0.0382	**0.0416***	4.79%
	N@20	0.0298	0.0387	0.0384	0.0254	0.0389	0.0385	0.0427	0.0496	0.0476	**0.0515***	3.83%
Electronics	R@10	0.0235	0.0363	0.0293	0.0207	0.0349	0.0363	0.0420	0.0439	0.0382	**0.0464***	5.69%
	R@20	0.0367	0.0540	0.0458	0.0331	0.0529	0.0541	0.0628	0.0643	0.0588	**0.0687***	6.84%
	N@10	0.0127	0.0204	0.0159	0.0109	0.0195	0.0202	0.0234	0.0245	0.0209	**0.0259***	5.71%
	N@20	0.0161	0.0250	0.0202	0.0141	0.0241	0.0248	0.0288	0.0298	0.0262	**0.0316***	6.04%

Table 3: Ablation study on different variants and modalities.

Dataset	Variant	R@10	R@20	N@10	N@20
Baby	TMLP	**0.0671**	**0.1016**	**0.0360**	**0.0449**
	TMLP_T	0.0656	0.1010	0.0354	0.0445
	TMLP_V	0.0538	0.0847	0.0296	0.0375
	Rand Pruning	0.0645	0.0993	0.0354	0.0443
	w/o Pruning	0.0602	0.0926	0.0323	0.0406
	w/o NA Loss	0.0519	0.0812	0.0281	0.0356
Sports	TMLP	**0.0769**	**0.1152**	**0.0416**	**0.0515**
	TMLP_T	0.0752	0.1128	0.0405	0.0502
	TMLP_V	0.0644	0.0981	0.0349	0.0436
	Rand Pruning	0.0742	0.1111	0.0402	0.0497
	w/o Pruning	0.0538	0.0838	0.0285	0.0362
	w/o NA Loss	0.0617	0.0948	0.0337	0.0422

5.3 Ablation Study (RQ2)

In this section, we conduct extensive experiments to analyze the performance of TMLP under different settings. The results on *Baby* and *Sports* are shown in Table 3.

Effect of Different Components of TMLP To examine the role of various components, we design the following variations: (i) **w/o NA Loss** disables correlation modeling by setting $\alpha = 0$ in Equation (13), and uses a vanilla MLP to capture item relations; (ii) **w/o Pruning** excludes TPS, relying on original item relations from FREEDOM for supervising correlation learning; (iii) **Rand Pruning** randomly samples the same number of neighbors as TMLP.

From Table 3, we observe: (i) NA loss is crucial in TMLP, significantly enhancing the model's ability to learn inter- and intra-modality correlations. Without it, an MLP fails to capture these relationships. (ii) Topological Pruning Strategy (TPS) is also essential. While acceptable on the smaller *Baby* dataset, omitting it on the larger *Sports* dataset leads to significant performance degradation due to increased noise. (iii) We find that random pruning performs slightly better due to small sample size ($K = 5$ in *Sports*) filtering out noise. However, the gap between random pruning and TMLP highlights the strength of TPS.

Effect of Single versus Multi-Modalities We compare TMLP's performance in uni-modal and multimodal settings. $TMLP_V$, $TMLP_T$ and $TMLP$ represent models using visual, textual, and both modalities. Table 3 indicates that both modalities are indispensable, while Amazon's textual features contribute more to performance than visual features.

5.4 Hyperparameter Sensitivity Study (RQ3)

MLP-related Hyperparameters We explore the impact of MLP-related hyperparameters like width, depth, and activation functions. Figure 3(a) shows that on larger datasets like *Electronics*, TMLP's performance improves with greater depth. In Figure 3(b), TMLP shows stable Recall@20 on smaller datasets like *Baby*, indicating its robustness to these parameters. This highlights TMLP's effectiveness in capturing complex cross-modality correlations.

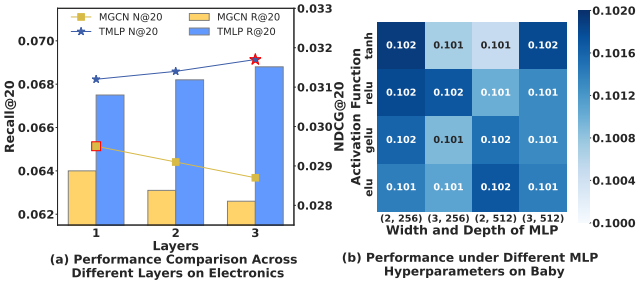

Figure 3: Hyperparameter study on width, depth and activation functions of MLP on *Baby* and *Electronics*.

Number of NA Loss Weight α We study the effect of NA loss weight α on *Baby* by tracking changes in Recall@20 and NDCG@20 in Figure 4(a). The metrics initially increase and then decrease, indicating an optimal α. A small α underfits inter- and intra-modality correlations, while a large α neglects the BPR loss, harming performance.

Number of Sampling Size K We examine the effect of sampling size K on *Sports* by tracking changes in Recall@20 and NDCG@20 in Figure 4(b). With an average node degree of 17-18, TMLP performs best at $K = 5$. A smaller K slightly reduces performance due to insufficient neighbors capturing modality correlations, yet TMLP still

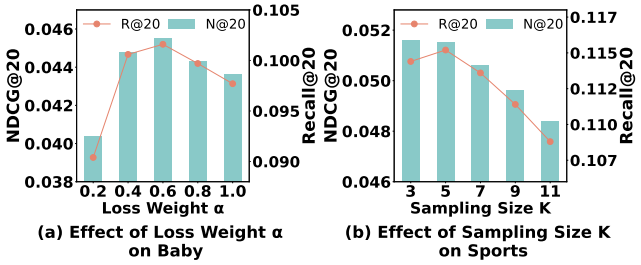

Figure 4: Hyperparameter study on NA loss weight α and sampling size K on *Baby* and *Sports*.

outperforms other models. Conversely, increasing K introduces noise, leading to a sharp performance drop.

5.5 Training Efficiency (RQ4)

We compare TMLP's efficiency with other models. Figure 5 shows Recall@20 across epochs on *Sports* and *Electronics*. With consistent hyperparameters (learning rate: 0.001, batch size: 2048), TMLP converges faster, reaching peak performance by the 30^{th} on *Sports* and 10^{th} epoch on *Electronics*. In contrast, FREEDOM and BM3 take around 200 epochs and still underperform. MGCN converges quickly but experiences a sharp performance drop, triggering early stopping. Notably, TMLP also has a shorter time per epoch compared to existing SOTA models, as shown in Figure 1(a). Additionally, TMLP's performance curve is more stable and smooth.

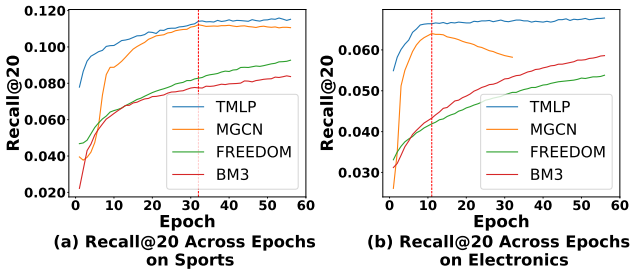

Figure 5: Comparison of training efficiency.

5.6 Resilience to Corrupted Data (RQ5)

In real world, mismatches and misalignments often introduce noise in item adjacency matrix A, which we refer to as corrupted connections. Our aim is for multimodal models to still learn effective cross-modality correlations in these scenarios. We define the corruption ratio ϵ as the proportion of noise, with the corrupted relations $A^{'}$ expressed as:

$$A^{'} = P \odot A + (1 - P) \odot N \tag{14}$$

$$P_{ij} = \begin{cases} 0, & \text{probability=}\epsilon \\ 1, & \text{probability=}1 - \epsilon \end{cases} \tag{15}$$

A is the original item adjacency matrix and N represents random noise. The operation $\odot$ indicates element-wise multiplication and P is a binary matrix determined by Equation (15). In Figure 6, TMLP shows strong performance

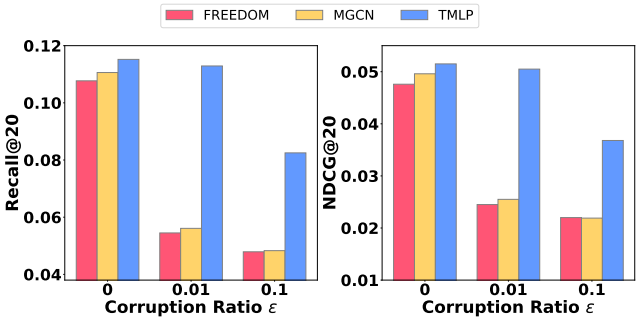

Figure 6: Performance on corrupted data on Sports.

on *Sports* despite noise. With a 1% noise ratio, TMLP effectively filters out irrelevant information via its topological pruning strategy, maintaining performance. In contrast, models like FREEDOM and MGCN suffer significant performance drops with even the slightest disturbance. These models even fall behind LightGCN, showing that multimodal integration with minor noise severely impacts recommendation quality, highlighting their lack of robustness.

5.7 Case Study

We conduct a case study to evaluate TPS and demonstrate TMLP's discriminative ability. In Figure 7(a), clustered circles represent items connected in the adjacency matrix of existing MMRS. TPS prunes these connections, retaining truly similar items while removing links between unrelated ones. Figure 7(b) presents a heatmap of cosine similarity in TMLP, illustrating that TMLP captures the similarity between boxing gloves while distancing other items, highlighting its strong discriminative ability.

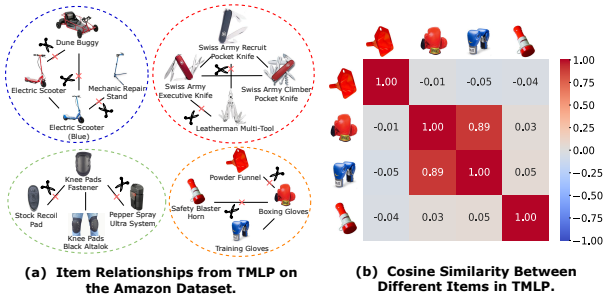

Figure 7: Case study on TMLP's discriminative power.

6 Conclusion

In this paper, we introduce TMLP, a novel topology-aware MLP framework that enhances scalability by replacing GCNs in MMRS. TMLP captures high-order and complex multimodal correlations through intra (inter)-modality learning, which injects topological dependencies between items into MLPs. To mitigate the impact of noisy connections that could degrade performance, TMLP employs a topological pruning strategy to effectively denoise item relations, improving robustness against modality mismatches.

Our evaluation on three real-world datasets shows that TMLP outperforms nine baseline models, demonstrating its effectiveness, superior training efficiency and robustness.

Acknowledgements

This work is partially supported by Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102) and National Natural Science Foundation of China (62322603, 62177033).

References

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *International conference on machine learning*, 531–540. PMLR.

Chen, D.; Lin, Y.; Li, W.; Li, P.; Zhou, J.; and Sun, X. 2020. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 3438–3445.

Chen, X.; Chen, H.; Xu, H.; Zhang, Y.; Cao, Y.; Qin, Z.; and Zha, H. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 765–774.

Dong, W.; Wu, J.; Luo, Y.; Ge, Z.; and Wang, P. 2022. Node representation learning in graph via node-to-neighbourhood mutual information maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16620–16629.

Du, K.; Chen, J.; Lin, J.; Xi, Y.; Wang, H.; Dai, X.; Chen, B.; Tang, R.; and Zhang, W. 2024. DisCo: Towards Harmonious Disentanglement and Collaboration between Tabular and Semantic Space for Recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 666–676.

Du, K.; Zhang, W.; Zhou, R.; Wang, Y.; Zhao, X.; Jin, J.; Gan, Q.; Zhang, Z.; and Wipf, D. P. 2022. Learning enhanced representation for tabular data via neighborhood propagation. *Advances in Neural Information Processing Systems*, 35: 16373–16384.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 249–256. JMLR Workshop and Conference Proceedings.

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.

Hu, Y.; You, H.; Wang, Z.; Wang, Z.; Zhou, E.; and Gao, Y. 2021. Graph-mlp: Node classification without message passing in graph. *arXiv preprint arXiv:2106.04051*.

Huang, J.; Chen, J.; Lin, J.; Qin, J.; Feng, Z.; Zhang, W.; and Yu, Y. 2024. A Comprehensive Survey on Retrieval Methods in Recommender Systems. *arXiv preprint arXiv:2407.21022*.

Huang, Q.; He, H.; Singh, A.; Lim, S.-N.; and Benson, A. R. 2020. Combining label propagation and simple models out-performs graph neural networks. *arXiv preprint arXiv:2010.13993*.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Lin, J.; Dai, X.; Xi, Y.; Liu, W.; Chen, B.; Zhang, H.; Liu, Y.; Wu, C.; Li, X.; Zhu, C.; et al. 2023. How can recommender systems benefit from large language models: A survey. *arXiv preprint arXiv:2306.05817*.

Liu, C.; Lin, J.; Wang, J.; Liu, H.; and Caverlee, J. 2024. Mamba4rec: Towards efficient sequential recommendation with selective state space models. *arXiv preprint arXiv:2403.03900*.

Liu, Q.; Wu, S.; and Wang, L. 2017. Deepstyle: Learning user preferences for visual recommendation. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, 841–844.

McAuley, J.; Targett, C.; Shi, Q.; and Van Den Hengel, A. 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.

Monti, F.; Bronstein, M.; and Bresson, X. 2017. Geometric matrix completion with recurrent multi-graph neural networks. *Advances in neural information processing systems*, 30.

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

Rosenfeld, J. S.; Rosenfeld, A.; Belinkov, Y.; and Shavit, N. 2019. A Constructive Prediction of the Generalization Error Across Scales. In *International Conference on Learning Representations*.

Rusch, T. K.; Bronstein, M. M.; and Mishra, S. 2023. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Su, X.; and Khoshgoftaar, T. M. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009.

Tao, Z.; Liu, X.; Xia, Y.; Wang, X.; Yang, L.; Huang, X.; and Chua, T.-S. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*.

Tian, Y.; Zhang, C.; Guo, Z.; Zhang, X.; and Chawla, N. 2022. Learning MLPs on graphs: A unified view of effectiveness, robustness, and efficiency. In *The Eleventh International Conference on Learning Representations*.

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.

Wang, X.; He, X.; Wang, M.; Feng, F.; and Chua, T.-S. 2019. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, 165–174.

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.

Wu, Y.; Liu, H.; and Yang, Y. 2018. Graph Convolutional Matrix Completion for Bipartite Edge Prediction. In *KDIR*, 49–58.

Xi, Y.; Liu, W.; Lin, J.; Cai, X.; Zhu, H.; Zhu, J.; Chen, B.; Tang, R.; Zhang, W.; and Yu, Y. 2024. Towards open-world recommendation with knowledge augmentation from large language models. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 12–22.

Yang, C.; Wu, Q.; Wang, J.; and Yan, J. 2022. Graph neural networks are inherently good generalizers: Insights by bridging gnns and mlps. *arXiv preprint arXiv:2212.09034*.

Yu, P.; Tan, Z.; Lu, G.; and Bao, B.-K. 2023. Multi-View Graph Convolutional Network for Multimedia Recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6576–6585.

Zhang, H.; Wang, S.; Ioannidis, V. N.; Adeshina, S.; Zhang, J.; Qin, X.; Faloutsos, C.; Zheng, D.; Karypis, G.; and Yu, P. S. 2023. Orthoreg: Improving graph-regularized mlps via orthogonality regularization. *arXiv preprint arXiv:2302.00109*.

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

Zhang, S.; Liu, Y.; Sun, Y.; and Shah, N. 2021b. Graph-less neural networks: Teaching old mlps new tricks via distillation. *arXiv preprint arXiv:2110.08727*.

Zhou, X. 2023. MMRec: Simplifying Multimodal Recommendation. *arXiv preprint arXiv:2302.03497*.

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. 2023. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, 845–854.

Appendix

Compared Baselines

- **BPR** (Rendle et al. 2012): This model captures latent factors representing both users and items from historical user-item interactions. It is known for introducing the widely adopted BPR loss function.
- **LightGCN** (He et al. 2020): LightGCN is a popular GCN-based collaborative filtering method that streamlines GCN design for improved recommendation performance.
- **VBPR** (He and McAuley 2016b): VBPR integrates visual features with ID embeddings to form item representations. It utilizes the BPR loss to capture user preferences.
- **MMGCN** (Wei et al. 2019): The model discerns modality-specific preferences by creating distinct graphs for each modality and employing these user-item graphs to learn representations within each modality. Finally, the model aggregates all modality-specific representations to get the final user and item embeddings.
- **GRCN** (Wei et al. 2020): This model employs the same user-item graphs as prior GCN-based models, but it identifies and eliminates false-positive edges within the graph. The refined graph is subsequently used to generate new representations through aggregation and information propagation.
- **DualGNN** (Wang et al. 2021): This model introduces a novel user-user co-occurrence graph by utilizing representations derived from modality-specific graphs and integrating neighbor representations.
- **BM3** (Zhou et al. 2023): This method streamlines the self-supervised approach by eliminating the need for randomly sampled negative examples. It uses the dropout technique to generate contrastive views and applies three contrastive loss functions to optimize the resulting representations.
- **MGCN** (Yu et al. 2023): This model addresses modality noise by refining modality features using item behavior information. It aggregates multi-view representations to enhance recommendation performance.
- **FREEDOM** (Zhou and Shen 2023): This model uncovers latent structures between items by constructing an item-item graph and freezing the graph before training. It introduces degree-sensitive edge pruning techniques to reduce noise in the user-item interaction graph.

Hyperparameter Settings

During the training process of TMLP, the specific hyperparameters are detailed in Table 4, with some basic settings adapted from MMRec (Zhou 2023).

Mean and STD Values in Ablation Study

To further verify the effectiveness of our framework, we conduct ablation studies using five different seeds, and present the mean and standard deviation values in Table 5. The results demonstrate that each component contributes to the performance of TMLP.

Table 4: Detailed hyperparameters when training TMLP.

Hyperparameters	Values
Embedding dimension	64
MLP hidden size	512
MLP layers	{2,3,4}
LightGCN layers on interaction graph	2
kNN K' when fusing A_v and A_t	10
β_m when fusing A_v and A_t	0.1
Learning rate	{1e-4, 5e-4, 1e-3, 5e-3}
Batch size	2048
Optimizer	Adam
NA loss weight α	0 to 2 (interval 0.1)
Temperature in NA loss τ	1
r-hop neighbors r	1
L_2 regularizer weight λ	{1e-3, 1e-2, 1e-1, 0}

Further Hyperparameter Analysis

Building on Section 5 in the original paper, Figure 8 presents additional hyperparameter analysis across various datasets, examining the effect of sampling size K on *Baby* and NA loss weight α on *Sports*.

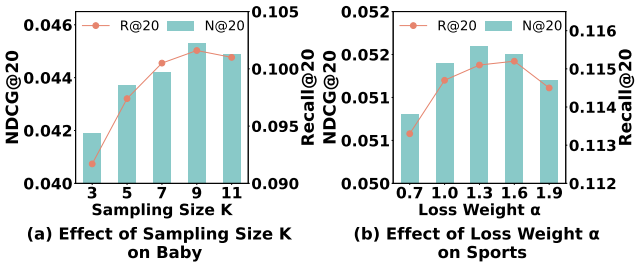

Figure 8: Hyperparameter study on NA loss weight α and sampling size K on *Baby* and *Sports*.

Visualization Analysis

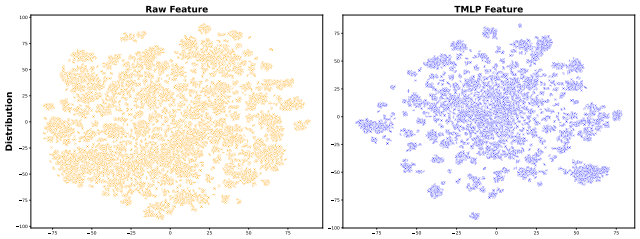

Figure 9: Item representation distribution. The left side of the figure presents the t-SNE visualization of raw features, while the right side showcases the features generated by TMLP.

The t-SNE visualization on the left in Figure 9 illustrates the distribution of raw features, where data points are uniformly spread, forming a homogeneous cluster. This lack of distinct clusters suggests limited discriminative power, which is crucial for accurate item differentiation in recommendations. Conversely, the t-SNE visualization on the

Table 5: Ablation study with mean and std values.

Dataset	Variant	R@10	R@20	N@10	N@20
Baby	TMLP	**0.0671±0.0011**	**0.1016±0.0008**	**0.0360±0.0005**	**0.0449±0.0004**
	TMLP_T	0.0656±0.0003	0.1010±0.0007	0.0354±0.0003	0.0445±0.0002
	TMLP_V	0.0538±0.0009	0.0847±0.0006	0.0296±0.0006	0.0375±0.0003
	Rand Pruning	0.0645±0.0006	0.0993±0.0009	0.0354±0.0002	0.0443±0.0002
	w/o Pruning	0.0602±0.0010	0.0926±0.0007	0.0323±0.0004	0.0406±0.0003
	w/o NA Loss	0.0519±0.0004	0.0812±0.0012	0.0281±0.0005	0.0356±0.0006
Sports	TMLP	**0.0769±0.0004**	**0.1152±0.0005**	**0.0416±0.0001**	**0.0515±0.0001**
	TMLP_T	0.0752±0.0006	0.1128±0.0006	0.0405±0.0003	0.0502±0.0003
	TMLP_V	0.0644±0.0003	0.0981±0.0002	0.0349±0.0001	0.0436±0.0001
	Rand Pruning	0.0742±0.0005	0.1111±0.0005	0.0402±0.0003	0.0497±0.0003
	w/o Pruning	0.0538±0.0007	0.0838±0.0008	0.0285±0.0004	0.0362±0.0004
	w/o NA Loss	0.0617±0.0005	0.0948±0.0008	0.0337±0.0004	0.0422±0.0005

right, representing features generated by TMLP, shows well-defined clusters. This indicates that TMLP enhances item representation by effectively capturing underlying relationships and distinctions, thereby improving the model's ability to differentiate between items for personalized recommendations.

Runtime Efficiency

Model	Epochs	Total Training Cost	Performance (R@20)
BM3	264	32 min	0.0956
FREEDOM	260	32.3 min	0.1077
MGCN	64	8.9 min	0.1106
TMLP	73	10.2 min	0.1152

Table 6: Efficiency comparison across models.

As noted in Section 4.5, topological pruning strategy (TPS) relies on pre-trained modality structures and is completed before training, so it does not impact efficiency. During forward propagation, we employ MLPs and refer to topological dependencies only in calculating the NA loss, allowing flexibility. At inference time, TMLP requires only node features without graph topology, achieving higher performance and efficiency than GCN-based methods. Efficiency based on convergence speed is shown in Figure 5, with total training cost included in Table 6.