Multi-Modal Self-Supervised Learning for Recommendation

Wei Wei
University of Hong Kong
weiweics@connect.hku.hk

Chao Huang*
University of Hong Kong
chaohuang75@gmail.com

Lianghao Xia
University of Hong Kong
aka_xia@foxmail.com

Chuxu Zhang
Brandeis University
chuxuzhang@brandeis.edu

ABSTRACT

The online emergence of multi-modal sharing platforms (*e.g.*, TikTok, Youtube) is powering personalized recommender systems to incorporate various modalities (*e.g.*, visual, textual and acoustic) into the latent user representations. While existing works on multi-modal recommendation exploit multimedia content features in enhancing item embeddings, their model representation capability is limited by heavy label reliance and weak robustness on sparse user behavior data. Inspired by the recent progress of self-supervised learning in alleviating label scarcity issue, we explore deriving self-supervision signals with effectively learning of modality-aware user preference and cross-modal dependencies. To this end, we propose a new Multi-Modal Self-Supervised Learning (MMSSL) method which tackles two key challenges. Specifically, to characterize the inter-dependency between the user-item collaborative view and item multi-modal semantic view, we design a modality-aware interactive structure learning paradigm via adversarial perturbations for data augmentation. In addition, to capture the effects that user's modality-aware interaction pattern would interweave with each other, a cross-modal contrastive learning approach is introduced to jointly preserve the inter-modal semantic commonality and user preference diversity. Experiments on real-world datasets verify the superiority of our method in offering great potential for multimedia recommendation over various state-of-the-art baselines. The implementation is released at: https://github.com/HKUDS/MMSSL.

CCS CONCEPTS

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

KEYWORDS

Self-Supervised Learning, Multi-Modal Recommendation

ACM Reference Format:
Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. 2023. Multi-Modal Self-Supervised Learning for Recommendation. In *Proceedings of the ACM Web Conference 2023 (WWW'23), May 1–5, 2023, Austin, TX, USA.* ACM, Austin, TX, USA, 13 pages. https://doi.org/10.1145/3543507.3583206

*Chao Huang is the corresponding author.

WWW'23, May 1–5, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583206

1 INTRODUCTION

Multimedia recommender systems play a crucial role in a wide range of content-sharing and e-commerce applications with massive web multimedia content, including micro-videos, images and songs [28]. In multimedia recommendation scenarios, various modalities of item content are involved, such as the visual, acoustic, and textual features of items [35]. Such multi-modal data characteristics may reflect users' preferences with fine-grained modality level.

Several research lines have emerged to incorporate multi-modal content into multimedia recommendation. For example, VBPR [10] as an early study to extend matrix decomposition framework to deal with the modality features of items. ACF [2] proposes to identify the component-level user preference via a hierarchically structured attention network. To explore high-order connectivity among user-item interactions, recently proposed methods (*e.g.*, MMGCN [45], GRCN [44], LATTICE [51]) adopt Graph Neural Networks (GNNs) to incorporate modality information into the message passing for inferring user and item representations. However, the satisfied performance of most existing multimedia recommenders usually requires sufficient high-quality label data (*i.e.*, observed user interactions) to train the models in a supervised manner. In real-life recommendation scenarios, interaction labels between users and items are very sparse compared with the whole interaction space [21, 46], which limits the representation ability of current fully supervised models to generate accurate embeddings to represent complex user preference in multimedia recommendation.

Inspired by the recent success of self-supervised learning (SSL) for data augmentation [17, 53], an antidote to mitigate the data sparsity limitation in multimedia recommendation is to generate supervisory signals from the unlabeled data. While some recent studies (*e.g.*, SGL [46], NCL [21], HCCF [47]) attempt to incorporate SSL into the modeling of user-item interactions for collaborative filtering, they fail to adapt the augmentation schemes to the specific multimedia recommendation task. For example, SGL [46] directly performs stochastic noise perturbation to dropout nodes and edges for graph augmentation. NCL [21] and HCCF [47] propose to discover implicit semantic node correlations over the observed user-item interactions. The overlook of multi-modal characteristics for augmentation may hinder the effectiveness of their introduced auxiliary SSL tasks to distill modality-aware signals.

Consider the presented the results on Amazon-baby data in Figure 1, several state-of-the-art multimedia methods are experimentally compared to make recommendation with respect to different

sparsity degrees from user and item side. To be specific, we notice that user interaction sparsity hinders the representation ability of compared methods to capture the genuine multi-modal preference of users. Significant performance gain is achieved by our MMSSL with highly sparse user interaction data (*e.g.*, < 4). By visualizing the performance distribution of item-specific prediction results, our method can be observed with more long-tail items distributed in the highlighted area. This gives an illustrated understanding of our model superiority in alleviating long-tail issue for recommendation.

To address the above limitations, this work explores multi-modal self-supervised learning paradigm to benefit multimedia recommendation from two perspectives of modality-aware augmentation.

Modality-aware Collaborative Relation Learning. In multimedia recommendation, the diversity of modalities can reflect different user preferences over items, which could be the interests in visual features of videos, acoustic characteristics of songs, and textual description of products [37]. Failing to inject modality-aware collaborative signals into the self-supervised learning task is insufficient to distill effective SSL signals. Hence, to perform the data augmentation with the awareness of modality-aware user preference, we propose an adversarial self-supervised learning method to explore the implicit relationships between users and items at the fine-grained modality-specific level. To distill the self-supervision signals pertinent to modality-aware user-item interaction patterns, our generative SSL-based augmentor integrates the modality-guided collaborative relation generator and discriminator, so as to model the influence of multi-modal content on user interaction behaviors.

Cross-Modality Dependency Modeling. Given that different modality-specific user preferences would interweave with each other in an implicit manner [34], leaving this fact untouched can easily lead to less accurate representations in preserving multi-modal user-item relationships. For instance, users may get attracted by a micro-video due to both its amazing visual content and instrumental background music. Additionally, purchase behaviors of customers in online retailers could be influenced by the presented product images as well as the posted product reviews. To enable our auxiliary supervision signals to be reflective of the modality-wise pattern influence, we propose to enhance our self-supervised learning paradigm with cross-modal dependency modeling.

In light of the aforementioned motivations for model design, we develop MMSSL, a new multimedia recommender that unifies the generative and contrastive SSL for modality-aware augmentation. We pursue a generic solution to jointly capture modality-specific collaborative effects and cross-modality interaction dependency, in recognizing multi-modal user preference. In particular, at the first stage of our SSL paradigm, we propose to train an adversarial relation learning network with the perturbed modality-aware user-item dependency weights. This scheme allows us to distill the useful multi-modal information and encode them into the latent user (item) embeddings facing the limitation of label scarcity. To tackle the challenge of adversarial learning over a sparse user-item connection matrix, we integrate the Gumbel-based projection and Wasserstein adversarial generation to mitigate the distribution gap. At the second stage of MMSSL, we introduce a cross-view multi-modal contrastive learning scheme to pursue encoding the influence among modality-wise preference into representations with model

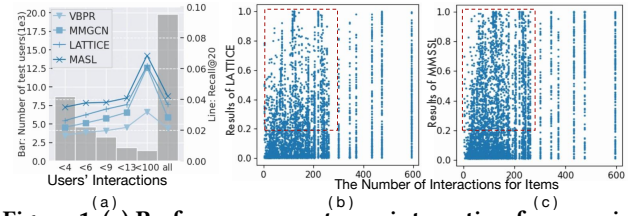

Figure 1: (a) Performance *w.r.t* user interaction frequencies. Left side y-axis: # of users in each group; Right side y-axis: performance of different methods. (b)-(c) Performance *w.r.t* item sparsity degrees. Each point represents item-specific recommendation accuracy averaged across epochs.

robustness. Furthermore, we offer the theoretical discussion of our self-supervised learning paradigm from viewpoints of: i) enhancing the multi-modal knowledge transfer to the collaborative view via adversarial self-augmentation; ii) benefiting the gradient learning with the cross-modal contrastive augmentation.

In a nutshell, our contributions can be summarized as follows.

- We tackle the label scarcity issue of multimedia recommendation with dual-stage self-supervised learning for modality-aware data augmentation. By unifying the recommendation task and our augmented generative/contrastive multi-modal SSL signals, significant performance gains can be achieved by our MMSSL.
- We propose a new recommender system MMSSL which integrates the generative modality-aware collaborative self-augmentation and the contrastive cross-modality dependency encoding.
- We extensively evaluate the proposed MMSSL to justify the model's effectiveness and robustness. In-depth and visual analyses demonstrate the rationality of our MMSSL method.

2 PRELIMINARY

Interaction Graph with Multi-Modality. In the context of graph learning, GNN-based collaborative filtering paradigms have offered state-of-the-art results [11, 39]. Inspired by this, MMSSL is built over the graph-structured interaction data. Specifically, we generate an user-item graph $G = \{(u, i)|u \in \mathcal{U}, i \in \mathcal{I}\}$, where $\mathcal{U}$ and $\mathcal{I}$ represents the set of users and items, respectively. We construct edge in G between user u and item i if an interaction is observed. We incorporate the multi-modal information (*e.g.*, textual, visual, acoustic modality) into the generated user-item interaction graph G with modality-aware characteristic features $\bar{\mathbf{F}} = \{\bar{\mathbf{f}}_i^1, ..., \bar{\mathbf{f}}_i^m, ..., \bar{\mathbf{f}}_i^{|\mathcal{M}|}\}$ of item i. Here, $\bar{\mathbf{f}}_i^m \in \mathbb{R}^{d_m}$ represents the raw feature embedding (with dimensionality d_m) of item i with modality $m \in \mathcal{M}$.

Task Formulation. We formulate our multi-modal recommender system that captures user-item relations with modality-aware user preference learning. In particular, given the generated multi-modal interaction graph $G = \{(u, i)|u \in \mathcal{U}, i \in \mathcal{I}, \bar{\mathbf{F}}\}$, our task of multi-modal recommender is to learn a function that forecasts how likely an item will be adopted by an user.

3 METHODOLOGY

3.1 Multi-Modal Self-Augmentation

In our model, we design a self-supervised learning task to supplement the user-item interaction modeling with multi-modal adversarial user-item relation learning. Benefiting from such design, our

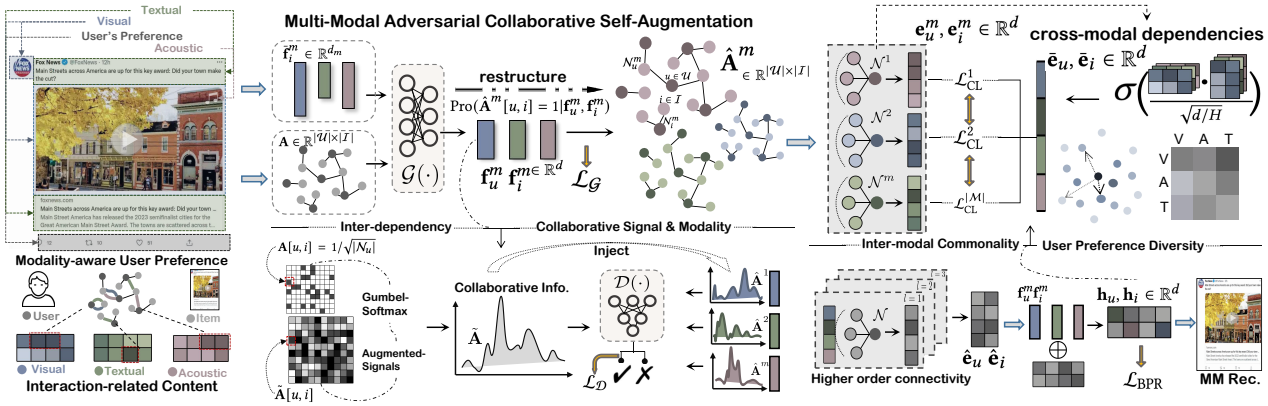

Figure 2: The model flow of our MMSSL. Gumbel-based transformation is integrated with Wasserstein adversarial generation to mitigate distribution gap between our augmented user-item relation matrix $\hat{\mathbf{A}}^m$ generated via $\mathcal{G}(\cdot)$ and the original A.

MMSSL not only captures the modality-aware user preference over items but also effectively alleviates the data scarcity with the derived self-supervision signals from multi-modal context. To achieve our goal, we propose an adversarial self-augmentation method which is composed of the modality-guided collaborative relation generator $\mathcal{G}(\cdot)$ and graph structure discriminator $\mathcal{D}(\cdot)$.

3.1.1 **Modality-guided Collaborative Relation Generator.** In the generative stage of our self-augmentation scheme, we aim to perform modality-aware graph structure learning over user-item interactions, so as to excavate modality-specific user preferences. In other words, we aim to derive the likelihood of user u interacting with item i given the corresponding multi-modal context:

$$\text{Pro}(\hat{\mathbf{A}}^m|\mathbf{F}^m) = \text{Pro}(\hat{\mathbf{A}}^m[u,i] = 1|\mathbf{f}_u^m, \mathbf{f}_i^m) \tag{1}$$

where $\hat{\mathbf{A}}^m \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$ represents the learned user-item interaction matrix under the modality m. Each element in $\hat{\mathbf{A}}^m$ is denoted as $\hat{\mathbf{A}}^m[u,i] \in \mathbb{R}$. To generate the input modality-aware user and item representations $\mathbf{f}_u^m$ and $\mathbf{f}_i^m$, we incorporate the collaborative effects into embeddings based on raw multi-modal feature vector $\bar{\mathbf{f}}_i^m$:

$$\mathbf{f}_u^m = \sum_{i\in\mathcal{N}_u} \bar{\mathbf{f}}_i^m/\sqrt{|\mathcal{N}_u|}; \quad \mathbf{f}_i^m = \sum_{u\in\mathcal{N}_i} \mathbf{f}_u^m/\sqrt{|\mathcal{N}_i|} \tag{2}$$

Here, $\mathcal{N}_u$ and $\mathcal{N}_i$ denotes the neighborhood set of user $u \in \mathcal{U}$ and item $i \in \mathcal{I}$ in the user-item interaction graph G. Before feeding the multi-modal feature vectors into our generator, we use the fully-connected layer with dropout [12] to perform a modality-specific transformation to map raw multi-modal features into latent embedding space, *i.e.*, $\bar{\mathbf{f}}_i^m \in \mathbb{R}^{d_m} \longrightarrow \bar{\mathbf{f}}_i^m \in \mathbb{R}^d, d \ll d_m$.

To capture the inter-dependency between the collaborative relationships and multi-modal contextual features, the modality-aware graph structure learning is conducted by learning the preference matrix $\hat{\mathbf{A}}^m \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$ with multi-modal representations:

$$\hat{\mathbf{A}}^m[u,i] = \mathcal{G}(\mathbf{f}_u^m, \mathbf{f}_i^m) = \mathbf{f}_u^{m\top} \cdot \mathbf{f}_i^m/(\|\mathbf{f}_u^m\|_2 \cdot \|\mathbf{f}_i^m\|_2) \tag{3}$$

To avoid calculating the entire interaction matrix, we conduct batch-based block matrix multiplication for memory-efficient implementation. With our collaborative relation generator $\mathcal{G}(\cdot)$, the learned modality-specific user-item relation matrix $\hat{\mathbf{A}}^m$ captures the inter-dependency between the collaborative view and multi-modal context view, which is reflective of modality-aware user preference.

3.1.2 **Discriminator with Adversarial Generation.** In our adversarial SSL paradigm, the discriminator $\mathcal{D}(\cdot)$ is designed to distinguish the generated modality-aware user-item relations $\hat{\mathbf{A}}^m$ and the observed user-item interaction matrix $\mathbf{A}$ from graph G. Through our relation discrimination process, our generator tries to confuse the discriminator by refining the learned modality-aware relation matrix. By doing so, the implicit inter-dependency modeling between the collaborative view and multi-modal context view can be enhanced by achieving adversarial robustness in representations. Formally, the discriminator $\mathcal{D}(\cdot)$ is built as follows:

$$\mathcal{D}(\mathbf{a}) = \delta(\Gamma^2(\mathbf{a})); \ \Gamma(\mathbf{a}) = \text{Drop}(\text{BN}(\text{LeakyRelu}(\text{Linear}(\mathbf{a})))) \tag{4}$$

Following the learning strategy in [25, 29], the input embedding $\mathbf{a} \in \mathbb{R}^{|\mathcal{I}|}$ corresponds to each matrix row, which is selected from either the generated relations $\hat{\mathbf{A}}^m$ or observed interactions $\mathbf{A}$, *i.e.*, $\{\mathbf{a}\,|\mathbf{a} \in \mathbf{A}\ or\ \mathbf{a} \in \hat{\mathbf{A}}^m, m \in \mathcal{M}\}$. $\Gamma(\cdot)$ denotes the discrimination neural layer by utilizing i) batch-normalization $\text{BN}(\cdot)$ for preventing the gradient vanishing issue [15]; ii) dropout $Drop(\cdot)$ for alleviating overfitting [12]; iii) non-linear activation $LeakyRelu(\cdot)$ for facilitating model convergence with more complete gradients [13, 23]. We stack two fully connected layers in our discriminator and adopt the sigmoid function $\delta(\cdot)$ to approximate the distribution of binary interaction data. $Linear(\cdot)$ applies a linear transformation to $\mathbf{a}$.

3.1.3 **Adversarial SSL against Distribution Gap.** Different from the dense pixel matrix of vision data, the observed user-item interaction matrix is highly sparse by including large number of zero values. Such sparsity poses unique challenges for our adversarial modality-aware relation learning, due to the data distribution difference between our generator $\mathcal{G}(\cdot)$ and discriminator $\mathcal{D}(\cdot)$. In particular, the user-item relation matrix $\hat{\mathbf{A}}^m$ learned by the generator with dense values is inevitably quite different from the observed interaction matrix $\mathbf{A}$, which may lead to the mode collapse and difficulty with convergence [31].

To address this challenge, we leverage *Gumbel-Softmax* [16] to transform the original interaction data into a dense matrix based on the Gumbel distribution, and bridge the distribution gap limitation.

To be specific, the adversarial enhancement is presented:

$$\tilde{\mathbf{A}}[u,i] = \underbrace{\frac{\exp\left(\left(\mathbf{A}[u,i]+g\right)/\tau\right)}{\sum_{i'}\exp\left(\left(\mathbf{A}[u,i']+g\right)/\tau\right)}}_{transformation} + \underbrace{\zeta \times \frac{\mathbf{h}_u^\top \mathbf{h}_i}{\|\mathbf{h}_u\|_2\|\mathbf{h}_i\|_2}}_{augmented\ signals} \tag{5}$$

Here, the perturbation factor g is calculated as $g = -log(-log(u))$ based on Gumbel (0, 1) distribution, where u ~ Uniform(0, 1) [16]. With the Gumbel-based transformation, the original observed interaction matrix $\mathbf{A}$ can be projected into a closely distributed matrix $\tilde{\mathbf{A}} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$. τ is the temperature factor to adjust the smoothness. To further improve the robustness of our adversarial SSL, we inject the auxiliary signals with the final multi-modal collaborative embeddings $\mathbf{h}_u, \mathbf{h}_i$ derived from Eq. 12. ζ is a weight parameter.

3.1.4 **Adversarial SSL Loss.** In our adversarial SSL task, we aim to capture the inter-dependency between the collaborative view and multi-modal view by aligning the distributions between our learned modality-aware user-item relations $\hat{\mathbf{A}}^m$ and the proxy $\tilde{\mathbf{A}}$ of the observed user-item interactions $\mathbf{A}$. Towards this end, we define our adversarial SSL loss to optimize our relation generator $\mathcal{G}(\cdot)$ and discriminator $\mathcal{D}(\cdot)$ in a minimax optimization manner as follows:

$$\min_{\mathcal{G}}\max_{\mathcal{D}} \mathbb{E}_{\tilde{\mathbf{A}}\sim\mathbb{P}_r}[\mathcal{D}(\tilde{\mathbf{A}})] - \mathbb{E}_{\hat{\mathbf{A}}^m\sim\mathbb{P}_f}[(\mathcal{D}(\mathcal{G}(\mathbf{f}^m))] \tag{6}$$

we separately present the optimized objectives $\mathcal{L}_\mathrm{G}$ and $\mathcal{L}_\mathrm{D}$ corresponding to the generator $\mathcal{G}$ and discriminator $\mathcal{D}$, respectively.

$$\begin{cases} \mathcal{L}_\mathcal{G} = -\mathbb{E}_{\hat{\mathbf{A}}^m}[\mathcal{D}(\hat{\mathbf{A}}^m)] \\ \mathcal{L}_\mathcal{D} = \mathbb{E}_{\tilde{\mathbf{A}}}[\mathcal{D}(\tilde{\mathbf{A}})] - \mathbb{E}_{\hat{\mathbf{A}}^m}[\mathcal{D}(\hat{\mathbf{A}}^m)] + \lambda_1\mathbb{E}_{\ddot{\mathbf{A}}}[(\|\nabla_{\mathcal{D}(\ddot{\mathbf{A}})}\|-1)^2] \end{cases} \tag{7}$$

To further enhance the robustness of our adversarial self-supervised learning against distribution gap and data sparsity, WassersteinGAN-GP [9] is introduced to add the gradient penalty with the balance weight λ_1. Here, $\ddot{\mathbf{A}}$ denotes the interpolation of $\hat{\mathbf{A}}^m$ and $\tilde{\mathbf{A}}$ matrix.

3.2 Cross-Modal Contrastive Learning

In multimedia recommendation scenarios, user interaction behavioral patterns with different item modalities (*e.g.*, visual, textual and acoustic) will influence each other. For example, the visual and acoustic features of a short video can jointly attract users to view it. Thus, the visual-specific and acoustic-specific preferences of users may interweave in a complex way. To capture such implicit dependencies among user modality-specific preferences, we design a cross-modal contrastive learning paradigm with modality-aware graph contrastive augmentation.

3.2.1 **Modality-aware Contrastive View.** To inject the modality-specific semantics into our contrastive learning component, we perform the information aggregation over the modality-aware semantic neighbors $\mathcal{N}_u^m$ and $\mathcal{N}_i^m$ of user u and item i. It can be derived from relational matrix $\hat{\mathbf{A}}^m$ (Eq. 3) learned in our generator.

$$\mathbf{e}_u^m = \sum_{i\in\mathcal{N}_u^m} \mathbf{e}_i/\sqrt{|\mathcal{N}_u^m|}; \qquad \mathbf{e}_i^m = \sum_{u\in\mathcal{N}_i^m} \mathbf{e}_u/\sqrt{|\mathcal{N}_i^m|} \tag{8}$$

$\mathbf{e}_u, \mathbf{e}_i \in \mathbb{R}^d$ are Xavier-initialized [8] id-corresponding embeddings. The multi-modal contextual information can be preserved in the modality-aware latent representations $\mathbf{e}_u^m, \mathbf{e}_i^m \in \mathbb{R}^d$.

Modality-wise Dependency Modeling. To capture the correlations between each pair of modality-specific user preferences, we design our modality-wise dependency encoder with a multi-head self-attention mechanism using the following formula:

$$\bar{\mathbf{e}}_u^m = \sum_{m'\in\mathcal{M}}^{|\mathcal{M}|} \Big\|_{h=1}^{H} \sigma\left(\frac{(\mathbf{e}_u^m\mathbf{W}_h^Q)^\top\cdot(\tilde{\mathbf{e}}_u^{m'}\mathbf{W}_h^K)}{\sqrt{d/H}}\right)\cdot\mathbf{e}_u^{m'} \tag{9}$$

where $\mathbf{W}_h^Q, \mathbf{W}_h^K \in \mathbb{R}^{d/H\times d}$ denote the h-head query and the key transformations for calculating the relation between modality pair (m, m'). H denotes the number of attention heads. $\sigma(\cdot)$ denotes the softmax function. Embeddings of both user and item side are refined in an analogous way. We then fuse the modality-specific embeddings through mean-pooling (*e.g.* $\bar{\mathbf{e}}_u = \sum_{m\in\mathcal{M}}^{|\mathcal{M}|}\bar{\mathbf{e}}_u^m/|\mathcal{M}|$), to generate the multi-modal user/item representations $\bar{\mathbf{e}}_u, \bar{\mathbf{e}}_i \in \mathbb{R}^d$.

Multi-Modal High-Order Connectivity. To explore the high-order collaborative effects with the awareness of multi-modal information, we build our encoder upon the graph neural network for recursive message passing with the matrix form as follows:

$$\hat{\mathbf{E}}_u^{l+1} = \mathbf{A}\cdot\hat{\mathbf{E}}_u^l; \qquad \hat{\mathbf{E}}_u^0 = \mathbf{E}_u + \eta\cdot\bar{\mathbf{E}}_u/\|\bar{\mathbf{E}}_u\|_2^2 \tag{10}$$

where $\hat{\mathbf{e}}_u^l \in \hat{\mathbf{E}}_u^l, \hat{\mathbf{e}}_u^{l+1} \in \hat{\mathbf{E}}_u^{l+1}$ denote the embeddings for the l-th and the $(l+1)$-th layer, respectively. The node degree-normalized matrix element $\mathbf{A}[u,i] = 1/\sqrt{|\mathcal{N}_u|}$ if user u has interacted with item i from the observed data. The zero-order embeddings $\hat{\mathbf{E}}_u^0$ are obtained by combining the initial id-corresponding embeddings $\mathbf{E}_u$ with the normalized modality-aware embeddings $\bar{\mathbf{E}}_u$ using the weight parameter η. In our multi-layer GNNs, the layer-specific embeddings are aggregated through mean-pooling to yield the output: $\hat{\mathbf{E}}_u = \sum_{l=0}^{L}\hat{\mathbf{E}}_u^l/L$, where L is the number of graph layers.

3.2.2 **Cross-Modal Contrastive Augmentation.** In this module, we introduce our multi-modal contrastive learning paradigm which distills the self-supervision signals for dependency modeling among modality-specific user preferences. Specifically, we adopt the InfoNCE loss function to maximize the mutual information between the modality-specific embeddings $\mathbf{e}_u^m$ and the overall user embedding $\mathbf{h}_u$ of the same user u. With the self-discrimination strategy [46, 47], embeddings from different users are treated as negative pairs. Our cross-modal contrastive loss is defined as:

$$\mathcal{L}_\mathrm{CL} = \sum_{m\in\mathcal{M}}^{|\mathcal{M}|}\sum_{u\in\mathcal{U}}^{|\mathcal{U}|} \log\frac{\exp s(\mathbf{h}_u, \mathbf{e}_u^m)}{\sum_{u'\in\mathcal{U}}\left(\exp s(\mathbf{h}_{u'}, \mathbf{e}_u^m) + \exp s(\mathbf{e}_{u'}^m, \mathbf{e}_u^m)\right)}$$
$$s(\mathbf{h}_u, \mathbf{e}_u^m) = \mathbf{h}_u^\top\cdot\mathbf{e}_u^m/(\tau'\cdot\|\mathbf{h}_u\|_2\cdot\|\mathbf{e}_u^m\|_2) \tag{11}$$

where $s(\cdot)$ denotes the similarity function, and τ' is the temperature coefficient. Our cross-modal contrastive learning aims to learn an embedding space in which different user representations are far apart, which allows our model to capture diverse user modality-specific preferences with uniformly-distributed embeddings.

3.3 Multi-Task Model Training

To generate our final user (item) representations $\mathbf{h}_u, \mathbf{h}_i \in \mathbb{R}^d$ for making predictions, we promote the cooperation between the collaborative view and multi-modal view by aggregating their corresponding encoded embeddings as follows:

$$\mathbf{h}_u = \hat{\mathbf{e}}_u + \omega \sum_{m \in \mathcal{M}}^{|\mathcal{M}|} \frac{\mathbf{f}_u^m}{\|\mathbf{f}_u^m\|_2}; \quad \mathbf{h}_i = \hat{\mathbf{e}}_i + \omega \sum_{m \in \mathcal{M}}^{|\mathcal{M}|} \frac{\mathbf{f}_i^m}{\|\mathbf{f}_i^m\|_2} \tag{12}$$

ω is the aggregation weight. Normalization is performed to alleviate the value scale difference between the embeddings of collaborative view ($\hat{\mathbf{e}}_u, \hat{\mathbf{e}}_i$) and multi-modal views ($\mathbf{f}_u^m, \mathbf{f}_i^m$). With the final embeddings, our MMSSL model makes predictions on the unobserved interaction between user u and item i through $\hat{y}_{u,i} = \mathbf{h}_u^\top \cdot \mathbf{h}_i$.

Model Optimization. We train our recommender with a multi-task learning scheme to jointly optimize MMSSL with i) the main user-item interaction prediction task $\mathcal{L}_{\text{BPR}}$; ii) adversarial modality-aware user-item relation learning task $\mathcal{L}_{\text{G}}$; iii) cross-modal contrastive learning task $\mathcal{L}_{\text{CL}}$. Formally, the jointly optimized objective is given ($\lambda_2, \lambda_3, \lambda_4$ are hyperparameters for loss term weighting):

$$\mathcal{L} = \mathcal{L}_{\text{BPR}} + \lambda_2 \cdot \mathcal{L}_{\text{CL}} + \lambda_3 \cdot \mathcal{L}_{\text{G}} + \lambda_4 \cdot \|\Theta\|^2 \tag{13}$$

$$\mathcal{L}_{\text{BPR}} = \sum_{(u,i_p,i_n)}^{|\mathcal{E}|} -\log\left(\text{sigm}(\hat{y}_{u,i_p} - \hat{y}_{u,i_n})\right) \tag{14}$$

i_p, i_n denotes the positive and negative samples for user u. The last term in $\mathcal{L}$ is the weight-decay regularization against over-fitting.

3.4 Theoretical Discussion of MMSSL

We offer theoretical analyzes to discuss the benefits of our multi-modal SSL: i) conducting collaborative knowledge transfer; ii) capturing the cross-modality commonality for each instance.

3.4.1 **Theoretical Analysis of Knowledge Transfer.** The purpose of the adversarial SSL task is to inject collaborative signals into the modality-specific distribution (*i.e.*, knowledge transfer). Theoretical analysis here supports for the associativity of adversarial transferability and knowledge transferability. In particular, we first provide formal definitions of those two transferability and then show their correlations. For knowledge transferability, we care about whether the source $\hat{\mathbf{A}}^m$ on data G can achieve low loss $\mathcal{L}(.;G)$ when combining with a trainable model $\mathcal{G}$ comparing with target $\tilde{\mathbf{A}}$, which is formally shown as follows:

$$\min_{\mathcal{G}} \mathcal{L}(\mathbf{G} \circ \hat{\mathbf{A}}^m; G) \quad \textit{compare with} \quad \min \mathcal{L}(\tilde{\mathbf{A}}; G) \tag{15}$$

We can observe that $\mathcal{G}$ will determine how to solve this optimization problem. For adversarial transferability, we introduce dedicated metrics ϱ_1, ϱ_2 (detailed in Appendix A.2), to offer a quantitative picture. Then, we can denote the upper bound of knowledge transferability by adversarial transferability (Appendix A.2) as:

$$\|\nabla\tilde{\mathbf{A}} - \nabla(\mathbf{G} \circ \hat{\mathbf{A}}^m)\|^2 \leq (1 - \varrho_1\varrho_2)L^2 \tag{16}$$

where the composite function $\mathbf{G} \circ \hat{\mathbf{A}}^m$ formally denotes the knowledge transferability on G to approximate $\tilde{\mathbf{A}}$ (assuming that $\tilde{\mathbf{A}}$ is L-Lipschitz continuous, *i.e.*, $\|\nabla\tilde{\mathbf{A}}\|_2 \leq L$). The adversarial transferability can be measured by the angle difference (ϱ_1) between source target gradients and the norm difference (ϱ_2). The inequality suggests that if adversarial transferability is high, there exists an affine transformation with the bounded norm. It provides a theoretical underpinning for self-augmentation in Sec. 3.1 and emphasizes the necessity of Sec. 3.1.3. We bridge the distribution gap to prevent mode collapse [31] in recommendation task.

Table 1: Statistics of experimented datasets with multi-modal item Visual(V), Acoustic(A), Textual(T) contents.

Dataset	Amazon				Tiktok			Allrecipes	
	Sports		Baby						
Modality	V	T	V	T	V	A	T	V	T
Embed Dim	4096	1024	4096	1024	128	128	768	2048	20
User	35598		19445		9319			19805	
Item	18357		7050		6710			10067	
Interactions	256308		139110		59541			58922	
Sparsity	99.961%		99.899%		99.904%			99.970%	

User-specific Patterns Modeling Through Gradient. Stacked GNN architecture may lead to performance degradation due to over-smoothing [19]. In contrast, our multi-modal contrastive learning is endowed with the capacity of encoding indistinguishable embeddings. Theoretically, by pushing the hard negatives away from the anchor, greater gradients can be obtained [36]. After obtaining the gradient of the negative sample (Eq. 18), we can get function $\phi(\cdot)$ by the proportional approximation (Eq. 20). It maps the similarity of the negative sample pair to the gradient of the negative node:

$$\phi(x) \propto \sqrt{1 - (x)^2} \cdot \exp(x/\tau) \tag{17}$$

where x is the input of $\phi(\cdot)$ calculated by normalized embedding of anchor node and negative sample. τ is the temperature coefficient. By analyzing the gradient with Appendix.Fig. 5, we can conclude that our contrastive learning paradigm will assign larger gradients to hard negative samples (other users) to enhance representation discrimination, *i.e.*, facilitating to model user-specific preference.

4 EVALUATION

In this section, we conduct experiments to validate the effectiveness of MMSSL method by answering the following research questions:

- **RQ1**: Can the proposed MMSSL method achieve performance improvement over various state-of-the-art (SOTA) baselines?
- **RQ2**: For key learning components in our MMSSL, what are their impacts in boosting the recommendation performance?
- **RQ3**: How effective is MMSSL in alleviating the sparsity issue?
- **RQ4**: How is training efficiency of MMSSL as epochs increases?
- **RQ5**: How does the hyperparameter settings impact the results?

4.1 Experimental Settings

4.1.1 **Dataset.** Experiments are conducted on four publicly available multi-modal recommendation datasets, *i.e.*, Tiktok, Amazon-Baby, Amazon-Sports, and Allrecipes. Data statistics with multi-modal feature embedding dimensionality are reported in Table 1.

- **TikTok**. This data is collected from TikTok platform to log the viewed short-videos of users. The multi-modal characteristics are visual, acoustic, and title textual features of videos. The textual embeddings are encoded with Sentence-Bert [30].
- **Amazon**. We adopt two benchmark datasets from Amazon with two item categories Amazon-Baby and Amazon-Sports [24]. In

Table 2: Performance comparison of baselines on different datasets in terms of *Recall*@20, *Precision*@20 and *NDCG*@20.

Baseline	Tiktok			Amazon-Baby			Amazon-Sports			Allrecipes		
	R@20	P@20	N@20	R@20	P@20	N@20	R@20	P@20	N@20	R@20	P@20	N@20
MF-BPR	0.0346	0.0017	0.0130	0.0440	0.0024	0.0200	0.0430	0.0023	0.0202	0.0137	0.0007	0.0053
NGCF	0.0604	0.0030	0.0238	0.0591	0.0032	0.0261	0.0695	0.0037	0.0318	0.0165	0.0008	0.0059
LightGCN	0.0653	0.0033	0.0282	0.0698	0.0037	0.0319	0.0782	0.0042	0.0369	0.0212	0.0010	0.0076
SGL	0.0603	0.0030	0.0238	0.0678	0.0036	0.0296	0.0779	0.0041	0.0361	0.0191	0.0010	0.0069
NCL	0.0658	0.0034	0.0269	0.0703	0.0038	0.0311	0.0765	0.0040	0.0349	0.0224	0.0010	0.0077
HCCF	0.0662	0.0029	0.0267	0.0705	0.0037	0.0308	0.0779	0.0041	0.0361	0.0225	0.0011	0.0082
VBPR	0.0380	0.0018	0.0134	0.0486	0.0026	0.0213	0.0582	0.0031	0.0265	0.0159	0.0008	0.0056
LightGCN-*M*	0.0679	0.0034	0.0273	0.0726	0.0038	0.0329	0.0705	0.0035	0.0324	0.0235	0.0011	0.0081
MMGCN	0.0730	0.0036	0.0307	0.0640	0.0032	0.0284	0.0638	0.0034	0.0279	0.0261	0.0013	0.0101
GRCN	0.0804	0.0036	0.0350	0.0754	0.0040	0.0336	0.0833	0.0044	0.0377	0.0299	0.0015	0.0110
LATTICE	0.0843	0.0042	0.0367	0.0829	0.0044	0.0368	0.0915	0.0048	0.0424	0.0268	0.0014	0.0103
CLCRec	0.0621	0.0032	0.0264	0.0610	0.0032	0.0284	0.0651	0.0035	0.0301	0.0231	0.0010	0.0093
MMGCL	0.0799	0.0037	0.0326	0.0758	0.0041	0.0331	0.0875	0.0046	0.0409	0.0272	0.0014	0.0102
SLMRec	0.0845	0.0042	0.0353	0.0765	0.0043	0.0325	0.0829	0.0043	0.0376	0.0317	0.0016	0.0118
MMSSL	**0.0921**	**0.0046**	**0.0392**	**0.0962**	**0.0051**	**0.0422**	**0.0998**	**0.0052**	**0.0470**	**0.0367**	**0.0018**	**0.0135**
p-value	$1.28e^{-5}$	$7.12e^{-6}$	$6.55e^{-6}$	$2.23e^{-6}$	$7.69e^{-6}$	$8.65e^{-7}$	$7.75e^{-6}$	$6.48e^{-6}$	$6.78e^{-7}$	$3.94e^{-4}$	$5.06e^{-6}$	$4.31e^{-5}$
Improv.	8.99%	9.52%	6.81%	16.04%	15.91%	14.67%	9.07%	8.33%	10.85%	15.77%	12.50%	14.40%

those datasets, textual feature embeddings are also generated via Sentence-Bert based on the extracted text from product title, description, brand and categorical information. The product images are used to generate 4096-d visual feature embeddings of items.

- **Allrecipes**. This dataset comes from one of the largest food-oriented social network platform by including 52,821 recipes in 27 different categories. For each recipe, its image and ingredients are considered as the visual and textual features. Following the setting in [7], 20 ingredients are sampled for each recipe.

4.1.2 **Evaluation Protocols.** To evaluate the accuracy of top-K recommendation results, we adopt three widely used metrics: Recall@K (R@K), Precision@K (P@K), and Normalized Discounted Cumulative Gain (N@K). Following the settings in [42, 44], we all-rank item evaluation strategy is used to measure the accuracy. The average scores over all users in the test set are reported.

4.1.3 **Baselines.** We compare MMSSL with SOTA multi-modal recommender systems, popular GNN-based collaborative filtering models, recently proposed SSL-based recommendation solutions.

i) **GNN-based Collaborative Filtering Models**

- **NGCF** [39]: The method leverages graph convolutional network to inject high-order collaborative signals into representations.
- **LightGCN** [11]: By removing the redundant transformation and non-linear activation, LightGCN simplifies the message passing for graph neural network-based recommendation.

ii) **SSL-based Recommendation Solutions**

- **SGL** [46]: This model improves the graph collaborative filtering with the incorporated contrastive learning signals using different data augmentation operators, *e.g.*, randomly node/edge dropout and random walk, to construct contrastive representation views.
- **NCL** [21]: In this approach, constrastive views are generated by identifying semantic and structural neighboring nodes with EM-based clustering, to generate the positive contrastive pairs.
- **HCCF** [47]: To supplement main recommendation objective with SSL task, HCCF leverages the hypergraph neural encoder to inject the global collaborative relations into the recommender.

iii) **Multi-Modal Recommender Systems**

- **VBPR** [10]: This is a representative work to incorporate multi-media features into the matrix decomposition framework.
- **LightGCN**-*M*: This baseline is generated by using SOTA GNN-based CF model LightGCN as backbone and incorporate multi-modal item features during the graph-based message passing.
- **MMGCN** [45]: It leverages graph convolutional network to propagate the modality-specific embedding and capture the modality-related user preference for micro-video recommendation.
- **GRCN** [44]: It is a structure-refined GCN multimedia recommender system, which yields the refined interactions to identify false-positive feedback and eliminate noisy with pruning edges.
- **LATTICE** [51]: It proposes to identify latent item-item relations with the generated item homogeneous graph. Connections will be added among items with similar modality features.
- **CLCRec** [43]: This model addresses the item cold-start issue by enhancing item embeddings with multi-modal features using mutual information-based contrastive learning.
- **MMGCL** [49]: It incorporates the graph contrastive learning into recommender through modality edge dropout and masking.
- **SLMRec** [33]: This method designs data augmentation on multi-modal content with two components, *i.e.*, noise perturbation over features and multi-modal pattern uncovering augmentation.

4.1.4 **Hyperparameter Settings.** Our MMSSL model is implemented with Pytorch. We adopt AdamW[22] and Adam[18] as the optimizer for generator and discriminator with the learning rate search range of $\{4.5e^{-4}, 5e^{-4}, 5.4e^{-3}, 5.6e^{-3}\}$ and $\{2.5e^{-4}, 3e^{-4}, 3.5e^{-4}\}$, respectively. The decay of L_2 regularization term is searched in $\{1.2e^{-2}, 1.4e^{-2}, 1.6e^{-2}\}$. For fair comparison, the embedding dimensionality of our MMSSL and other compared methods is set as 64. For our MMSSL model, the number of propagation layers over graph structure is tuned from {1, 2, 3, 4}.

4.2 Performance Comparison (RQ1)

Evaluation results are reported in Table 2, in which the performance of our MMSSL and the best-performed baseline are highlighted with bold and underlined, respectively. Key observations are as follows:

Table 3: Ablation study on key components of MMSSL

Data	Amazon-Baby		Allrecipes		Tiktok	
Metrics	Recall	NDCG	Recall	NDCG	Recall	NDCG
w/o-ASL	0.0907	0.0396	0.0326	0.0124	0.0801	0.0358
w/o-CL	0.0924	0.0408	0.0328	0.0130	0.0821	0.0351
w/o-GT	0.0929	0.0405	0.0325	0.0121	0.0815	0.0353
r/p-GAE	0.0931	0.0411	0.0331	0.0126	0.0843	0.0364
MMSSL	**0.0962**	**0.0422**	**0.0367**	**0.0135**	**0.0921**	**0.0392**

Table 4: Performance comparison *w.r.t* different interaction sparsity degrees on Amazon-Baby and Allrecipes datasets.

Baby (1e3)	[0,4) 9.783	[4,6) 4.081	[6,9) 2.823	[9,13) 1.649	[13,100) 1.109
VBPR	0.0166↑110.8%	0.0187↑101.6%	0.0197↑93.4%	0.0218↑87.6%	0.0319↑115.1%
MMGCN	0.0219↑59.8%	0.0247↑52.6%	0.0277↑37.6%	0.0315↑29.8%	0.0608↑12.8%
LATTICE	0.0263↑33.1%	0.0300↑25.7%	0.0337↑13.1%	0.0366↑11.8%	0.0619↑10.8%
SLMRec	0.0269↑30.1%	0.0291↑29.6%	0.0302↑26.2%	0.0318↑28.6%	0.0611↑12.3%
Ours	0.0350	0.0377	0.0381	0.0409	0.0686
Allrecipes (1e3)	[0,2) 8.720	[2,3) 4.643	[3,4) 3.221	[4,5) 1.790	[5,10) 1.431
VBPR	0.0039↑164.1%	0.0051↑132.7%	0.0056↑141.1%	0.0057↑127.6%	0.0069↑175.4%
MMGCN	0.0067↑53.7%	0.0084↑44.1%	0.0100↑35.0%	0.0104↑26.9%	0.0134↑41.8%
LATTICE	0.0072↑43.1%	0.0088↑37.5%	0.0105↑28.6%	0.0099↑33.3%	0.0159↑19.5%
SLMRec	0.0085↑21.2%	0.0099↑22.2%	0.0116↑16.4%	0.0114↑15.8%	0.0174↑9.2%
Ours	0.0103	0.0121	0.0135	0.0132	0.0190

- MMSSL shows promising performance by consistently outperforming all baselines on different datasets, which demonstrates the effectiveness of our proposed new model. We attribute the performance improvement to the integrated generative and contrastive SSL components for modality-aware data augmentation. Most of recently proposed multimedia recommendation methods perform better than graph-based collaborative filtering models, which indicates the effectiveness of incorporating multi-modal context in learning modality-aware collaborative relationships.
- While recent studies (*i.e.*, SGL, NCL, HCCF) attempt to augment user-item interaction modeling in a contrastive fashion, only marginal performance gains are achieved by them compared with NGCF and LightGCN. We postulate the reason is the ignorance of multi-modal contextual information for generating self-supervision signals. With the guidance of multi-modal patterns (*e.g.*, modality-specific user preference, cross-modal relatedness), our MMSSL distills modality-aware self-supervised signals to supplement the supervised task of multimedia recommendation.
- In comparison with multimedia recommender systems, the obvious performance improvement of our MMSSL can be observed. As to state-of-the-art SSL-based methods (CLCRec, MMGCL, SLMRec), our framework MMSSL leverages the generative adversarial SSL to construct modality-speicific user-item relation and contrastive cross-modal relatedness learning, for effective multi-modal augmentation. However, directly masking the modality user/item features in MMGCL may cause the important information loss, which makes the sparsity issue of inactive users even worse. Furthermore, SLMRec generates the augmented views via the pre-defined hierarchical correlations among different modality data, which may hinder the effects of self-supervised signals across various multimedia recommendation datasets.

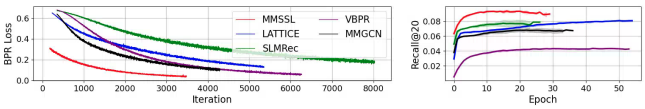

Figure 3: Training curves of MMSSL and compared methods.

4.3 In-Depth Analysis (RQ2 and RQ3)

4.3.1 **Ablation Study (RQ2).** Experiments are conducted to justify the importance of key components in MMSSL with the details. (1) We first disable the adversarial generative self-augmentation in the variant w/o-ASL. As shown in Table 3, the performance of w/o-ASL without our adversarial SSL decreases sharply compared with our MMSSL, demonstrating the strength of our designed generative augmentation with modality-guided self-supervised learning.

(2) We ablate the cross-modal contrastive learning paradigm with the variant w/o-CL. The superior model accuracy further reveals the significance of our contrastive augmentation by exploring the modality-wise dependencies for multimedia recommendation.

(3) To emphasize the rationality of our adversarial SSL method, we make another comparison between MMSSL and another variant (w/o-GT) without the Gumbel-based transformation. The observed performance gain reflects the improvements of our designed Gumbel-based transformation in enhancing the adversarial learning in addressing the distribution gap issue.

(4) Compared with r/p-GAE, which replaces our adversarial learning component with graph autoencoder for generating multi-modal user-item relational patterns, our MMSSL always performs the best. This indicates the superiority of capturing implicit user-item relations with the awareness of modality-aware preference using our framework for useful data augmentation in generative fashion.

4.3.2 **Evaluation *w.r.t* Data Sparsity (RQ3).** To examine the effectiveness of MMSSL in addressing the sparsity issue, we separately evaluate the recommendation accuracy with different interaction frequencies of users. According to the presented results (measured by NDCG@20) in Table 4, it is obvious that our MMSSL consistently outperforms the compared approaches under different interaction sparsity degrees, which further validates the rationality of our new self-supervised learning in mitigating the data sparsity issue to some degree. With the self-augmented multi-modal signals, our MMSSL can generate more accurate representations via effectively transferring multi-modal knowledge in an expressive manner.

4.3.3 **Model Convergence Analysis (RQ4).** In this section, we study the impact of our multi-modal self-supervised learning paradigm in model training efficiency with convergence analysis. In Figure 3, we show the training curves of MMSSL and compared methods on Amazon-Baby dataset, as the number of iterations and epochs increases. From the results, the faster converge speed of our MMSSL method is obviously observed, which suggests the advantage of MMSSL in training efficiency, meanwhile maintaining superior recommendation accuracy. This indicates that our incorporated multi-modal self-supervision signals bring positive effects to learn useful gradients during model optimization phase for fast convergence. The parameter inference procedures of compared learning methods (*e.g.*, LATTICE, MMGCN) require sufficient training labels (*i.e.*, user interactions) to learn good gradients, which are limited by the label shortage issue in practical scenarios.

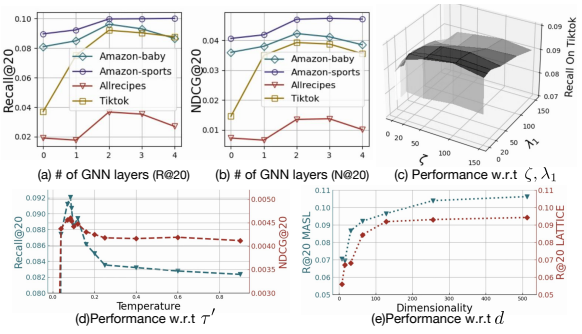

Figure 4: Impact study of hyperparameters in MMSSL.

4.4 Impact Study of Hyperparameters (RQ4)

In this section, we examine the sensitivity of several important parameters of our MMSSL model on different datasets.

- **Effect of # of GNN layers** L. We first investigate the influence of GNN model depth by varying the number of message passing layers L from 1 to 4. As we can see in Figure 4, our method performs the best with 2 or 3 graph layers. As the model goes deeper, the oversmoothing issue raises in the encoded embeddings, and thus decreasing the recommendation performance.
- **Effect of augmentation factors ζ, λ_1, in Adversarial SSL**. As presented in Section 3.1.3 and 3.1.4, ζ, λ_1 are augmentation factors in addressing the distribution gap in our multi-modal adversarial self-augmentation paradigm. Following similar settings in [27], ζ is selected from the range of [0, 10, 20, 50, 100, 150]. The best performance is obtained with $\zeta = 100$. λ_1 is tuned from [0, 1, 10, 20, 50, 100, 150] and performs the best when λ_1=1. The evaluation results indicate that the augmentation factors with appropriate values are effective to alleviate the distribution gap in our minmax optimization for adversarial self-augmentation.
- **Effect of temperature parameter** τ. We tune the hyperparameter τ from $(0, 0.9)$ to control the agreement strength between the instances of positive pairs. From the results on TikTok dataset, the best performance is achieved with $\tau = 0.085$.
- **Effect of latent dimensionality** d. The embedding dimensionality d of our model is searched from $(2^3, 2^4, 2^5, 2^6, 2^7, 2^8, 2^9)$. Comparing the performance of our MMSSL and the best-performed baseline (LATTICE), we observe the consistent performance improvement achieved by our method, which further justifies the effectiveness of our modality-aware self-supervision for enhancing model robustness by learning augmented representations.

5 RELATED WORK

GNN-based Recommender Systems. Graph neural networks have been widely adopted in recommender systems to model various relationships in different recommendation scenarios. For example, 1) user-item interactions in collaborative filtering (*e.g.*, LightGCN [11], LR-GCCF [3]); 2) user connections in social recommendation (*e.g.*, GraphRec [6], KCGN [14]); 3) item-item temporal transitional relationships in sequential recommendation (*e.g.*, GCE-GNN [40], SURGE [1]); and 4) entity-item dependencies in knowledge graph-enhanced recommender (*e.g.*, KGAT [38]). Motivated by these research studies, our MMSSL method adopts GNN as the backbone to model the high-order collaborative relationships with the injection of multi-modal contextual information.

Self-Supervised Learning for Recommendation. Recently, self-supervised learning (SSL) has shown its effectiveness in addressing label scarcity for recommendation [4, 21]. At its core is to augment original supervision signals with the incorporated auxiliary learning task. For graph augmentation with contrastive learning, NCL [21], CML [41] and HCCF [47] propose to generate SSL signals via contrasting positive node pairs based on various augmentation operators, *e.g.*, random walk graph sampling and semantic neighbor identification. For SSL-based sequence augmentation, CL4SRec [48] augments item sequence in three different ways, *i.e.*, crop, mask and reorder. S^3-Rec [52] performs contrastive learning among item sequence and attribute sequence. Additionally, for augmenting relational learning in social recommendation, MHCN [50] designs SSL task to capture high-order connectivity using mutual information maximization. Different from these works, for robust multi-modal user preference learning, we creatively design a new SSL recommender which adversarially trains a modality-aware neural graph generator to integrate with cross-modal contrastive augmentation.

Multimedia Recommendation. Many efforts have been devoted to enhancing recommender systems by incorporating multimedia content. One representative early study VBPR [10] extends matrix factorization to integrate both id-corresponding embeddings and multimedia feature embeddings of items. To improve the user-item relation modeling with multimedia content, attention mechanisms are used in ACF [2] and VECF [5] to capture complex user preference. In recent years, graph neural networks have been demonstrated as powerful solutions for multimedia recommendation by capturing high-order dependent structures among users and items. For example, graph convolutional network used in MMGCN [45] and GRCN [44], and graph attention mechanism applied in MK-GAT [32]. In this work, we propose a new multimedia recommender system which addresses the limitation of heavily rely on abundant labels in existing methods with a dual-stage SSL paradigm.

6 CONCLUSION

In this work, we tackle the problem of multimedia recommendation by proposing a multi-modal self-supervised learning model MMSSL. In MMSSL, we design a novel SSL task with modality-aware adversarial perturbation to capture multi-modal user preference under sparse interaction labels. In addition, we introduce the cross-modal contrastive learning paradigm to enable the dependency modeling across different modality-specific user interaction patterns. Through extensive experiments on several public datasets, the proposed MMSSL with effective self-supervision can achieve significant performance gains compared with various baselines. Future studies include extension of MMSSL to encode multi-interest preference of users with diverse latent embeddings. To incorporate multi-dimensional interests into our encoded user embeddings, the multi-modality information can be fed into our interest identification component by clustering multimedia items. In addition, in our future work, it is interesting to enhance the explainablity of our MMSSL by developing a GNN-based explainer to learn causal effects on modality-aware user-item interaction graph. It will facilitate the understanding of user preference with intuitive explanations.

REFERENCES

[1] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2021. Sequential recommendation with graph neural networks. In *SIGIR*. 378–387.

[2] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *SIGIR*. 335–344.

[3] Lei Chen, Le Wu, Richang Hong, Kun Zhang, and Meng Wang. 2020. Revisiting Graph Based Collaborative Filtering: A Linear Residual Graph Convolutional Network Approach. In *AAAI*, Vol. 34. 27–34.

[4] Mengru Chen, Chao Huang, Lianghao Xia, Wei Wei, Yong Xu, and Ronghua Luo. 2023. Heterogeneous Graph Contrastive Learning for Recommendation. In *WSDM*.

[5] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *SIGIR*. 765–774.

[6] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *WWW*. 417–426.

[7] Xiaoyan Gao, Fuli Feng, Xiangnan He, Heyan Huang, Xinyu Guan, Chong Feng, Zhaoyan Ming, and Tat-Seng Chua. 2019. Hierarchical attention network for visually-aware food recommendation. *Transactions on Multimedia (TMM)* 22, 6 (2019), 1647–1659.

[8] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*. JMLR Workshop and Conference Proceedings, 249–256.

[9] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. *Neurips* 30 (2017).

[10] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *AAAI*, Vol. 30.

[11] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *SIGIR*. 639–648.

[12] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CVPR* (2012).

[13] Hu Hu, Tian Tan, and Yanmin Qian. 2018. Generative adversarial networks based data augmentation for noise robust speech recognition. In *ICASSP*. IEEE, 5044–5048.

[14] Chao Huang, Huance Xu, Yong Xu, Peng Dai, Lianghao Xia, Mengyin Lu, Liefeng Bo, Hao Xing, Xiaoping Lai, and Yanfang Ye. 2021. Knowledge-aware coupled graph neural network for social recommendation. In *AAAI*. 4115–4122.

[15] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*. PMLR, 448–456.

[16] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).

[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *NeurIPS* 33 (2020), 18661–18673.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI*, Vol. 32.

[20] Kaizhao Liang, Jacky Y Zhang, Oluwasanmi O Koyejo, and Bo Li. 2020. Does Adversarial Transferability Indicate Knowledge Transferability? (2020).

[21] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. 2022. Improving Graph Collaborative Filtering with Neighborhood-enriched Contrastive Learning. In *WWW*. 2320–2329.

[22] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *ICLR*.

[23] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *ICML*.

[24] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*. 43–52.

[25] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. 2016. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163* (2016).

[26] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2018), 1979–1993.

[27] Henning Petzka, Asja Fischer, and Denis Lukovnicov. 2017. On the regularization of wasserstein gans. *ICLR* (2017).

[28] Ruihong Qiu, Sen Wang, Zhi Chen, Hongzhi Yin, and Zi Huang. 2021. Causalrec: Causal inference for visual debiasing in visually-aware recommendation. In *MM*. ACM, 3844–3852.

[29] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[30] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*.

[31] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. 2017. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems* 30 (2017).

[32] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal knowledge graphs for recommender systems. In *CIKM*. 1405–1414.

[33] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised Learning for Multimedia Recommendation. *Transactions on Multimedia (TMM)* (2022).

[34] Quoc-Tuan Truong, Aghiles Salah, and Hady Lauw. 2021. Multi-modal recommender systems: Hands-on exploration. In *Recsys*. ACM, 834–837.

[35] Di Wang, Quan Wang, Yaqiang An, Xinbo Gao, and Yumin Tian. 2020. Online collective matrix factorization hashing for large-scale cross-media retrieval. In *SIGIR*. 1409–1418.

[36] Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *CVPR*. 2495–2504.

[37] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. 2021. DualGNN: Dual Graph Neural Network for Multimedia Recommendation. *Transactions on Multimedia (TMM)* (2021).

[38] Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019. Kgat: Knowledge graph attention network for recommendation. In *KDD*. 950–958.

[39] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *SIGIR*.

[40] Ziyang Wang, Wei Wei, Gao Cong, Xiao-Li Li, Xian-Ling Mao, and Minghui Qiu. 2020. Global context enhanced graph neural networks for session-based recommendation. In *SIGIR*. 169–178.

[41] Wei Wei, Chao Huang, Lianghao Xia, Yong Xu, Jiashu Zhao, and Dawei Yin. 2022. Contrastive meta learning with behavior multiplicity for recommendation. In *WSDM*. 1120–1128.

[42] Yinwei Wei, Xiang Wang, Xiangnan He, Liqiang Nie, Yong Rui, and Tat-Seng Chua. 2021. Hierarchical user intent graph network for multimedia recommendation. *Transactions on Multimedia (TMM)* (2021).

[43] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive learning for cold-start recommendation. In *MM*. 5382–5390.

[44] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *MM*. 3541–3549.

[45] 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 *MM*. 1437–1445.

[46] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *SIGIR*. 726–735.

[47] Lianghao Xia, Chao Huang, Yong Xu, Jiashu Zhao, Dawei Yin, and Jimmy Xiangji Huang. 2022. Hypergraph Contrastive Collaborative Filtering. In *SIGIR*. 70–79.

[48] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *ICDE*. IEEE, 1259–1273.

[49] Zixuan Yi, Xi Wang, Iadh Ounis, and Craig Macdonald. 2022. Multi-modal Graph Contrastive Learning for Micro-video Recommendation. In *SIGIR*. 1807–1811.

[50] Junliang Yu, Hongzhi Yin, Jundong Li, Qinyong Wang, Nguyen Quoc Viet Hung, and Xiangliang Zhang. 2021. Self-Supervised Multi-Channel Hypergraph Convolutional Network for Social Recommendation. In *WWW*. 413–424.

[51] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *MM*. 3872–3880.

[52] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, et al. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *CIKM*. 1893–1902.

[53] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2021. Graph contrastive learning with adaptive augmentation. In *WWW*. 2069–2080.

A APPENDIX

In this section, in-depth details are included to: i) analyze the user preference diversity brought by contrastive learning paradigm from the perspective of gradients; ii) discuss the theoretical basis of self-augmented adversarial collaborative knowledge transfer; iii) analyze the time complexity of our MMSSL; iv) further justify the effectiveness of our proposed MMSSL with additional experiments.

A.1 Derivation of Negative Sample Gradient

Inspired by works [36, 46] which state contrastive loss with hardness-aware ability can push away the hard negative instances from the anchor by giving greater gradients, we leverage this property to tackle the over-smoothing issue [19] when encoding high-order collaborative signals and learning distinguishable embeddings.

Gradient of Positive and Negative Samples. We provide the gradient derivation procedure to support the point described in Sec. 3.1.1. The partial derivative of the anchor point is computed using $\mathcal{L}_{\mathrm{CL}_u}$ to examine the impact of samples on gradients:

$$
\begin{aligned}
\frac{\partial \mathcal{L}_{\mathrm{CL}_u}}{\partial \mathbf{q}_u} &= \frac{\partial}{\partial \mathbf{q}_u}\left(-log\frac{\exp\left(\mathbf{q}_u\cdot\mathbf{q}_P/\tau\right)}{\sum_{\mathcal{U}_A}\exp\left(\mathbf{q}_u\cdot\mathbf{q}_A/\tau\right)}\right)\\
&= \frac{\partial}{\partial \mathbf{q}_u}\left(-log\exp\left(\mathbf{q}_u\cdot\mathbf{q}_P/\tau\right)\right) + \frac{\partial}{\partial \mathbf{q}_u}\left(log\sum_{\mathcal{U}_A}\exp\left(\mathbf{q}_u\cdot\mathbf{q}_A/\tau\right)\right)\\
&= -\frac{1}{\tau}\cdot\mathbf{q}_P + \frac{1}{\tau}\frac{\sum_{\mathcal{U}_A}\mathbf{q}_A\cdot\exp\left(\mathbf{q}_u\cdot\mathbf{q}_A/\tau\right)}{\sum_{\mathcal{U}_A}\exp\left(\mathbf{q}_u\cdot\mathbf{q}_A/\tau\right)}\\
&= \frac{1}{\tau}\left(\frac{\sum_{\mathcal{U}_N}\mathbf{q}_N\cdot\exp\left(\mathbf{q}_u\cdot\mathbf{q}_N/\tau\right)}{\sum_{\mathcal{U}_A}\exp\left(\mathbf{q}_u\cdot\mathbf{q}_A/\tau\right)} + \frac{\mathbf{q}_P\cdot\exp\left(\mathbf{q}_u\cdot\mathbf{q}_P/\tau\right)}{\sum_{\mathcal{U}_A}\exp\left(\mathbf{q}_u\cdot\mathbf{q}_A/\tau\right)} - \mathbf{q}_P\right)\\
&= \underbrace{\frac{1}{\tau\cdot\parallel\mathbf{e}_u\parallel}\cdot\frac{\sum_{\mathcal{U}_N}\left(\mathbf{q}_N-\left(\mathbf{q}_u\cdot\mathbf{q}_N\right)\cdot\mathbf{q}_u\right)\cdot\exp\left(\mathbf{q}_u\cdot\mathbf{q}_N/\tau\right)}{\sum_{\mathcal{U}_A}\exp\left(\mathbf{q}_u\cdot\mathbf{q}_A/\tau\right)}}_{negative}\\
&+ \underbrace{\frac{1}{\tau\cdot\parallel\mathbf{e}_u\parallel}\cdot\left(\mathbf{q}_P-\left(\mathbf{q}_u\cdot\mathbf{q}_P\right)\cdot\mathbf{q}_u\right)\cdot\left(\frac{\exp\left(\mathbf{q}_u\cdot\mathbf{q}_P/\tau\right)}{\sum_{\mathcal{U}_A}\exp\left(\mathbf{q}_u\cdot\mathbf{q}_A/\tau\right)}-1\right)}_{positive}
\end{aligned}
\tag{18}
$$

where $\mathbf{q}_u$ represents the normalized (*e.g.*, $\mathbf{q}_u = \mathbf{e}_u/\parallel\mathbf{q}_u\parallel$) anchor node mapped into the same hyperspace with other nodes. $\mathbf{q}_P, \mathbf{q}_N, \mathbf{q}_A \in \mathbb{R}^{d\times 1}$ are the instances from the positive, negative, and the entire set, respectively. The chain rule employed in Eq. 18 is detailed in Eq. 22 and Eq. 23. Consequently, it can be clearly concluded that the gradient of user u can be determined using both the positive and negative pairs after the calculation in Eq. 18.

Proportional Relationship of Negative Sample Gradient. To get $\phi(x)$ which maps the similarity of the negative sample pair to the gradient of negative node, we focus on the gradient produced by negative cases shown as follows:

$$
\frac{\left(\mathbf{q}_N-\left(\mathbf{q}_u\cdot\mathbf{q}_N\right)\cdot\mathbf{q}_u\right)\cdot\exp\left(\mathbf{q}_u\cdot\mathbf{q}_N/\tau\right)}{\sum_{\mathcal{U}_A}\exp\left(\mathbf{q}_u\cdot\mathbf{q}_A/\tau\right)} \tag{19}
$$

The norm of the corresponding gradient is proportional to the term:

$$
\begin{aligned}
&\parallel\mathbf{q}_N-\left(\mathbf{q}_u\cdot\mathbf{q}_N\right)\cdot\mathbf{q}_u\parallel\left|\frac{\exp\left(\mathbf{q}_u\cdot\mathbf{q}_N/\tau\right)}{\sum_{\mathcal{U}_A}\exp\left(\mathbf{q}_u\cdot\mathbf{q}_A/\tau\right)}\right|\\
&\Rightarrow\sqrt{1-\left(\mathbf{q}_u\cdot\mathbf{q}_N\right)^2}\left|\frac{\exp\left(\mathbf{q}_u\cdot\mathbf{q}_N/\tau\right)}{\sum_{\mathcal{U}_A}\exp\left(\mathbf{q}_u\cdot\mathbf{q}_A/\tau\right)}\right|\\
&\propto\sqrt{1-\left(\mathbf{q}_u\cdot\mathbf{q}_N\right)^2}\cdot\exp\left(\mathbf{q}_u\cdot\mathbf{q}_N/\tau\right)
\end{aligned}
\tag{20}
$$

Given that both $\mathbf{q}_u$ and $\mathbf{q}_N$ are unit vectors, we introduce the variable x with the definition of $x = q_u\cdot q_N \in [-1, 1]$ to abbreviate the conclusion of Eq. 20 into function $\phi(\cdot)$:

$$
\phi(x) \propto \sqrt{1-(x)^2}\cdot\exp\left(x/\tau\right) \tag{21}
$$

To facilitate the analysis, we plot the gradient function $\phi(x)$ in Eq. 21 in Fig. 5. We can observe that the gradient of negative samples will rise as x increases. In other words, our contrastive learning paradigm will assign larger gradients to hard negative samples (other users) so as to enhance the discrimination of user representations.

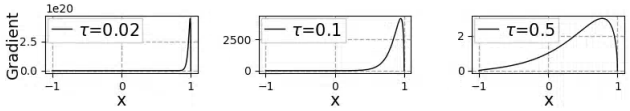

Figure 5: Gradient function $\phi(x)$ in Eq. 21 when τ = 0.02, τ = 0.1 and τ = 0.5. x is the similarity between positive and negative instances. This demonstrates that the gradient increases with decreasing temperature coefficient τ.

Chain Rule for Normalized Embedding. This part further discusses Eq.18. Specifically, the gradient of the loss with respect to $\mathbf{e}_u$ is related to that of $\mathbf{q}_u$ via the chain rule presented as:

$$
\frac{\partial \mathcal{L}_u(\mathbf{q}_u)}{\partial \mathbf{e}_u} = \frac{\partial \mathcal{L}_u(\mathbf{q}_u)}{\partial \mathbf{q}_u}\frac{\partial \mathbf{q}_u}{\partial \mathbf{q}_u} \tag{22}
$$

$$
\begin{aligned}
\frac{\partial \mathbf{q}_u}{\partial \mathbf{e}_u} &= \frac{\partial}{\partial \mathbf{e}_u}\left(\frac{\mathbf{e}_u}{\parallel\mathbf{e}_u\parallel}\right) = \frac{1}{\parallel\mathbf{e}_u\parallel}I - \mathbf{e}_u\left(\frac{\partial(1/\parallel\mathbf{e}_u\parallel)}{\partial \mathbf{e}_u}\right)^T\\
&= \frac{1}{\parallel\mathbf{e}_u\parallel}\left(I - \frac{\mathbf{e}_u\mathbf{e}_u^T}{\parallel\mathbf{e}_u\parallel^2}\right) = \frac{1}{\parallel\mathbf{e}_u\parallel}\left(I - \mathbf{q}_u\mathbf{q}_u^T\right)
\end{aligned}
\tag{23}
$$

Algorithm 1: The Learning Process of MMSSL Framework

Input: Adjacency matrix $\mathbf{A} \in \mathbb{R}^{|\mathcal{U}|\times|\mathcal{I}|}$ obtained by implicit feedback. Raw media feature $\bar{\mathbf{F}}$.
Output: User's interactive preference $\hat{y}_{u,i} (i \in \mathcal{I}, u \in \mathcal{U})$.

Initialize: Xavier initialized $\mathbf{e}_u, \mathbf{e}_i$, $\mathcal{G}$, $\mathcal{D}$, etc.
for *epoch* $\leftarrow 0, 1, \ldots$ **do**
  Update learning rate scheduler.
  for *step* $\leftarrow 0, 1, \ldots$ **do**
    for *D–step* **do**
      Prepare input for $\mathcal{D}$: $\hat{\mathbf{A}}^m, \tilde{\mathbf{A}} \longleftarrow$ Eq. 3, 5
      Get objective of $\mathcal{D}$: $\mathcal{L}_{\mathrm{D}} \longleftarrow$ Eq. 4, 7
      Gradient descent for $\mathcal{D}$ (with $\mathcal{G}$ no grad).
    end
    for *G–step* **do**
      Get quality prior for $\mathcal{G}$: $\mathbf{f}_u^m, \mathbf{f}_i^m \longleftarrow$ Eq. 2
      `// Modality-aware GNNs`
      Modality-aware ID embedding: $\mathbf{e}_u^m, \mathbf{e}_i^m \longleftarrow$ Eq. 8
      Inter-modal relation&High-order connectivity:
      $\bar{\mathbf{e}}_u^m, \bar{\mathbf{e}}_i^m \longleftarrow$ Eq. 9, $\hat{\mathbf{e}}_u, \hat{\mathbf{e}}_i \longleftarrow$ Eq. 10
      `// Objective of Multi-task Framework`
      Obtain final representations for recommendation: $\mathbf{h}_u, \mathbf{h}_i \longleftarrow$ Eq. 12
      Calculate the overall objective in Eq. 13:
      $\mathcal{L} \longleftarrow \mathcal{L}_{\mathrm{BPR}} + \mathcal{L}_{\mathrm{G}}$(Eq. 7) + $\mathcal{L}_{\mathrm{CL}}$(Eq. 11)
      Gradient descent back propagation.
    end
  end
end

A.2 Derivation of Transferability Bound

Our MMSSL introduces the multi-modal self-augmentation in Sec. 3.1 to characterize the inter-dependency between the collaborative view and modality-aware view. We will give the theoretical basis here for cross-view knowledge transfer with adversarial tasks.

Measurement of Adversarial Transferability. We offer metrics from the perspective of gradient difference ϱ_1 and norm difference ϱ_2 to provide quantitative understanding of transferability inspired by [20]. The two metrics can be calculated as follows:

$$\varrho_1(\mathbf{x}) = \frac{(\nabla f_S(\mathbf{x}) \cdot \nabla f_T(\mathbf{x}))^2}{\|\nabla f_S(\mathbf{x})\|_2^2 \cdot \|\nabla f_S(\mathbf{x})\|_2^2}, \varrho_2(\mathbf{x}) = \frac{(\Delta_{f_S,\delta_{f_S}} \cdot \Delta_{f_T,\delta_{f_S}})^2}{\|\Delta_{f_S,\delta_{f_S}}\|^2 \cdot \|\Delta_{f_T,\delta_{f_S}}\|^2} \tag{24}$$

where ϱ_1 is the gradient simplified with scale normalization which characterizes the similarity of the two inputs. ϱ_2 is information about the deviation of a function given attacks. $\delta(\cdot)$ indicates the attack difference $f(\mathbf{x}) - f(\mathbf{x}+\delta)$ on point $\mathbf{x}$ which can approximate as $\delta_f(\mathbf{x}) = \mathrm{argmax}\|\nabla f(\mathbf{x})^T\delta\|_2$ refer to [26], and $\Delta_{f,\delta(\mathbf{X})} = \nabla f(\mathbf{x})^T\delta(\mathbf{x})$. The adversarial attacks reflect comprehensive information of the gradients, enabling ϱ_1 and ϱ_2 to encode modality-aware collaborative relations. Both metrics are in [0, 1], where higher values indicate larger adversarial transferability.

Multi-Modal Knowledge Transfer with Adversarial Learning. The affine transformation with bounded norm (Ref to Eq 16) for transferability can be derived as follows:

$$\begin{aligned}
&\|\nabla f_T - \nabla(G \circ f_S)\|_G^2 = \|\nabla f_T - A\nabla f_S\|_G^2 \\
&= \mathbb{E}_{\mathbf{x}\sim G}\left[\|\nabla f_T(\mathbf{x}) - A\nabla f_S(\mathbf{x})\|_2^2\right] \\
&= \mathbb{E}_{\mathbf{x}\sim G}\left[\left\|cos(\Lambda(\mathbf{x}))\frac{\|\nabla f_T(\mathbf{x}))\|_2}{\|\nabla f_S(\mathbf{x}))\|_2}\nabla f_S(\mathbf{x}) + v(\mathbf{x}) - A\nabla f_S(\mathbf{x})\right\|_2^2\right] \\
&= \mathbb{E}_{\mathbf{x}\sim G}\left[\left(cos(\Lambda(\mathbf{x}))\frac{\|\nabla f_T(\mathbf{x}))\|_2}{\|\nabla f_S(\mathbf{x}))\|_2} - A\right)^2 \|\nabla f_S(\mathbf{x}))\|_2^2\right] \\
&+ \mathbb{E}_{\mathbf{x}\sim G}[(1-\varrho_1(\mathbf{x}))\|\nabla f_T(\mathbf{x})\|_2^2] \\
&= (1-\varrho_2)\mathbb{E}_{\mathbf{x}\sim G}\left[\varrho_1\|\nabla f_T(\mathbf{x}))\|_2^2\right] - \mathbb{E}_{\mathbf{x}\sim G}(1-\varrho_1(\mathbf{x}))\|\nabla f_T(\mathbf{x}))\|_2^2 \\
&= \mathbb{E}_{\mathbf{x}\sim G}[(1-\varrho_1\varrho_2)\|\nabla f_T(\mathbf{x})\|_2^2] \\
&\leq (1-\varrho_1\varrho_2)L^2
\end{aligned} \tag{25}$$

where $\Lambda(\mathbf{x})$ is the angle between $\nabla f_T(\mathbf{x})$ and $\nabla f_S(\mathbf{x})$ in Euclidean space. We define $v(\mathbf{x}) = \nabla f_T(\mathbf{x})) - cos(\Lambda(\mathbf{x}))\frac{\|\nabla f_T(\mathbf{x}))\|_2}{\|\nabla f_S(\mathbf{x}))\|_2}\nabla f_S(\mathbf{x}))$ for notation convenience. The matrix A degenerates to a scalar constant $A = (\Delta_{f_S,\delta_{f_S}}, \Delta_{f_T,\delta_{f_S}})/\|\Delta_{f_S,\delta_{f_S}}\|_G^2$.

A.3 Time Complexity Analysis

The time complexity of MMSSL mainly consists of three parts: i) Adversarial self-augmentation in $\mathcal{G}(\cdot)$ and $\mathcal{D}(\cdot)$. The transformations and matrix reconstruction in $\mathcal{G}(\cdot)$ take $O(|\mathcal{I}| \times \sum_m^{|\mathcal{M}|} d^m \times d)$ and $O(B\times|\mathcal{M}|\times|\mathcal{I}|\times d)$, with B denoting the batch size. $\mathcal{D}(\cdot)$ based on MLPs takes $O(B \times |\mathcal{I}|)$ complexity. ii) Modeling for cross-modality information and higher-order connectivity, containing view generation, scaled dot-product attention and high-order GNNs. This part requires $O(|\mathcal{M}| \times |\hat{\mathbf{A}}^m| \times d + |\mathcal{U}| \times |\mathcal{M}|^2 \times d + L \times |\mathbf{A}| \times d)$ computations, where $|\mathbf{A}|$ denotes the number of observed interactions. iii) Multi-task training with the BPR loss and contrastive loss. The recommendation takes $O(B \times d)$ complexity. The InfoNCE-based contrastive learning needs $O(|\mathcal{M}|\times B\times d)$ and $O(|\mathcal{M}|\times B\times|\mathcal{U}|\times d)$ cost for numerator and denominator calculation, respectively. The cost of MMSSL largely lies in the GNN calculation and model training, which is comparable with state-of-the-art SSL graph-based recommenders. Additionally, we employ batch reconstruction and lightweight GNN to further reduce the computational burden.

A.4 Supplementary Experiments

To suggest the model-agnostic property of our SSL components (*i.e.*, ASL and CL in Sec. 4.3.1) to enhance existing multimedia recommender, we add them to the best performed baseline SLMRec.

Table 5: Results of SLMRec equipped with our developed SSL modules in terms of Recall@20 and NDCG@20.

Data	Amazon-Baby		Allrecipes		Tiktok	
Metrics	R@20	N@20	R@20	N@20	R@20	N@20
SLMRec-*w-ASL*	0.0835	0.0351	0.0342	0.0125	0.0871	0.0366
SLMRec-*w-CL*	0.0790	0.0331	0.0327	0.0119	0.0853	0.0359

SLMRec-*w-CL* directly uses the output of GNNs, without using the spatial transformation in Eq.10 (SLMRec) [33]. SLMRec-ASL adds operations in Eq.2,3,4,5,7 of MMSSL. Performance improvement in Table 5 and Table 2 demonstrate the effectiveness of our two SSL modules designed for multimodal recommendation scenario.