Three Minds, One Student: Online Multi-Teacher Knowledge Distillation for Multimodal Recommendation

Hangtong Xu , **Yuanbo Xu*** , **En Wang***
College of Computer Science and Technology, Jilin University
xuht24@mails.jlu.edu.cn, {yuanbox, wangen}@jlu.edu.cn

Abstract

Existing multimodal recommendation models using complex fusion mechanisms (e.g., attention) or multi-stage processes (e.g., early or late fusion) integrate different modalities. However, attention-based adaptive fusion is prone to shortcut learning, where dominant collaborative signals (ID) can overshadow other modalities. This dominance affects the entire fusion process: early fusion often amplifies biases driven by identity, while late fusion struggles to extract preference-relevant signals from misaligned modalities, even with alignment regularization. To address these issues, we propose **M**ulti-**T**eacher **S**ingle-**S**tudent Online Distillation for **M**ulti**m**odal Recommendation (**MTS^{2}4MM**), which reframes the multimodal recommendation task from direct fusion to controllable knowledge transfer. Specifically, we construct multiple teachers to specialize in complementary perspectives, and a unified student distills their guidance via objectives at both the ranking and representation levels. This design explicitly controls modality contributions, improves robustness to modality noise and misalignment. Furthermore, we design a Modality-specific Preference Extractor to explicitly extract user preferences across different modalities equally. Extensive experiments across five real-world datasets demonstrate that MTS^{2}4MM consistently outperforms state-of-the-art baselines, achieving improvements of up to 7.22%.

1 Introduction

Multimodal recommendation seeks to improve collaborative filtering by utilizing rich item-level features, such as images and text. To achieve this, current approaches often rely on complex fusion methods, such as attention-based aggregation or multi-stage fusion processes, including early or late fusion, to integrate diverse modalities into a unified representation and improve recommendation performance.

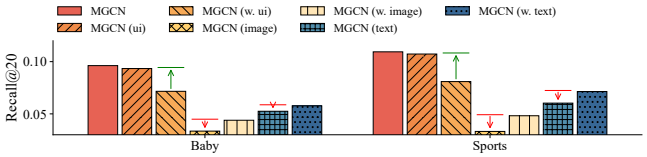

Figure 1: We present modality-specific variants of MGCN that operate under different training and inference settings. MGCN refers to the original model, which is trained and inferred using all modalities. MGCN (ui/image/text) performs inference using only the specified modality. In contrast, MGCN (w. ui/image/text) is a single-modality model in which both training and inference are conducted exclusively with the corresponding modality.

These methods implicitly assume that different modalities can be automatically balanced through learned fusion strategies, thereby leveraging their complementary strengths. However, recent empirical evidence suggests that this assumption is often violated in practice. In particular, multimodal recommendation models often suffer from modality imbalance, in which dominant collaborative signals–most notably item IDs–overshadow auxiliary modalities during joint learning. To gain a clearer understanding of this issue, we conducted a pilot study utilizing modality-specific variants of a representative multimodal graph model, MGCN, as shown in Figure 1. The results reveal two key findings: $\langle 1 \rangle$ the jointly fused model is predominantly influenced by the ID modality, with the contributions from image and text signals being minimal in the final representation; $\langle 2 \rangle$ current fusion strategies do not effectively leverage the complementary nature of different modalities; as a result, the fused model's performance is often worse than that achieved by independently modeling and inferring from each modality.

These findings expose fundamental limitations in fusion-based multimodal recommendation systems: $\langle 1 \rangle$ heterogeneous modalities carry preference knowledge with varying strengths and reliability, which makes them vulnerable to dominance effects during joint optimization; $\langle 2 \rangle$ fusion mechanisms implicitly assume that modality representations are well-aligned and equally informative, leaving little control over how modality-specific knowledge should influence recommendation outcomes. Consequently, fusion becomes a bottleneck that suppresses important modality signals rather than facilitating effective collaboration among them.

* Corresponding author.
https://github.com/MICLab-Rec/MTS4MM

These insights suggest that the core challenge of multimodal recommendation lies not in designing increasingly sophisticated fusion modules, but in regulating how preference knowledge from different modalities should be utilized and combined. Instead of forcing heterogeneous modalities to compete within a shared fusion space, it is more principled to treat each modality as an independent, potentially imbalanced source of preference knowledge, and to transfer what is genuinely beneficial for recommendation selectively.

We propose Multi-Teacher Single-Student Online Distillation for Multimodal Recommendation (**MTS^{2}4MM**), a new framework that rethinks multimodal recommendation by shifting from direct feature fusion to a controllable knowledge transfer approach. In this framework, we create multiple teacher models that specialize in complementary preference views, incorporating ID-based collaborative signals along with semantic information from images and texts. A unified student model selectively distills transferable knowledge from these teachers, focusing on both ranking and representation objectives. By separating modality learning from direct fusion and explicitly managing the contributions of each modality during the knowledge transfer process, MTS^{2}4MM effectively reduces modality dominance and enhances robustness against modality noise and misalignment. Additionally, we introduce a Modality-Specific Preference Extractor that differentiates between shared and modality-specific preference components, allowing for fair utilization of various modalities without increasing inference costs. Extensive experiments conducted on five real-world datasets show that MTS^{2}4MM consistently surpasses state-of-the-art baselines, achieving improvements of up to 7.22%. The contributions of this work can be summarized as follows:

- To the best of our knowledge, this work represents the first attempt to systematically reformulate multimodal recommendation as a controllable knowledge-transfer problem, rather than relying on conventional fusion-based paradigms.
- We propose MTS^{2}4MM, a multi-teacher single-student online distillation framework that reformulates multimodal recommendation as a selective knowledge transfer problem, explicitly regulating modality contributions without additional inference cost.
- Extensive experiments on five real-world datasets demonstrate that MTS^{2}4MM consistently outperforms state-of-the-art baselines and effectively mitigates the dominant-modality issue.

2 Problem Definition

We clarify the conventional multimodal recommendation problem and the knowledge-transfer-based multimodal recommendation problem proposed in this paper.

2.1 Conventional Multimodal Recommendation

Given a user set $\mathcal{U}$, an item set $\mathcal{I}$, and an interaction matrix $\mathbf{Y}$, each item i is associated with multiple observed modalities $\mathbf{X}_i^{(m)}$. Conventional multimodal recommendation systems aim to encode each modality and combine the representations into a single item latent space using a fusion function, such as attention or gating, using the merged item representation along with the user embedding to predict user preference:

$$\mathbf{z}_i = \mathcal{F}\left(\{\mathbf{z}_i^m\}_{m\in\mathcal{M}}\right), \quad \hat{y}_{u,i} = g(\mathbf{z}_u, \mathbf{z}_i). \tag{1}$$

The main goal of the conventional approach is to develop an effective fusion strategy $\mathcal{F}(\cdot)$ to combine diverse modality features to enhance recommendation performance.

2.2 Knowledge Transfer for Multimodal Recommendation

Given a user set $\mathcal{U}$, an item set $\mathcal{I}$, and an interaction matrix $\mathbf{Y}$, each item i is associated with multiple observed modalities $\mathbf{X}_i^{(m)}$. We consider a set of modality-specific teacher models $\{\mathcal{T}^{(k)}\}_{k=1}^{|\mathcal{M}|}$, where each teacher captures a distinct source of preference knowledge (e.g., ID-based collaborative signals or semantic information from images and texts). Given a user–item pair $\langle u, i\rangle$, the k-th teacher produces a modality-specific preference signal:

$$\pi_{u,i}^{(k)} = \mathcal{T}^k(u, i), \tag{2}$$

$\pi_{u,i}^{(k)}$ encodes the teacher's preference view over item i for user u. Our goal is to learn a unified student model $\mathcal{S}$ to predict user preference by selectively transferring knowledge from multiple teachers:

$$\min_{\theta_{\mathcal{S}}} \mathcal{L}_{rec}(\mathcal{S}) + \sum_{k=1}^{|\mathcal{M}|} \mathcal{D}(\mathcal{S}, \mathcal{T}^k), \tag{3}$$

where $\mathcal{L}_{\text{rec}}$ is the recommendation loss (e.g., BPR), $\mathcal{D}(\cdot)$ denotes a knowledge transfer (distillation) loss that enforces consistency between the student and the k-th teacher.

Compared to conventional work that focuses on how to fuse modality features, we explicitly address what preference-relevant knowledge should be transferred, enabling an effective balance of modality contributions and mitigating the dominance-signal problem.

3 Methodology

3.1 Modality Graph Construction

Following previous studies [Mao *et al.*, 2021; He *et al.*, 2023; Zhou *et al.*, 2023a], we construct modality-specific user–item graphs to encode high-order collaborative signals associated with each modality.

User–item collaborative graph. Given the user-item interaction matrix $\mathbf{Y} \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $\mathbf{Y}_{u,i} = 1$ signifies an observed user behavior on item i, we construct the user-item behavioral graph as follows:

$$\mathcal{A} = \begin{bmatrix} \mathbf{0} & \mathbf{Y} \\ \mathbf{Y}^\top & \mathbf{0} \end{bmatrix}, \quad \mathcal{D} = \begin{bmatrix} \mathcal{D}_u & \mathbf{0} \\ \mathbf{0} & \mathcal{D}_i \end{bmatrix}. \tag{4}$$

We explicitly decompose the degree matrix of the bipartite graph into user-side ($\mathcal{D}_u = \sum_i \mathbf{Y}_{u,i}$) and item-side ($\mathcal{D}_i = \sum_u \mathbf{Y}_{u,i}$) components when describing item-to-user aggregation. Furthermore, to alleviate the scale discrepancy caused by node degrees, we apply symmetric normalization:

$$\tilde{\mathcal{A}} = \mathcal{D}^{-\frac{1}{2}}\mathcal{A}\mathcal{D}^{-\frac{1}{2}}. \tag{5}$$

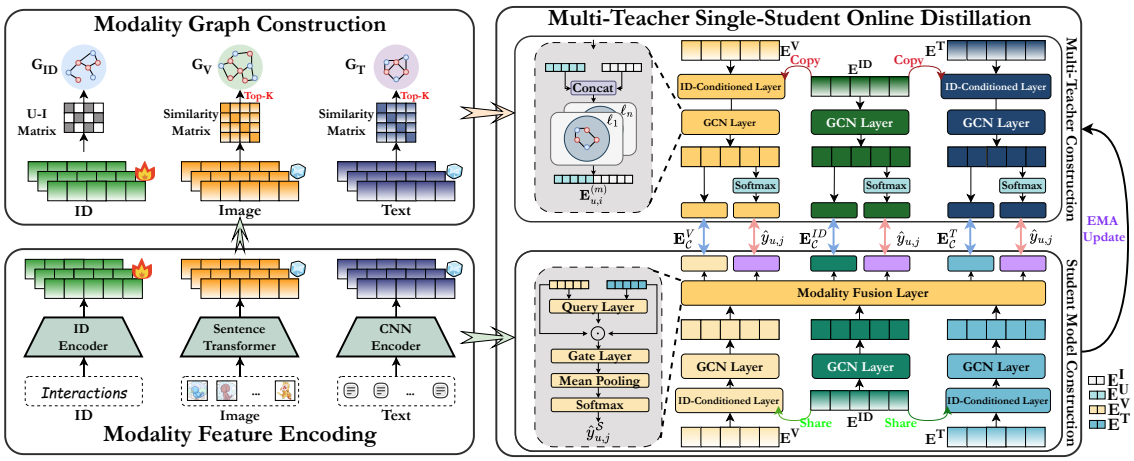

Figure 2: The overall architecture of MTS^{2}4MM consists of three modality-specific teachers (ID, Image, and Text) and a single fused student.

Modality-specific item–item graph. For each content modality $m \in \mathcal{M} = \{v, t\}$, we construct a modality-specific graph based on raw modality features $\mathbf{X}^{(m)} \in \mathbf{R}^{|\mathcal{I}| \times d_m}$. Specifically, we compute the cosine similarity between item features and use it to define edge weights:

$$\text{sim}_{i,j}^{(m)} = \frac{\mathbf{x}_i^{(m)} \cdot \mathbf{x}_j^{(m)}}{\|\mathbf{x}_i^{(m)}\|_2 \, \|\mathbf{x}_j^{(m)}\|_2}, \tag{6}$$

where $\mathbf{x}_i^{(m)}$ denotes the feature vector of item i under modality m. For each item i, we retain its top-k nearest neighbors $\mathcal{N}_k(i)$ and construct a sparse KNN adjacency matrix $\mathcal{A}^{(m)}$, where $\mathcal{A}_{i,j}^{(m)} = 1$ if $j \in \mathcal{N}_k(i)$, else 0. Similar to the user–item graph, we apply symmetric normalization to obtain:

$$\tilde{\mathcal{A}}^{(m)} = (\mathcal{D}^{(m)})^{-\frac{1}{2}} \mathcal{A}^{(m)} (\mathcal{D}^{(m)})^{-\frac{1}{2}}, \tag{7}$$

where $\mathcal{D}^{(m)}$ is the diagonal degree matrix of $\mathcal{A}^{(m)}$.

3.2 Modality Preference Extractor

ID preference encoder. Given the user embeddings $\mathbf{E}_u \in \mathbb{R}^{|\mathcal{U}| \times d}$ and item embeddings $\mathbf{E}_i^{id} \in \mathbb{R}^{|\mathcal{I}| \times d}$, we encode collaborative preference signals by propagating representations on the normalized user–item graph $\tilde{\mathcal{A}}$. Specifically, starting from the initial embeddings $\mathbf{E}_{ui}^{(0)} = [\mathbf{E}_u; \mathbf{E}_i^{id}]$, we perform ℓ layers of graph convolution:

$$\mathbf{E}_{ui}^{(\ell+1)} = \tilde{\mathcal{A}} \mathbf{E}_{ui}^{(\ell)}, \quad \ell = 0, \ldots, n. \tag{8}$$

To capture high-order collaborative dependencies while avoiding over-smoothing, we aggregate representations from all layers via layer-wise mean pooling. The final ID-based collaborative preference representation is given by:

$$\mathbf{E}^{ID} = \frac{1}{n+1} \sum_{\ell=0}^{\ell=n} \mathbf{E}_{ui}^{(\ell)} \in \mathbb{R}^{(|\mathcal{U}|+|\mathcal{I}|) \times d}. \tag{9}$$

ID-conditioned modality encoding (image/text views). For each content modality $m \in \mathcal{M} = \{t, v\}$, we construct modality-specific preference representations under the collaborative semantic space defined by item IDs. Following prior works, we first transform raw modality features into the latent space using modality-specific encoders. Specifically, we employ a Sentence Transformer $\mathcal{F}_{\text{text}}(\cdot)$ to encode textual content and a convolutional neural network $\mathcal{F}_{\text{img}}(\cdot)$ to extract high-level visual features from item images:

$$\mathbf{H}_i^t = \mathcal{F}_{\text{text}}(\mathbf{x}_i^t), \quad \mathbf{H}_i^v = \mathcal{F}_{\text{img}}(\mathbf{x}_i^v). \tag{10}$$

Rather than directly treating modality features as independent item representations, we condition them on item ID embeddings to inject collaborative preference bias. Concretely, given the item ID embedding $\mathbf{E}_i^{(m)}$, we apply a feature-wise gating mechanism:

$$\mathbf{E}_i^{(m)} = \mathbf{E}_i^{id,(m)} \odot \sigma\left(\mathbf{H}_i^{(m)} \mathbf{W}_g^{(m)}\right), \tag{11}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\odot$ is the element-wise product. This operation constrains modality information to modulate ID-based collaborative semantics, instead of forming standalone modality embeddings.

To capture high-order modality-induced relations, we propagate the conditioned item representations on the modality-specific graph $\tilde{\mathcal{A}}^{(m)}$ and aggregate representations from all propagation layers via layer-wise mean pooling:

$$\mathbf{E}_i^{(m,\ell+1)} = \tilde{\mathcal{A}}^{(m)} \mathbf{E}_i^{(m,\ell)}, \mathbf{E}_i^{(m)} = \frac{1}{n+1} \sum_{\ell=0}^{\ell=n} \mathbf{E}_i^{(m,\ell)}, \tag{12}$$

where $\mathbf{E}_i^{(m,0)} = \mathbf{E}_i^{(m)}$. Finally, item-side modality signals are propagated to the user side through the normalized user–item incidence matrix, and we obtain the ID-conditioned modality-specific preference representation over all nodes:

$$\mathbf{E}_u^{(m)} = (\mathcal{D}_u^{-\frac{1}{2}} \mathbf{Y} \mathcal{D}_i^{-\frac{1}{2}}) \mathbf{E}_i^{(m)}, \mathbf{E}^{(m)} = [\mathbf{E}_u^{(m)}; \mathbf{E}_i^{(m)}]. \tag{13}$$

3.3 Modality-Specific Teacher Construction

Based on the extracted preference representations from different views, we further construct a set of modality-specific

teachers to provide complementary supervision signals. Formally, we define a teacher set $\{\mathcal{T}^{(m)}\}_{m\in\mathcal{M}}$, where each teacher corresponds to a distinct preference view. In this paper, $\mathcal{M} = \{ID, V, T\}$, including the collaborative (ID) view and the available content views (V and T).

Teacher parameterization. Each teacher $\mathcal{T}^{(m)}$ is parameterized by a view-specific preference encoder and maps a user–item pair $\langle u, i\rangle$ to preference-aware representations. Importantly, teachers share the same encoder architectures as the student for their corresponding views, ensuring that teacher signals are structurally aligned with student representations.

Specifically, for the collaborative teacher $\mathcal{T}^{ID}$, user and item ID embeddings are propagated on the normalized user–item graph $\tilde{\mathcal{A}}$ using the ID preference encoder described in Eq. (9), yielding collaborative preference representations:

$$\mathbf{E}^{\mathcal{T}^{ID}} = [\mathbf{E}_u^{ID,\mathcal{T}}; \mathbf{E}_i^{ID,\mathcal{T}}]. \tag{14}$$

For each content modality $m \in \{v, t\}$, the modality teacher $\mathcal{T}^{(m)}$ adopts the ID-conditioned modality encoding described in Eqs. (11)– (13). Specifically, raw modality features are first projected and gated by item ID embeddings, propagated on the modality-specific graph $\tilde{\mathcal{A}}^{(m)}$, and then aggregated from the item side to the user side via the normalized incidence matrix $\mathbf{R}$, yields modality-specific preference representations:

$$\mathbf{E}^{\mathcal{T}^{(m)}} = [\mathbf{E}_u^{(m),\mathcal{T}}; \mathbf{E}_i^{(m),\mathcal{T}}]. \tag{15}$$

Teacher outputs. Different modalities encode preference knowledge with modality-specific inductive biases and uneven effectiveness at the representation and ranking levels; therefore, we design each teacher to expose complementary embedding-level and ranking-level outputs.

Given a user–item pair $\langle u, i\rangle$, each teacher provides two types of outputs to guide student learning:

(1) Embedding-level output. Each teacher exposes its preference representations $\mathbf{e}_u^{\mathcal{T}^{(m)}}$ and $\mathbf{e}_i^{\mathcal{T}^{(m)}}$, which correspond to the user-side and item-side embeddings extracted from $\mathbf{E}^{(\mathcal{T}^{(m)})}$. These embeddings convey view-specific semantic information for representation-level distillation.

(2) Rank-level output. Each teacher also induces a ranking preference over a candidate item set $\mathcal{I}$. Specifically, teacher logits are computed as:

$$\pi_{u,j}^{\mathcal{T}^{(m)}} = \mathrm{softmax}((\mathbf{e}_u^{\mathcal{T}^{(m)}})^\top \mathbf{e}_j^{\mathcal{T}^{(m)}})/\tau, \quad j \in \mathcal{I}, \tag{16}$$

$\mathrm{softmax}(\cdot)$ maps the logits over a candidate set into a normalized probability distribution, explicitly modeling relative preference ordering rather than independent scores, and τ smooths the output distribution, helping prevent it from collapsing to a one-hot vector dominated by the top-ranked item.

By modeling teachers as preference extractors specific to each view, we allow the student to selectively absorb transferable knowledge from multiple perspectives on equal footing.

3.4 Student Model Construction

Student parameterization. The student model adopts the same view-specific encoding architectures as the teachers, including the ID preference encoder and the ID-conditioned modality encoders. Given the normalized graphs constructed in Sec. 3.1, the student extracts collaborative and modality-specific preference representations following the same propagation and aggregation procedures:

$$\mathbf{E}^{\mathcal{S}^{ID}} = [\mathbf{E}_u^{ID,\mathcal{S}}; \mathbf{E}_i^{ID,\mathcal{S}}], \mathbf{E}^{\mathcal{S}^{(m)}} = [\mathbf{E}_u^{(m),\mathcal{S}}; \mathbf{E}_i^{(m),\mathcal{S}}]. \tag{17}$$

Modality Fusion with Shared Semantic Component. After extracting collaborative and modality-specific preference representations, we first compute a shared semantic component (C) between content modalities. Specifically, we compute scalar compatibility scores using a shared query network $q(\cdot)$ and obtain normalized weights α_V and α_T via softmax:

$$\alpha_V, \alpha_T = \mathrm{softmax}\left(q(\mathbf{E}^{\mathcal{S}^V}), q(\mathbf{E}^{\mathcal{S}^T})\right). \tag{18}$$

We calculate the shared semantic component as a weighted sum of modality-specific representations. Additionally, we decompose the modality-specific representations into residual components concerning the common content representation:

$$\mathbf{E}^{\mathcal{S}^C} = \alpha_V \mathbf{E}^{\mathcal{S}^V} + \alpha_T \mathbf{E}^{\mathcal{S}^T}, \mathbf{E}_{\mathcal{S}}^{(m),\mathrm{sp}} = \mathbf{E}^{\mathcal{S}^{(m)}} - \mathbf{E}^{\mathcal{S}^C}. \tag{19}$$

These residuals are modulated by collaborative preference representations via sigmoid-based preference gates:

$$\tilde{\mathbf{E}}_{\mathcal{S}}^{(m),\mathrm{sp}} = g_m(\mathbf{E}^{\mathcal{S}^{ID}}) \odot \mathbf{E}_{\mathcal{S}}^{(m),\mathrm{sp}}, \quad m \in \{v, t\}. \tag{20}$$

We further aggregate modality-aware preference representations using mean pooling and combine them with ID preferences to obtain a unified student representation:

$$\mathbf{E}^{\mathcal{S}^{\mathcal{M}}} = \mathrm{Avg}(\mathbf{E}^{\mathcal{S}^C}, \tilde{\mathbf{E}}_S^{V,\mathrm{sp}}, \tilde{\mathbf{E}}_S^{T,\mathrm{sp}}), \mathbf{E}^{\mathcal{S}} = \mathbf{E}^{\mathcal{S}^{ID}} + \mathbf{E}^{\mathcal{S}^{\mathcal{M}}}, \tag{21}$$

where $\mathrm{Avg}(\cdot)$ averages over existing terms and preference scores are computed as $\hat{y}_{u,i} = (\mathbf{e}_u^{\mathcal{S}})^\top \mathbf{e}_i^{\mathcal{S}}$.

Although attention and gating are applied, they do not induce modality dominance. Attention separates modality–content into both shared content and modality–specific components, while gating conditions modality-specific signals based on collaborative context.

3.5 Training and Inference Process

Training Process. While the teachers serve as knowledge providers, the student is optimized end-to-end under recommendation and distillation objectives.

(1) Student Training Objective. Following previous works [Ong and Khong, 2025; Xu *et al.*, 2025], we use ***BPR*** [Rendle *et al.*, 2012] as recommendation loss for the student model, for each training triplet (u, i^+, i^-):

$$\mathcal{L}_{\mathrm{rec}}^{\mathcal{S}} = -\log \sigma\left(\hat{y}_{u,i^+} - \hat{y}_{u,i^-}\right) + \lambda \|\Theta_{\mathcal{S}}\|_2^2, \tag{22}$$

where λ is the L_2 regularization weight. Furthermore, we apply ***InfoNCE*** [Oord *et al.*, 2018] to encourage consistency between different modalities for both users and items:

$$\mathcal{L}_{\mathrm{cl}}^{\mathcal{S}} = \mathrm{InfoNCE}(\mathbf{E}_u^{\mathcal{S}^{\mathcal{M}}}, \mathbf{E}_u^{\mathcal{S}^{ID}}) + \mathrm{InfoNCE}(\mathbf{E}_i^{\mathcal{S}^{\mathcal{M}}}, \mathbf{E}_i^{\mathcal{S}^{ID}}), \tag{23}$$

where $\mathrm{InfoNCE}(\cdot)$ is implemented with cosine similarity.

(2) Multi-Teacher Distillation. We distill knowledge from all available teachers into a unified student via two complementary objectives: rank-level and embedding-level distillation.

(i) Rank-level distillation. For a mini-batch of users, we construct an in-batch candidate set $\mathcal{C}$ as the union of positive and sampled negative items. Given a user u, the student computes logits over $\mathcal{C}$ and induces a soft distribution:

$$p_S(j \mid u) = \mathrm{softmax}(\frac{\hat{y}^{\mathcal{S}}_{u,j}}{\tau}). \tag{24}$$

Each teacher $k \in \{1, \ldots, |\mathcal{M}|\}$ produces its own distribution $p^{(k)}_{\mathcal{T}}(j \mid u)$ using embeddings from the same view. We uniformly ensemble all teachers to form the target distribution:

$$\bar{p}_{\mathcal{T}}(j \mid u) = \frac{1}{|\mathcal{M}|} \sum_{k=1}^{|\mathcal{M}|} p^{(k)}_{\mathcal{T}}(j \mid u). \tag{25}$$

The rank-level distillation objective is defined as

$$\mathcal{L}^{\mathrm{rank}}_{\mathrm{kd}} = \tau^2 \cdot \mathrm{KL}(\bar{p}_T(\cdot \mid u) \parallel p_S(\cdot \mid u)). \tag{26}$$

We use τ^2 as a scale to counteract the gradient shrinkage caused by temperature scaling, thereby preventing the distillation signal from diminishing at high temperatures.

(ii) Embedding-level distillation. We further align student and teacher representations on candidate items. For each view k, let $\mathbf{E}^{\mathcal{S},(k)}_{\mathcal{C}}$ and $\mathbf{E}^{\mathcal{T},(k)}_{\mathcal{C}}$ denote the matrices of student and teacher item embeddings restricted to $\mathcal{C}$, respectively. We minimize a normalized mean squared error:

$$\mathcal{L}^{\mathrm{emb}}_{\mathrm{kd}} = \sum_{k=1}^{|\mathcal{M}|} \left\| \mathrm{norm}\left(\mathbf{E}^{\mathcal{S},(k)}_{\mathcal{C}}\right) - \mathrm{norm}\left(\mathbf{E}^{\mathcal{T},(k)}_{\mathcal{C}}\right) \right\|^2_2, \tag{27}$$

where $\mathrm{norm}(\cdot)$ denotes ℓ_2-normalization to mitigate scale mismatch across teachers.

(3) Teacher Online Update Process. Each teacher is designed to preserve a *modality-specialized* preference function. To this end, we adopt a two-stage online update scheme that ensures each teacher remains a strong specialist for its own view, while being continuously aligned with the student's evolving cross-modal representations.

(i) Single-modality preference learning. For the k-th teacher, we optimize its view-specific encoder using the Bayesian Personalized Ranking (BPR) objective on observed training triplets (u, i^+, i^-):

$$\mathcal{L}^{\mathcal{T}}_{\mathrm{rec}} = \frac{1}{|\mathcal{M}|} \sum_{k=1}^{|\mathcal{M}|} -\log \sigma(\hat{y}^{(k)}_{u,i^+} - \hat{y}^{(k)}_{u,i^-}), \tag{28}$$

where $\hat{y}^{(k)}_{u,i}$ denotes the prediction produced by the k-th teacher under its corresponding modality. This objective enables each teacher to explicitly capture modality-specific preference cues, avoiding premature dominance by ID-based collaborative signals during multi-view learning.

(ii) Cross-modal synchronization. After updating the teacher with $\mathcal{L}^{\mathcal{T}}_{\mathrm{rec}}$, we further synchronize it with the student via an exponential moving average (EMA) [Tarvainen and Valpola, 2017]. Let $\Theta^{(k)}_{\mathcal{T}}$ denote the parameters of the k-th teacher and $\Theta^{(k)}_{\mathcal{S}}$ the corresponding parameter subset of the student. The update rule is given by:

$$\Theta^{(k)}_{\mathcal{T}} \leftarrow \mu\, \Theta^{(k)}_{\mathcal{T}} + (1-\mu)\, \Theta^{(k)}_{\mathcal{S}}, \tag{29}$$

where $\mu \in (0, 1)$ is the momentum coefficient. EMA acts as a temporally smoothed information bridge that injects the student's cross-modal knowledge into the teacher and stabilizes teacher representations across mini-batches.

The overall training loss is:

$$\mathcal{L} = \mathcal{L}^{\mathcal{S}}_{\mathrm{rec}} + \lambda_{\mathrm{cl}}\mathcal{L}^{\mathcal{S}}_{\mathrm{cl}} + \mathcal{L}^{\mathcal{T}}_{rec} + \lambda_{\mathrm{rank}}\mathcal{L}^{\mathrm{rank}}_{\mathrm{kd}} + \lambda_{\mathrm{emb}}\mathcal{L}^{\mathrm{emb}}_{\mathrm{kd}}. \tag{30}$$

Inference Process. At inference time, we discard all teachers and only run the student forward pass to obtain $\mathbf{e}^{\mathcal{S}}_u$ and $\mathbf{e}^{\mathcal{S}}_i$, and rank items by:

$$\hat{y}_{u,i} = (\mathbf{e}^{\mathcal{S}}_u)^\top \mathbf{e}^{\mathcal{S}}_i. \tag{31}$$

Dataset	Baby		Sports		Clothing		Elec		MicroLens	
Modality	V	T	V	T	V	T	V	T	V	T
Embed Dim	4,096	384	4,096	384	4,096	384	4,096	384	4,096	384
# Users	19,445		35,598		39,387		192,403		98,129	
# Items	7,050		18,357		23,033		63,001		17,228	
# Interactions	160,792		296,337		278,677		1,689,188		705,174	
Sparsity	99.88%		99.95%		99.97%		99.99%		99.96%	

Table 1: Statistics of the datasets.

4 Experiments and Results

4.1 Experimental Settings

Dataset. To comprehensively and fairly evaluate the models' effectiveness, we conducted experiments using five publicly available datasets encompassing a variety of recommendation scenarios (such as movies) and different densities from the Amazon review [Lakkaraju *et al.*, 2013] and a video platform [Ni *et al.*, 2025]. We select five datasets of varying sizes: `Baby`, `Sports`, `Clothing`, `Electronics`, and `MicroLens`.

Metric. To ensure a fair comparison, we align our approach with the settings used in prior studies [Zhou and Shen, 2023; Ong and Khong, 2025]. We utilize two widely accepted evaluation metrics for top-K recommendation: Recall@K and NDCG@K. We present the average scores for all users in the test dataset for both K=10 and K=20.

Baselines. We compare MTS^{2}4MM against several state-of-the-art (SOTA) recommender models. **(1) General Model:** The following SOTA models include the matrix factorization (MF, BPR) [Rendle *et al.*, 2012] and a graph-based model LightGCN [He *et al.*, 2020], which are chosen for comparison. **(2) Multimodal model:** To ensure robust evaluation of the proposed model, several state-of-the-art multimodal recommendation systems have been selected for comparison, including the matrix factorization model (**VBPR** [He and McAuley, 2016]) and various graph-based models (**MMGCN** [Wei *et al.*, 2019], **GRCN** [Wei *et al.*, 2020], **DualGNN** [Wang *et al.*, 2023], **SLMRec** [Tao *et al.*, 2022], **BM3** [Zhou *et al.*, 2023b], **MGCN** [Yu *et al.*, 2023], **FREEDOM** [Zhou and Shen, 2023], **LGMRec** [Guo *et al.*, 2024], **DAMRS** [Xv *et al.*, 2024], **SMORE** [Ong and Khong, 2025]), **COHESION** [Xu *et al.*, 2025].

Dataset	Metric	General Model		Multimodal Model														Imp.(%)
		BPR UAI'09	LightGCN SIGIR'19	VBPR AAAI'16	MMGCN MM'19	GRCN MM'20	DualGNN TMM'21	SLMRec TMM'22	BM3 WWW'23	MGCN MM'23	FREEDOM MM'23	LGMRec AAAI'24	DAMRS KDD'24	PGL AAAI'25	SMORE WSDM'25	COHESION SIGIR'25	MTS^{2}4MM Ours	
Baby	R@10	0.0357	0.0479	0.0423	0.0378	0.0539	0.0448	0.0529	0.0564	0.0631	0.0627	0.0630	0.0612	0.0601	0.0655	0.0655	**0.0694**	+5.95%
	R@20	0.0575	0.0754	0.0663	0.0615	0.0833	0.0716	0.0775	0.0883	0.0960	0.0992	0.0957	0.0944	0.0956	0.1027	0.1034	**0.1093**	+5.71%
	N@10	0.0192	0.0257	0.0223	0.0200	0.0288	0.0240	0.0290	0.0301	0.0344	0.0330	0.0342	0.0326	0.0328	0.0357	0.0349	**0.0381**	+6.72%
	N@20	0.0249	0.0328	0.0284	0.0261	0.0363	0.0309	0.0353	0.0383	0.0429	0.0424	0.0426	0.0411	0.0419	0.0453	0.0446	**0.0482**	+6.40%
Sports	R@10	0.0432	0.0569	0.0558	0.0370	0.0598	0.0568	0.0663	0.0656	0.0723	0.0717	0.0662	0.0689	0.0681	0.0735	0.0715	**0.0772**	+5.03%
	R@20	0.0653	0.0864	0.0856	0.0605	0.0915	0.0859	0.0990	0.0980	0.1093	0.1089	0.1012	0.1022	0.1026	0.1103	0.1094	**0.1165**	+5.62%
	N@10	0.0241	0.0311	0.0307	0.0193	0.0332	0.0310	0.0365	0.0355	0.0395	0.0385	0.0357	0.0372	0.0367	0.0401	0.0391	**0.0425**	+5.99%
	N@20	0.0298	0.0387	0.0384	0.0254	0.0414	0.0385	0.0450	0.0438	0.0490	0.0481	0.0447	0.0458	0.0456	0.0496	0.0488	**0.0524**	+5.65%
Elec	R@10	0.0235	0.0363	0.0293	0.0213	0.0389	0.0365	0.0443	0.0437	0.0424	0.0396	0.0433	0.0408	0.0394	0.0423	0.0408	**0.0475**	+7.22%
	R@20	0.0367	0.0540	0.0458	0.0343	0.0590	0.0542	0.0651	0.0648	0.0626	0.0601	0.0639	0.0610	0.0597	0.0631	0.0628	**0.0694**	+6.61%
	N@10	0.0127	0.0204	0.0159	0.0112	0.0216	0.0206	0.0249	0.0247	0.0237	0.0220	0.0242	0.0227	0.0219	0.0236	0.0228	**0.0267**	+7.23%
	N@20	0.0161	0.0250	0.0202	0.0146	0.0268	0.0252	0.0303	0.0302	0.0289	0.0273	0.0295	0.0279	0.0271	0.0290	0.0265	**0.0322**	+6.27%
Clothing	R@10	0.0202	0.0351	0.0343	0.0235	0.0440	0.0251	0.0468	0.0443	0.0646	0.0646	0.0518	0.0601	0.0545	0.0654	0.0594	**0.0683**	+4.43%
	R@20	0.0301	0.0525	0.0530	0.0376	0.0660	0.0392	0.0707	0.0651	0.0950	0.0940	0.0776	0.0893	0.0807	0.0967	0.0877	**0.1017**	+5.17%
	N@10	0.0113	0.0195	0.0185	0.0122	0.0230	0.0132	0.0254	0.0242	0.0355	0.0347	0.0280	0.0322	0.0294	0.0355	0.0318	**0.0371**	+4.51%
	N@20	0.0139	0.0239	0.0232	0.0157	0.0286	0.0168	0.0315	0.0295	0.0432	0.0422	0.0346	0.0396	0.0361	0.0435	0.0390	**0.0456**	+4.83%
MicroLens	R@10	0.0603	0.0728	0.0682	0.0459	0.0727	0.0556	0.0762	0.0601	0.0719	0.0635	0.0704	0.0733	0.0705	0.0727	0.0628	**0.0815**	+6.96%
	R@20	0.0975	0.1079	0.1033	0.0746	0.1128	0.0832	0.1163	0.0966	0.1105	0.0999	0.1077	0.1112	0.1079	0.1116	0.1002	**0.1241**	+6.71%
	N@10	0.0308	0.0382	0.0351	0.0230	0.0379	0.0312	0.0393	0.0308	0.0368	0.0324	0.0366	0.0380	0.0364	0.0370	0.0326	**0.0416**	+5.85%
	N@20	0.0403	0.0473	0.0441	0.0304	0.0482	0.0402	0.0496	0.0401	0.0467	0.0417	0.0462	0.0478	0.0460	0.0470	0.0435	**0.0526**	+6.05%

Table 2: The overall performance comparison results of applying our model and baselines on five real-world datasets. The result is calculated based on the mean of five repetitions with different random seeds for all models.

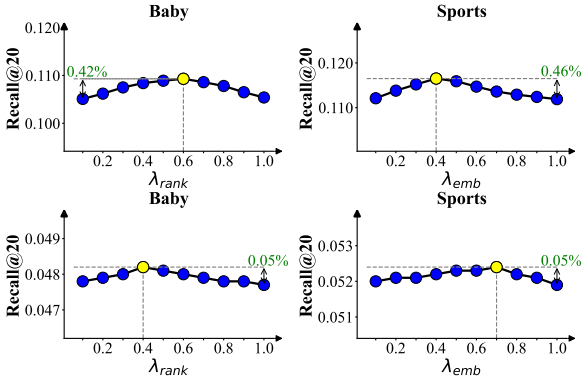

Figure 3: Sensitivity analysis of the distillation weights.

	Baby		Sports	
	R@20	N@20	R@20	N@20
w/o KD$_{\text{rank}}$	0.1062	0.0463	0.1125	0.0510
w/o KD$_{\text{emb}}$	0.1085	0.0475	0.1142	0.0518
w/o mg	0.0983	0.0451	0.1135	0.0506
w/o all	0.0963	0.0425	0.1095	0.0495
Original	**0.1093**	**0.0482**	**0.1165**	**0.0524**

Table 3: Ablation study of MTS^{2}4MM on Baby and Sports. "KD$_{\text{rank}}$" refers to rank-level distillation, "KD$_{\text{emb}}$" to embedding-level distillation, and "mg" to the modality-preference gating module. w/o all removes all three components and trains the student using independent BPR objectives for each modality.

Implementation Details. To ensure a fair comparison, we utilized the unified open-source MMRec framework [Zhou, 2023] to develop our proposed model and replicate existing recommendation systems. For each chosen baseline, the hyperparameters were adjusted according to the optimal configurations reported in the relevant published papers. Further details can be found in the supplementary material file.

4.2 Overall Performance

The comparison between MTS^{2}4MM and baselines is shown in Table 2. In summary, we have the following observations: **(1) Overall Effectiveness.** The results show that MTS^{2}4MM consistently outperforms all baselines in terms of Recall and NDCG across five real-world datasets. The stable improvements indicate its superior ability to capture user preference signals and recommend relevant items. **(2) Comparison with Fusion-Based Multimodal Models.** State-of-the-art multimodal baselines (e.g., SMORE) achieve competitive performance by integrating multiple modalities. However, their gains are often limited by modality imbalance and ineffective fusion, where dominant collaborative signals overshadow auxiliary modalities. In contrast, MTS^{2}4MM consistently achieves higher performance, with relative improvements of approximately 4%–7%, highlighting the limitations of direct fusion strategies.

4.3 Hyper-parameter Analysis

Figure 3 analyzes the sensitivity of MTS^{2}4MM to the distillation weights λ_{rank} and λ_{emb}. We observe that moderate values of both hyper-parameters consistently yield the best performance across datasets, while overly small or large weights lead to performance degradation. This trend indicates that effective multimodal learning requires a balanced integration of ranking-level preference transfer and embedding-level representation alignment. Overall, MTS^{2}4MM shows stable performance within a broad range of hyper-parameter settings, demonstrating its robustness to hyper-parameter choices.

4.4 Ablation Study

The ablation results of MTS^{2}4MM are reported in Table 3. We summarize the key observations as follows: **(1) Effectiveness of multi-teacher distillation.** Removing either rank-level distillation (KD$_{\text{rank}}$) or embedding-level distillation (KD$_{\text{emb}}$) consistently degrades performance, indicating that preserving both preference ordering and representation alignment is crucial for effective transfer of modality-specific knowledge. **(2) Importance of modality-aware student modeling.** Eliminating the modality-preference gating module or all components leads to substantial performance drops, demonstrating that explicitly regulating modality contributions in the student is essential for mitigating modality imbalance and achieving optimal recommendation performance.

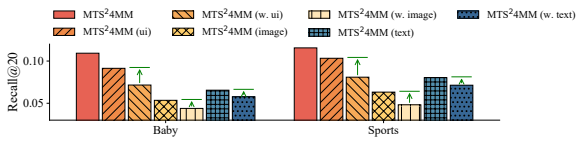

Figure 4: Modality contribution analysis.

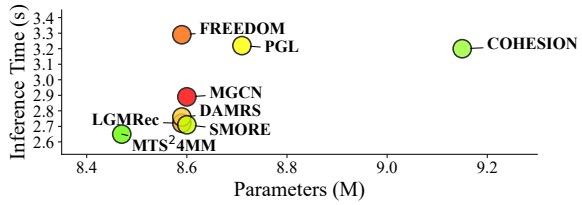

Figure 5: Comparison of model efficiency against baseline models.

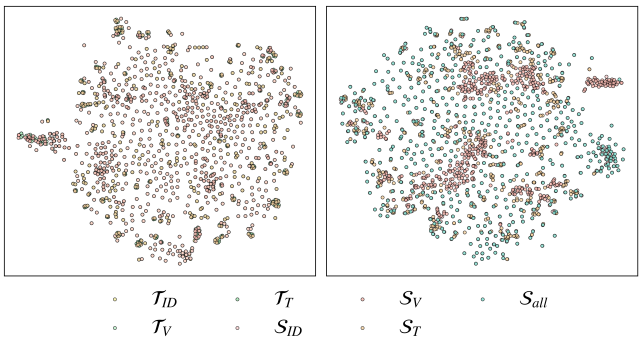

Figure 6: Visualization of Student and Teacher Model Embeddings.

4.5 Case Study

Figure 4 compares MTS^{2}4MM with its modality-specific variants on the Baby and Sports datasets. Models relying on a single modality, either during inference or throughout both training and inference, consistently underperform the full model, indicating that no individual modality is sufficient for accurate preference modeling. In contrast, MTS^{2}4MM achieves substantial improvements over all single-modality variants, with relative gains of up to 27.7% on Baby and 31.1% on Sports. These results highlight both the presence of modality imbalance and the effectiveness of the proposed multi-teacher distillation framework in selectively integrating complementary modality signals.

4.6 Time Complexity Analysis

We analyze the time complexity of MTS^{2}4MM by its main components. For the ID-based user–item graph, one GCN layer costs $\mathcal{O}(|\mathcal{E}_{ui}|\cdot d/B)$, where $|\mathcal{E}_{ui}|$ is the number of edges, d is the embedding dimension, and B is the batch size. With L_{ui} layers, the total cost is $\mathcal{O}(L_{ui}|\mathcal{E}_{ui}|d/B)$. For each content modality m, propagation on the item–item graph costs $\mathcal{O}(L_m|\mathcal{E}_{ii}^{(m)}|d)$. The BPR objective incurs a cost of $\mathcal{O}(2dB)$ per mini-batch. Rank- and embedding-level distillation introduce a linear overhead $\mathcal{O}(dB|\mathcal{C}|)$, where $|\mathcal{C}|$ is the in-batch candidate size. Overall, as shown in Figure 5, MTS^{2}4MM maintains *lower* computation and storage requirements than competitive baselines such as COHESION, and SMORE.

4.7 t-SNE Visualization Analysis

Figure 6 presents t-SNE visualizations of the embeddings for both the teacher and student models across different modalities. The left panel shows individual modality embeddings for the Teacher and Student models, while the right panel displays the fused embeddings ($\mathcal{S}_{all}$) of the Student model after modality integration, we have the following key observations: ⟨1⟩ The student model successfully integrates modality-specific representations into a shared space without any single modality dominating, indicating balanced fusion; ⟨2⟩ Both teacher and student models show well-aligned embeddings, demonstrating effective cross-modal knowledge transfer through EMA; ⟨3⟩ Randomly selecting a user's two positive samples and one negative sample, the t-SNE visualization confirms that the model can effectively distinguish between positive and negative samples, highlighting its ability to learn meaningful representations.

5 Related Work

Early approaches treat modality features as auxiliary signals, while recent methods emphasize modality-specific modeling and graph-based integration to capture high-order relations. Self-supervised objectives are widely adopted to improve cross-view consistency and robustness to noisy or incomplete modalities, and advanced designs further mitigate fusion bias through multi-stage fusion and graph or spectral regularization, as exemplified by COHESION [Xu *et al.*, 2025], SMORE [Ong and Khong, 2025], and PGL [Yu *et al.*, 2025]. However, these fusion-centric paradigms tightly couple modalities within unified architectures, implicitly assuming balanced and reliable modality contributions, which often leads to dominant-modality bias and limited controllability in practice. In contrast, *multimodal distillation* reframes multimodal learning as a knowledge transfer problem, enabling flexible and efficient utilization of heterogeneous expertise. Although this direction remains under-explored in recommender systems, recent studies such as MMKD (PromptMM) [Wei *et al.*, 2024] and multi-teacher routing or consistency-based distillation methods [Sun *et al.*, 2024; Feng *et al.*, 2025] demonstrate its potential. Our work addresses the critical gap by systematically modeling modality-specific preference knowledge and enabling selective, controllable transfer via multi-teacher online distillation.

6 Conclusion

In this paper, we proposed MTS^{2}4MM, a multi-teacher single-student online distillation framework that reframes multimodal recommendation as a selective knowledge transfer problem. By treating each modality as an independent preference teacher and distilling both ranking and representation knowledge into a unified student, MTS^{2}4MM effectively mitigates modality imbalance and avoids dominant-modality bias. These results highlight the promise of knowledge-driven multimodal learning as a robust alternative to conventional fusion-based recommendation paradigms.

Acknowledgements

This work is supported by the Natural Science Foundation of China under Grant No. 92567204 and No. 62472196.

References

[Feng *et al.*, 2025] Kaidong Feng, Zhu Sun, Hui Fang, Jie Yang, Wenyuan Liu, and Yew-Soon Ong. Routing distilled knowledge via mixture of lora experts for large language model based bundle generation, 2025.

[Guo *et al.*, 2024] Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. Lgmrec: local and global graph learning for multimodal recommendation. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024.

[He and McAuley, 2016] Ruining He and Julian McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[He *et al.*, 2020] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 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*, pages 639–648, 2020.

[He *et al.*, 2023] Li He, Xianzhi Wang, Dingxian Wang, Haoyuan Zou, Hongzhi Yin, and Guandong Xu. Simplifying graph-based collaborative filtering for recommendation. In *Proceedings of the sixteenth ACM international conference on web search and data mining*, pages 60–68, 2023.

[Lakkaraju *et al.*, 2013] Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. What's in a name? understanding the interplay between titles, content, and communities in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 311–320, 2013.

[Mao *et al.*, 2021] Kelong Mao, Jieming Zhu, Xi Xiao, Biao Lu, Zhaowei Wang, and Xiuqiang He. Ultragcn: Ultra simplification of graph convolutional networks for recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 1253–1262, New York, NY, USA, 2021. Association for Computing Machinery.

[Ni *et al.*, 2025] Yongxin Ni, Yu Cheng, Xiangyan Liu, Junchen Fu, Youhua Li, Xiangnan He, Yongfeng Zhang, and Fajie Yuan. A content-driven micro-video recommendation dataset at scale. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6486–6491, 2025.

[Ong and Khong, 2025] Rongqing Kenneth Ong and Andy WH Khong. Spectrum-based modality representation fusion graph convolutional network for multimodal recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 773–781, 2025.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

[Sun *et al.*, 2024] Wenqi Sun, Ruobing Xie, Junjie Zhang, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. Curriculum-scheduled knowledge distillation from multiple pre-trained teachers for multi-domain sequential recommendation, 2024.

[Tao *et al.*, 2022] Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:5107–5116, 2022.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2023] Qifan Wang, Yinwei Wei, Jianhua Yin, Jianlong Wu, Xuemeng Song, and Liqiang Nie. Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:1074–1084, 2023.

[Wei *et al.*, 2019] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445, 2019.

[Wei *et al.*, 2020] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3541–3549, 2020.

[Wei *et al.*, 2024] Wei Wei, Jiabin Tang, Yangqin Jiang, Lianghao Xia, and Chao Huang. Promptmm: Multi-modal knowledge distillation for recommendation with prompt-tuning, 2024.

[Xu *et al.*, 2025] Jinfeng Xu, Zheyu Chen, Wei Wang, Xiping Hu, Sang-Wook Kim, and Edith CH Ngai. Cohesion: Composite graph convolutional network with dual-stage fusion for multimodal recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1830–1839, 2025.

[Xv *et al.*, 2024] Guipeng Xv, Xinyu Li, Ruobing Xie, Chen Lin, Chong Liu, Feng Xia, Zhanhui Kang, and Leyu Lin. Improving multi-modal recommender systems by denoising and aligning multi-modal content and user feedback. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page

3645–3656, New York, NY, USA, 2024. Association for Computing Machinery.

[Yu *et al.*, 2023] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6576–6585, 2023.

[Yu *et al.*, 2025] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. Mind individual information! principal graph learning for multimedia recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12):13096–13105, Apr. 2025.

[Zhou and Shen, 2023] Xin Zhou and Zhiqi Shen. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM international conference on multimedia*, pages 935–943, 2023.

[Zhou *et al.*, 2023a] Xin Zhou, Aixin Sun, Yong Liu, Jie Zhang, and Chunyan Miao. Selfcf: A simple framework for self-supervised collaborative filtering. *ACM Trans. Recomm. Syst.*, 1(2), June 2023.

[Zhou *et al.*, 2023b] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM web conference 2023*, pages 845–854, 2023.

[Zhou, 2023] Xin Zhou. Mmrec: Simplifying multimodal recommendation. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, pages 1–2, 2023.