FreDRec: Frequency-Decoupled Knowledge Distillation for Multimodal Recommendation in Missing Modalities Scenarios

Qixian Chen[1], **Xiaolong Xu**[1*], **Haolong Xiang**[1], **Kun Yi**[3], **Xiaoyu Xia**[4], **Lianyong Qi**[5], **Wei Fan**[6]

[1]School of Software, Nanjing University of Information Science and Technology, China
[3]State Information Center, China [4]School of Computing Technologies, RMIT University, Australia
[5]College of Computer Science and Technology, China University of Petroleum (East China), China
[6]School of Computer Science, University of Auckland, New Zealand

Abstract

Multimodal Recommender Systems (MRSs) have achieved significant success in personalized recommendations by integrating diverse modalities such as images, text, and audio. While in complex real-world industrial scenarios where MRS has usually been deployed, the presence of missing modalities is very common. Most existing MRS methods recover missing data by disentangling available modality features into general and specific representations, thereby generating missing modalities based on personalized aligned feature. However, these methods rely on implicit constraints within pre-extracted multimodal feature spaces, which usually introduce noise that prevents the effective decoupling of general and specific semantics, thereby diminishing recommendation accuracy. Moreover, unguided linear transformations of available modality features for generation lead to semantic hallucinations, which inevitably degrade the ranking quality of the recommendation results. To address these issues, we propose the Frequency-Decoupled Knowledge Distillation Framework for Multimodal Recommendation (**FreDRec**), which conducts effective and robust missing feature generation. Specifically, motivated by the physical characteristics of multimodal signals, FreDRec utilizes Fourier and Discrete Wavelet Transforms to explicitly decouple raw multimodal signals into low-frequency core semantics and high-frequency details, which is mathematically proven to effectively minimize noise introduction at the source. Next, to ensure robust feature generation, we pre-train a teacher model based on decoupled multi-view graph propagation to guide a student model via multi-level knowledge distillation, which establishes an end-to-end highly efficient lightweight architecture and achieves precise missing modality generation with expert collaborative bounds. Finally, extensive experiments on multiple public benchmark datasets demonstrate that FreDRec achieves superior performance under various modality missing ratios.

1 Introduction

Multimodal Recommender Systems (MRSs) enhance item representations by integrating diverse modalities such as images, text and audio, which significantly outperform traditional ID-based collaborative filtering models [19, 7, 24, 8]. To capture high-order collaborative signals, multimodal graph recommenders have become the main method, constructing user-item bipartite graphs to explicitly mine latent structural dependencies from modality features [27, 26, 37, 6, 35, 36].

However, most existing methods are based on the idealistic assumption of completeness of the modality. In real-world scenarios, modalities are often missing due to data collection flaws, transmission errors, or content sparsity [23, 29, 5]. Consequently, conventional models suffer substantial performance degradation under missing modality, long-tailed, or cold-start scenarios [13, 30, 32, 16].

*Corresponding author: xlxu@nuist.edu.cn

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Current research tackles this challenge primarily via invariant learning and modality generation. The former uses information bottlenecks or invariant strategies to extract robust, incompleteness-insensitive preference features [2, 3]. The latter reconstructs the missing modality information through generative modeling [4, 14, 12]. In particular, disentanglement-based generation has emerged as a prevalent solution [11], which decouples available features into shared and specific components to cross-modally infer missing representations. Despite these initial successes, most current disentanglement-based generation methods still suffer from the following fundamental limitations.

• **Noise Entangled Feature Disentanglement:** To mitigate noise, recent methods decouple multimodal features into modality-general and modality-specific representations [11]. However, these methods typically perform disentanglement on pre-extracted feature space. Operating on previously extracted semantic spaces is suboptimal since environmental noise becomes intrinsically coupled during initial extraction. Moreover, enforcing separation via contrastive learning or mutual information lacks physical interpretability and fails to guarantee absolute representation purity.

• **Unconstrained Generation without Prior Guidance:** During the generative phase, most existing methods aim to reconstruct missing representations by merely minimizing latent feature-level reconstruction losses. However, lacking expert knowledge or collaborative structural regularization, this unguided process easily degenerates into blind generation. Consequently, to satisfy optimization objectives, models become highly susceptible to producing semantic hallucinations that yield misleading artifacts instead of authentic signals, which ultimately degrades recommendation robustness.

To address these fundamental issues, we propose FreDRec, frequency-decoupled knowledge distillation framework for incomplete multimodal recommendation. It comprehensively comprises three core modules: Physical-Level Frequency Decoupling, Multi-View Graph Learning, and Multi-Level Knowledge Distillation. A robust teacher model is first constructed using the initial two modules under complete modalities. Subsequently, the distillation module effectively transfers this expert knowledge to a lightweight student network, ensuring robust recommendation under missing modalities. Through this elegant paradigm, FreDRec achieves highly accurate and semantically constrained representations for items with missing modalities, without compromising strict low-latency deployment requirements. In summary, the primary contributions of this work are outlined as follows:

- We propose FreDRec, a novel Frequency Decoupled Knowledge Distillation framework adopting a teacher student paradigm. By synergistically combining physical level frequency domain disentanglement with global collaborative priors, our approach explicitly isolates robust core semantics and specific fine details directly from raw multimodal signals. This fundamental separation effectively prevents the blind generation of missing modalities, successfully eliminating semantic hallucinations and ensuring highly accurate feature reconstruction.
- We provide a rigorous mathematical derivation from a manifold geometry perspective to prove the optimality of our physical-level decoupling. Our analysis demonstrates that this approach effectively suppresses noise and representation drift at the source, achieving a strictly tighter error bound compared to the prevalent mutual information-based disentanglement methods.
- We develop a lightweight end to end student model specifically optimized for highly complex real world industrial scenarios. This architecture satisfies stringent online inference deployment constraints and strict efficiency requirements while maintaining consistently robust performance, achieving a precise balance between model complexity and recommendation efficiency.
- We conducted extensive experiments on multiple real-world benchmark datasets. The results demonstrate that FreDRec consistently and significantly outperforms state-of-the-art baselines under various modality missing ratios. The source code is available at `https://github.com/Maplefallin/FreDRec.git`.

2 Related Work

Graph-based Multimodal Recommendation. Traditional collaborative filtering methods like BPR [19] provide foundational user-item modeling but struggle to capture complex high-order connectivity. To address this limitation, Graph Neural Networks such as NGCF [24] and LightGCN [8], alongside self-supervised contrastive paradigms like SGL [28] and SimGCL [34], have been proposed to explicitly model high-order structural signals. Concurrently, these paradigms have been extended to

tackle extreme sparsity, including user cold-start via C2LRec [31] and CMCLRec [30], as well as long-tailed behaviors through HPSERec [32] and CLLMRec [15]. To further enrich isolated item representations, integrating multimodal features has become standard practice, evolving from early matrix factorization approaches like VBPR [7] to sophisticated multimodal graph networks. For instance, MMGCN [27] constructs modality-specific graphs, while GRCN [26] adaptively refines interaction structures. Recent advancements further boost performance through latent structure mining frameworks like LATTICE [37] and MGCN [35], alongside topology aggregation algorithms such as LGMRec [6] and GUME [13]. However, these advanced methods implicitly assume complete modality availability and treat pre-extracted multimodal features as noise-free inputs. Consequently, they remain vulnerable to semantic hallucinations and severe performance degradation when deployed in real-world scenarios plagued by noisy or missing modalities.

Recommendation under Missing Modalities. To counteract the performance degradation caused by incomplete data, handling missing modalities has emerged as a critical research focus. Early approaches mitigate this issue by learning robust modality-specific representations [23]. Subsequent studies have explored unified single-branch networks [5] and contrastive cross-modal generation frameworks. Notably, Lin et al. proposed CI2MG [14], which reconstructs missing modalities by translating available features across modal boundaries. To bridge the semantic gap between complete and incomplete data, knowledge distillation has also proven highly effective. Existing methods leverage distillation to transfer privileged representations from full-modality to missing-modality scenarios, demonstrating strong performance in visual question answering [4] and general representation learning [1, 10]. More recently, state-of-the-art methods have shifted toward robust feature generation and denoising. Specifically, Chen et al. proposed I3-MRec [3], which utilizes information bottlenecks for invariant learning, while Kim et al. introduced DGMRec [11], employing disentangled generation to impute absent signals. Following this trajectory, Zhuang et al. developed TMDC [38], which incorporates a two-stage denoising and complementation framework.

However, these methods overlook two critical aspects: (1) how to perform disentanglement at the explicit physical level to guarantee absolute signal purity rather than relying on already-noisy pre-extracted features, and (2) how to incorporate expert structural guidance into the cross-modal generation process to prevent blind generation, thereby avoiding severe semantic hallucinations.

3 Problem Definition

Let $\mathcal{U}$ and $\mathcal{I}$ denote the user and item sets, with historical interactions represented by a binary matrix $Y \in \{0,1\}^{|\mathcal{U}|\times|\mathcal{I}|}$, where $y_{ui} = 1$ indicates an observed interaction. Each item $i \in \mathcal{I}$ is associated with a set of modalities $\mathcal{M}$ featuring raw vectors $x_{i,m} \in \mathbb{R}^{d_m}$. To formulate the missing modality problem, we introduce an indicator matrix $O \in \{0,1\}^{|\mathcal{I}|\times|\mathcal{M}|}$, where $o_{i,m} = 1$ if modality m for item i is available, and 0 otherwise.

Our goal is to learn a robust recommendation model $f_\theta(\cdot)$ under knowledge distillation that maps users and items into a d-dimensional latent space ($e_u, e_i \in \mathbb{R}^d$). The matching probability $\hat{y}_{ui}$ for unobserved interactions is estimated as:

$$\hat{y}_{ui} = f_\theta(u, i|Y, \{X_m\}_{m\in\mathcal{M}}, O) = e_u^\top e_i. \tag{1}$$

Finally, the top-K items with the highest predicted scores $\hat{y}_{ui}$ are recommended to user u.

4 Methodology

4.1 Overview

This section introduces the FreDRec framework, which is designed to resolve semantic representation hallucinations and deep noise entanglement caused by missing modalities in MRSs. As illustrated in Figure 1, FreDRec mainly comprises three core modules.

First, the **Physical-Level Frequency Decoupling** module (Section 4.2) employs spectral transforms and attention mechanisms to explicitly separate raw multimodal signals into low-frequency core semantics and high-frequency details, filtering redundant noise at the source. Second, the **Multi-View Graph Learning** module (Section 4.3) constructs a robust teacher network under complete modality conditions. It captures global collaborative priors to generate high-quality expert representations

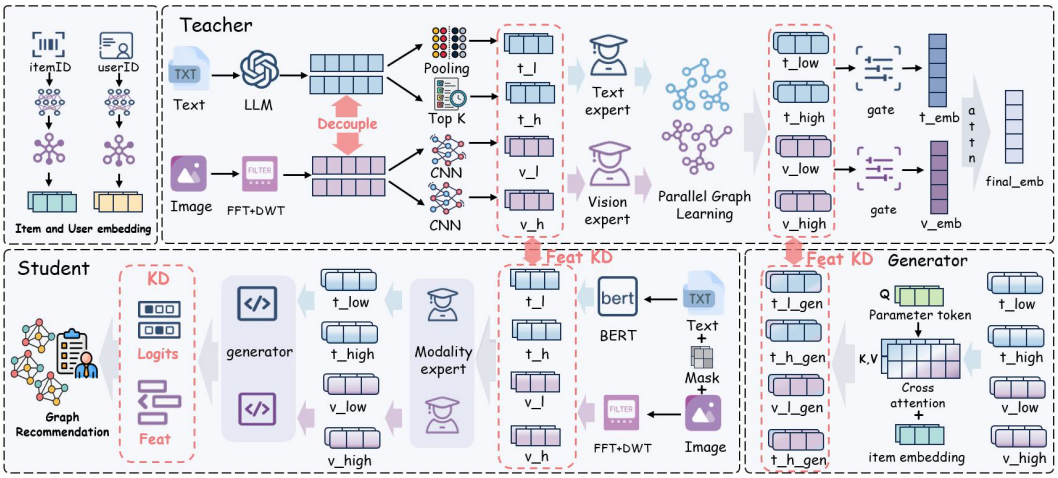


Figure 1: The architecture of FreDRec. The Physical-Level Frequency Decoupling module decouple raw signals to eliminate noise. The Multi-View Graph Learning module constructs a teacher network to generate expert representations. The Multi-Level Knowledge Distillation module trains a lightweight student network to generate missing modalities.

via decoupled graph propagation. Third, the **Multi-Level Knowledge Distillation** module (Section 4.4) tackles missing of modality using an asymmetric generation strategy. It leverages deterministic mapping for low-frequency bases and Cross-Attention retrieval for high-frequency details, strictly bounded by a hierarchical distillation objective to eliminate semantic hallucinations. Finally, the sequential expert-guided training strategy is elaborated in Section 4.5.

4.2 Physical-Level Frequency Decoupling for Multimodal Signal

To address noise pollution and semantic entanglement in multimodal features, FreDRec introduces an asymmetric frequency-domain decoupling architecture. This module decouples raw data into Low-Frequency Core Semantics and High-Frequency Specific Features, mapping them into a unified collaborative space.

Vision Modality Decoupling. Let $\boldsymbol{I}_v \in \mathbb{R}^{C\times H\times W}$ be the input image tensor. To filter global noise, we apply an Adaptive Frequency Filter [17]. It first transforms the spatial image into the frequency spectrum $\boldsymbol{S}(u,v)$ via 2D Fast Fourier Transform (FFT). The image is then modulated by a learnable complex weight matrix $\boldsymbol{W}_{complex}$ and reconstructed back to the spatial domain via 2D Inverse Fast Fourier Transform (IFFT). The denoised image $\tilde{\boldsymbol{I}}_v(x,y)$ at spatial coordinates (x,y) is rigorously formulated as:

$$\tilde{\boldsymbol{I}}_v(x,y) = \boldsymbol{I}_v(x,y) + \frac{1}{HW}\sum_{u=0}^{H-1}\sum_{v=0}^{W-1}\left(\boldsymbol{S}(u,v)\odot \boldsymbol{W}_{complex}(u,v)\right)e^{j2\pi\left(\frac{ux}{H}+\frac{vy}{W}\right)}, \tag{2}$$

where u,v represent the frequency coordinates, $\odot$ denotes the Hadamard product, and j is the imaginary unit.

Subsequently, a 2D Discrete Wavelet Transform (DWT) decomposes $\tilde{\boldsymbol{I}}_v$ using a low-pass scaling filter g and high-pass directional wavelet filters ψ_c. The low-frequency core semantics $\boldsymbol{V}_{LL}$ and high-frequency directional details $\boldsymbol{V}^c_{HF}$ are extracted via 2D strided convolutions:

$$\boldsymbol{V}_{LL}(p,q) = \sum_m\sum_n g[2p-m]g[2q-n]\tilde{\boldsymbol{I}}_v(m,n), \tag{3}$$

$$\boldsymbol{V}^c_{HF}(p,q) = \sum_m\sum_n \psi_c[2p-m,2q-n]\tilde{\boldsymbol{I}}_v(m,n), \tag{4}$$

where $c\in\{LH,HL,HH\}$ represents horizontal, vertical, and diagonal subbands. We then use a ConvNeXt-based extractor for $\boldsymbol{V}_{LL}$ and a Dilated Inverted Bottleneck for $\boldsymbol{V}^c_{HF}$ to expand the receptive field. These processed features are ultimately linearly projected into a d-dimensional space to obtain the visual decoupled vectors $\boldsymbol{v}_{LF}$ and $\boldsymbol{v}_{HF}$.

Textual Modality Decoupling. For discrete textual sequences, we propose an Large language Model (LLM)-based semantic decoupling paradigm. A frozen pre-trained LLM extracts raw embeddings $\boldsymbol{E}_t \in \mathbb{R}^{N \times d_{LLM}}$. To quantify token contributions, a Multi-Head Attention (MHA) mechanism computes the normalized attention weight distribution $\boldsymbol{w}_{attn} \in \mathbb{R}^N$. Defining low-frequency as universal base semantics, we obtain the low-frequency vector $\boldsymbol{h}_{t,LF}$ via weighted mean pooling:

$$\boldsymbol{h}_{t,LF} = \sum_{k=1}^{N} w_{attn}^{(k)} \boldsymbol{E}_t^{(k)}, \tag{5}$$

where $w_{attn}^{(k)}$ is the attention score for the k-th token. For high-frequency details, we employ a Top-K routing mechanism to dynamically capture the most discriminative specific tokens. These selected embeddings are refined through a non-linear compression layer to form $\boldsymbol{h}_{t,HF}$. Finally, the textual components are layer-normalized and projected into the d-dimensional collaborative space. To prevent feature collapse caused by padding bias—where LayerNorm inadvertently introduces non-zero offsets at sequence padding positions. We apply a strict masking mechanism by re-weighting the outputs after the encoder, yielding the purified vectors $\boldsymbol{t}_{LF}$ and $\boldsymbol{t}_{HF}$.

Through this pipeline, multimodal inputs are decoupled into four aligned vectors $\{\boldsymbol{v}_{LF}, \boldsymbol{v}_{HF}, \boldsymbol{t}_{LF}, \boldsymbol{t}_{HF}\}$. To ensure semantic consistency while preventing redundant noise entanglement, we optimize a joint loss $\mathcal{L}_{cl}$ involving both intra-item spectral alignment and inter-item preference separation:

$$\mathcal{L}_{cl} = -\sum_{i \in \mathcal{B}} \log \frac{\exp(\text{sim}(\boldsymbol{v}_{i,LF}, \boldsymbol{t}_{i,LF})/\tau)}{\sum_{j \in \mathcal{B}} \exp(\text{sim}(\boldsymbol{v}_{i,LF}, \boldsymbol{t}_{j,LF})/\tau)} + \lambda \|\boldsymbol{v}_{HF} \boldsymbol{t}_{HF}^{\top}\|_F^2, \tag{6}$$

where $\mathcal{B}$ denotes the batch, $\text{sim}(\cdot)$ is the cosine similarity, and τ is the temperature hyperparameter. The Frobenius norm regularizer in the second term enforces orthogonality between high-frequency specificities, thereby strictly eliminating spurious cross-modal noise correlations.

4.3 Multi-View Graph Learning for Collaborative Prior Extraction

This module constructs a robust Teacher Model using complete multimodal data. To prevent information loss from uniform propagation, we build independent graphs for distinct frequency components, shielding high-frequency specificities from over-smoothing. By integrating the decoupled frequency features with graph priors via a gating mechanism and contrastive learning, the teacher extracts high-quality Expert Representations to supervise the student model in missing-modality scenarios.

Prior-Guided Frequency Gating. Let $\boldsymbol{x}_{i,LF}^m, \boldsymbol{x}_{i,HF}^m \in \mathbb{R}^{d_m}$ be the low- and high-frequency features of item i for modality $m \in \{v, t\}$, and $\boldsymbol{e}_i^{cf} \in \mathbb{R}^d$ be its collaborative filtering (CF) ID embedding. Low-frequency features undergo a standard affine transformation $\tilde{\boldsymbol{x}}_{i,LF}^m = \boldsymbol{W}_L^m \boldsymbol{x}_{i,LF}^m + \boldsymbol{b}_L^m$ to map them into the shared collaborative manifold. High-frequency features encapsulate both crucial discriminative semantics and volatile noise. To adaptively weigh the necessity of these details against the risk of noise based on an item's collaborative confidence, high-frequency features are dynamically modulated by a Popularity-Aware Frequency Gate guided by the CF prior. This is formulated as a diagonalized information bottleneck:

$$\tilde{\boldsymbol{x}}_{i,HF}^m = \text{diag}\Big(\sigma(\boldsymbol{W}_{gate} \boldsymbol{e}_i^{cf} + \boldsymbol{b}_{gate})\Big) \cdot \Big(\boldsymbol{W}_H^m \boldsymbol{x}_{i,HF}^m + \boldsymbol{b}_H^m\Big), \tag{7}$$

where $\text{diag}(\cdot)$ constructs a diagonal filtering matrix, $\sigma(\cdot)$ is the Sigmoid function, and $\boldsymbol{W}_{gate}, \boldsymbol{b}_{gate}$ map the CF prior into a gating vector to explicitly suppress non-collaborative noise.

Decoupled Multi-View Graph Propagation. To prevent premature coupling of modalities and frequency bands, we construct independent graph propagation views. We initialize the node state for a frequency representation as $\boldsymbol{e}_{i,m,\phi}^{(0)} = \tilde{\boldsymbol{x}}_{i,\phi}^m$ ($\phi \in \{LF, HF\}$), and for users as $\boldsymbol{e}_{u,m,\phi}^{(0)} = \boldsymbol{e}_u$. A symmetric Graph Convolutional Network (GCN) is executed independently across the user-item bipartite graph $\mathcal{G}$:

$$\boldsymbol{e}_{i,m,\phi}^{(l+1)} = \sum_{j \in \mathcal{N}_i \cup \{i\}} \frac{1}{\sqrt{|\mathcal{N}_i||\mathcal{N}_j|}} \boldsymbol{e}_{j,m,\phi}^{(l)}, \quad \forall l \in \{0, 1, \ldots, L-1\}, \tag{8}$$

where $\mathcal{N}_i$ denotes the first-order topological neighborhood. The high-order features of the L -th layer of all views are then aggregated with the residual CF embedding to form the unified expert representation item $\boldsymbol{e}_i^{joint} = \boldsymbol{e}_i^{cf} + \sum_m \sum_\phi \boldsymbol{e}_{i,m,\phi}^{(L)}$.

Optimization Objectives. To enforce semantic orthogonality between intra-modality high- and low-frequency features, we introduce a *Frequency Decouple Loss* ($\mathcal{L}_{dec}$):

$$\mathcal{L}_{dec} = \mathbb{E}_{i\sim\mathcal{B}}\left[\sum_m \frac{\left|\langle \tilde{\boldsymbol{x}}_{i,LF}^m, \tilde{\boldsymbol{x}}_{i,HF}^m \rangle\right|}{\|\tilde{\boldsymbol{x}}_{i,LF}^m\|_2 \|\tilde{\boldsymbol{x}}_{i,HF}^m\|_2}\right], \tag{9}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product. Furthermore, to semantically anchor the fused representation, which is defined as $\tilde{\boldsymbol{x}}_{i,fused}^m = \tilde{\boldsymbol{x}}_{i,LF}^m + \tilde{\boldsymbol{x}}_{i,HF}^m$ to the structural CF view, we apply an InfoNCE-based *Contrastive Alignment Loss* $\mathcal{L}_{align}$:

$$\mathcal{L}_{align} = -\mathbb{E}_{i\sim\mathcal{B}}\left[\sum_m \log \frac{\exp\left(\text{sim}(\boldsymbol{e}_i^{cf}, \tilde{\boldsymbol{x}}_{i,fused}^m)/\tau\right)}{\sum_{j\in\mathcal{B}} \exp\left(\text{sim}(\boldsymbol{e}_i^{cf}, \tilde{\boldsymbol{x}}_{j,fused}^m)/\tau\right)}\right], \tag{10}$$

where τ is the temperature parameter. Finally, the teacher model is jointly optimized using the standard Bayesian Personalized Ranking (BPR) loss [19], formulated as $\mathcal{L}_{bpr} = -\mathbb{E}_{(u,i,j)\sim\mathcal{D}}[\ln\sigma(\langle \boldsymbol{e}_u^{joint}, \boldsymbol{e}_i^{joint}\rangle - \langle \boldsymbol{e}_u^{joint}, \boldsymbol{e}_j^{joint}\rangle)]$, which encourages observed interactions to score higher than unobserved ones.

4.4 Multi-Level Knowledge Distillation for Modality Generation

The teacher's decoupled training phase inherently forms a computationally heavy, non-end-to-end pipeline. To overcome this deployment bottleneck, we design a parameter-efficient, end-to-end student model utilizing lightweight, modality-specific encoders. Guided by the frozen teacher, the student mimics expert experience to generate robust representations for the missing modality m using the available features $\{\boldsymbol{x}_{i,LF}^{m'}, \boldsymbol{x}_{i,HF}^{m'}\}$ extracted from the observed modality m'.

Asymmetric Completion Strategy. We adopt an asymmetric generation strategy tailored to the semantic density of different frequency bands. To capture global collaborative information, both generation paths explicitly incorporate the item ID embedding $\boldsymbol{e}_i^{cf}$. For the missing low-frequency core semantics $\hat{\boldsymbol{x}}_{i,LF}^m$, a deterministic non-linear projection is performed on the concatenated features:

$$\hat{\boldsymbol{x}}_{i,LF}^m = \mathcal{F}_\theta\left([\boldsymbol{x}_{i,LF}^{m'} \parallel \boldsymbol{e}_i^{cf}]\right) \triangleq \boldsymbol{W}_2 \cdot \text{ReLU}\left(\boldsymbol{W}_1[\boldsymbol{x}_{i,LF}^{m'} \parallel \boldsymbol{e}_i^{cf}] + \boldsymbol{b}_1\right) + \boldsymbol{b}_2, \tag{11}$$

where $\parallel$ denotes concatenation, $\boldsymbol{x}_{i,LF}^{m'}$ represents the available source modality, and $\boldsymbol{W}_*$, $\boldsymbol{b}_*$ are trainable parameters of the Multi-Layer Perceptron (MLP). Conversely, the high-frequency component contains specific, volatile details. To prevent semantic hallucinations, we introduce a Cross-Attention retrieval mechanism. Using a dynamic query formulated by learnable tokens $\boldsymbol{Q}_{token}$ and the projected ID embedding, missing details are cross-retrieved directly from the available source modality's high-frequency features $\boldsymbol{x}_{i,HF}^{m'}$:

$$\hat{\boldsymbol{x}}_{i,HF}^m = \text{softmax}\left(\frac{(\boldsymbol{Q}_{token} + \boldsymbol{W}_{id}\boldsymbol{e}_i^{cf})(\boldsymbol{x}_{i,HF}^{m'})^\top}{\sqrt{d_k}}\right)\boldsymbol{x}_{i,HF}^{m'}, \tag{12}$$

where $\boldsymbol{W}_{id}$ is the ID projection matrix and d_k is the dimensionality scaling factor. This asymmetric design ensures that generated signals strictly adhere to reliable cross-modal evidence.

Hierarchical Distillation Objective. To strictly enforce these expert boundaries, we construct a comprehensive multi-level distillation objective $\mathcal{L}_{distill}$:

$$\mathcal{L}_{distill} = \sum_{\kappa\in\{sub,exp,joint\}} \lambda_\kappa \|\boldsymbol{Z}_{stu}^{(\kappa)} - \boldsymbol{Z}_{tea}^{(\kappa)}\|_F^2 + \lambda_{kd} D_{KL}(\hat{\boldsymbol{y}}_{stu} \parallel \hat{\boldsymbol{y}}_{tea}) + \lambda_{align}\mathcal{L}_{align}, \tag{13}$$

where $\boldsymbol{Z}_{stu}^{(\kappa)}$ and $\boldsymbol{Z}_{tea}^{(\kappa)}$ denote the feature representations of the student and teacher models at the κ-th spatial hierarchy, respectively. $\|\cdot\|_F^2$ denotes the squared Frobenius norm, which minimizes the mean squared error across three progressively deeper hierarchies: the physical subband level (sub), the graph-structured expert level (exp), and the final multimodal joint representation level ($joint$). Furthermore, D_{KL} is the Kullback-Leibler divergence for matching the prediction logits, and $\mathcal{L}_{align}$ is the InfoNCE-based alignment loss which bridges the cross-modal semantic gap of the generated features.

4.5 Expert-Guided Training Strategy

FreDRec follows a sequential training paradigm that first establishes a full-modality teacher and then transfers its knowledge to a student under incomplete observations. This separation provides stable supervision for modality completion while allowing the student to adapt its representations to the recommendation objective.

Teacher Construction Stage. We first pre-train the frequency decoupling module (Section 4.2) to obtain low- and high-frequency representations for each modality. These features subsequently serve as inputs to the multi-view graph learning module (Section 4.3), which integrates modality-specific content with collaborative signals from the training interactions. The teacher is optimized using the recommendation objective together with frequency decoupling and contrastive alignment constraints. After training, its parameters are frozen to provide consistent supervision at the subband, graph-expert, and joint representation levels, as well as recommendation score distributions.

Teacher Construction Stage. We simulate modality incompleteness and train the student to perform recommendation using the observed modalities and collaborative priors. The asymmetric generators reconstruct missing low-frequency semantics and high-frequency details, while hierarchical distillation constrains the generated representations to remain consistent with the teacher. The student is jointly optimized by $\mathcal{L}_{student} = \mathcal{L}_{bpr} + \mathcal{L}_{distill}$, balancing ranking accuracy with feature reconstruction and prediction consistency. Complete-modality teacher representations are used only as training targets, rather than additional student inputs. During inference, the student performs modality completion and recommendation without invoking the teacher network. A two-step variant that retains the teacher's decoupling module is evaluated in Appendix C.1.

5 Experiment

5.1 Experimental Settings

Datasets. We evaluate FreDRec and all baseline models using three real-world e-commerce datasets: Baby, Sports, and Clothing. For data pre-processing, we follow previous studies by filtering out users and items with sparse interactions to ensure data quality. More details about the datasets and preprocessing can be seen in Appendix B.1.

Evaluation Metrics. The performance is evaluated under the missing modality setting using widely adopted top-K ranking metrics, specifically Recall (R@20, R@50) and Normalized Discounted Cumulative Gain (N@20, N@50). Consistent with prior research in multimodal recommendation, we rank the ground-truth items among all un-interacted items for each user to accurately compute these evaluation metrics.

Compared Methods. To assess the effectiveness of FreDRec, we compare it against a comprehensive set of baselines, which are grouped into three categories: **Standard Collaborative Filtering (CF) Models:** MF [19], NGCF [24], LightGCN [8], SGL [28], and SimGCL [34]. **Multi-Modal Recommenders:** VBPR [7], MMGCN [27], GRCN [26], SLMRec [20], BM3 [25], LATTICE [37], MGCN [35], LGMRec [6], DAMRS [33], GUME [13], and DGMRec [11] (the strongest baseline). **Missing Modality-Aware (MMA) Recommender Systems:** CI2MG [14], MILK [2], and SiBraR [5]. More details about these baselines are provided in Appendix B.2.

Implementation Details. To ensure a fair comparison, we select the optimal hyperparameters for each baseline through a systematic grid search on the validation set. To simulate realistic data scarcity, we randomly partition the items in each dataset into three equally sized subsets according to modality availability: (1) **Complete Modality Subset**, where samples possess both visual and textual information; (2) **Single-Modality Missing Subset**, where one modality (either vision or text) is randomly absent; and (3) **All-Modality Missing Subset**, where all modalities are unavailable. FreDRec and its variants are evaluated end-to-end under this challenging hybrid missing-modality evaluation setting. Further experimental details are provided in Appendix B.3.

Table 1: The overall performance comparison of FreDRec and competing baselines on the Baby, Sports, and Clothing datasets. The boldface refers to the highest score, and the underline indicates the best result among all the baselines.

	Dataset	Baby				Sports				Clothing			
	Metric	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50	R@20	R@50	N@20	N@50
CF	SGL	0.0804	0.1422	0.0348	0.0473	0.0917	0.1492	0.0414	0.0531	0.0600	0.0936	0.0271	0.0338
	NGCF	0.0602	0.1137	0.0258	0.0366	0.0701	0.1215	0.0304	0.0408	0.0422	0.0728	0.0176	0.0237
	SimGCL	0.0809	0.1409	0.0349	0.0471	0.0910	0.1465	0.0410	0.0523	0.0542	0.0833	0.0252	0.0310
	MF	0.0611	0.1091	0.0273	0.0370	0.0707	0.1112	0.0327	0.0416	0.0346	0.0533	0.0164	0.0201
	LightGCN	0.0733	0.1323	0.0320	0.0440	0.0829	0.1369	0.0379	0.0488	0.0514	0.0818	0.0227	0.0288
Multi-Modal Recommenders	BM3	0.0683	0.1235	0.0296	0.0408	0.0908	0.1466	0.0400	0.0513	0.0591	0.0920	0.0268	0.0334
	VBPR	0.0514	0.0937	0.0213	0.0299	0.0741	0.1229	0.0328	0.0427	0.0462	0.0737	0.0207	0.0226
	LGMRec	0.0813	0.1410	0.0352	0.0471	0.0906	0.1496	0.0403	0.0522	0.0624	0.1015	0.0277	0.0355
	MMGCN	0.0519	0.0991	0.0215	0.0310	0.0509	0.0913	0.0215	0.0297	0.0289	0.0530	0.0120	0.0168
	DAMRS	0.0804	0.1390	0.0355	0.0474	0.0941	0.1526	0.0416	0.0534	0.0670	0.1066	0.0301	0.0380
	SLMRec	0.0753	0.1254	0.0340	0.0422	0.0914	0.1462	0.0415	0.0526	0.0624	0.0979	0.0281	0.0351
	GUME	0.0835	0.1429	0.0369	0.0489	0.0947	0.1554	0.0424	0.0546	0.0639	0.1016	0.0291	0.0366
	LATTICE	0.0738	0.1297	0.0319	0.0432	0.0867	0.1401	0.0306	0.0384	0.0581	0.0929	0.0262	0.0332
	GRCN	0.0644	0.1151	0.0274	0.0377	0.0681	0.1157	0.0300	0.0397	0.0381	0.0644	0.0161	0.0214
	MGCN	0.0833	0.1389	0.0366	0.0481	0.0941	0.1525	0.0425	0.0544	0.0665	0.1052	0.0300	0.0377
MMA RSs	SIBRAR	0.0480	0.0888	0.0207	0.0289	0.0434	0.0758	0.0190	0.0255	0.0264	0.0453	0.0110	0.0148
	MILK	0.0427	0.0763	0.0182	0.0250	0.0362	0.0626	0.0155	0.0209	0.0226	0.0376	0.0094	0.0124
	CI2MG	0.0720	0.1285	0.0305	0.0420	0.0717	0.1179	0.0331	0.0425	0.0523	0.0845	0.0237	0.0301
	DGMRec	0.0897	0.1531	0.0404	0.0528	0.1024	0.1625	0.0462	0.0584	0.0725	0.1134	0.0324	0.0406
Ours	**FreDRec-Stu.**	**0.0942**	**0.1574**	**0.0450**	**0.0589**	**0.1063**	**0.1678**	**0.0526**	**0.0661**	**0.0837**	**0.1272**	**0.0418**	**0.0511**
	FreDRec-Tec.	0.0888	0.1454	0.0440	0.0565	0.1045	0.1646	0.0524	0.0657	0.0746	0.1149	0.0375	0.0462
	Improv.	5.02%	2.81%	11.39%	11.55%	3.81%	3.26%	13.85%	13.18%	15.45%	12.17%	29.01%	25.86%

Table 2: The ablation study on the Baby, Sports, and Clothing datasets. The boldface refers to the highest score. The results demonstrate the effectiveness of each proposed module. (FD: Frequency-domain Decoupling Module, PG: Prior Collaborative Module, KD: Knowledge Distillation Module)

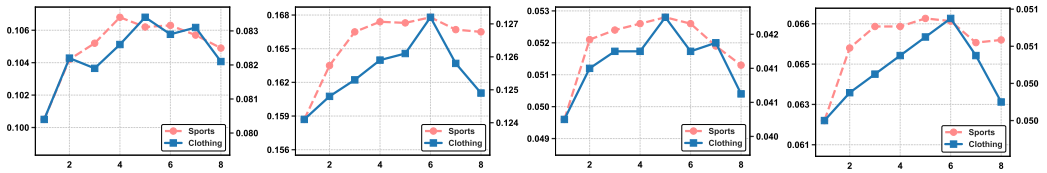
Dataset	**Metric**	**FreDRec**	**FD Module**			**PG Module**			**KD Module**		
			w/o FD	-Freq-V	-Freq-T	w/o PG	-Guard	-ItemID	w/o KD	-Feat	-Logit
Baby	R@20	**0.0942**	0.0796	0.0836	0.0861	0.0823	0.0925	0.0935	0.0647	0.0748	0.0827
	N@20	**0.0450**	0.0392	0.0406	0.0418	0.0401	0.0449	0.0443	0.0323	0.0369	0.0415
Sports	R@20	**0.1058**	0.0971	0.0993	0.1002	0.0881	0.1028	0.1045	0.0804	0.0934	0.1007
	N@20	**0.0526**	0.0484	0.0493	0.0499	0.0437	0.0514	0.0514	0.0412	0.0470	0.0503
Clothing	R@20	**0.0819**	0.0712	0.0737	0.0764	0.0560	0.0806	0.0803	0.0580	0.0742	0.0760
	N@20	**0.0411**	0.0334	0.0353	0.0371	0.0280	0.0400	0.0404	0.0293	0.0371	0.0382

5.2 Overall Performance Comparison

As shown in Table 1, FreDRec significantly boost overall performance, particularly in the top-ranking metric N@20. Specifically, our lightweight model (**FreDRec-Stu.**) comprehensively outperforms the strongest baseline, DGMRec, achieving substantial improvements of **11.39 %**, **13.85 %**, and **29.01 %** on the Baby, Sports, and Clothing datasets, respectively. Unlike existing models that suffer from noise entanglement under missing modalities, FreDRec achieves dual gains in Recall and NDCG. Our physical-level frequency decoupling significantly enhances ranking quality: low-frequency skeletons maintain a reliable global ranking baseline, while pure high-frequency details enable precise local fine-tuning, effectively securing top positions for highly relevant items. Notably, even without the missing-modality generation module, our teacher model (**FreDRec-Tec.**) remains highly robust. This confirms that our decoupling strategy alone effectively eliminates noise and preserves ranking fidelity under extreme feature sparsity.

5.3 Ablation Study and Hyper-parameter Analysis

Table 2 details the ablation studies on the core components of FreDRec. Removing frequency decoupling (**FD**) entirely or partially degrades performance significantly, confirming that physical domain disentanglement is essential for effective noise reduction. Discarding the prior generation (**PG**) module and its spatial or collaborative subcomponents causes severe accuracy drops, proving that reliable completion requires joint guidance from expert structures and collaborative signals. Additionally, the inferiority of distillation (**KD**) ablated variants verifies that our framework deeply absorbs structural knowledge rather than merely imitating prediction logits. Finally, sensitivity analysis (Figure2) reveals that the expert alignment weight w_{exp} requires careful tuning. Small values



(a) Recall 20 (b) Recall 50 (c) NDCG 20 (d) NDCG 50

Figure 2: The parameter sensitivity results of the proposed FreDRec regarding the expert constraint weight w_{exp}. The results are based on the Sports and Clothing datasets across four ranking metrics.



(a) Original (b) FFT Filtered (c) HF: Details (d) LF: Core

Figure 3: Frequency decoupling visualization.

provide insufficient boundary constraints during missing feature completion, while excessively large weights overly suppress the primary recommendation objective, making a moderate optimal value crucial to balance expert knowledge transfer with the main ranking task.

5.4 Efficiency and Robustness Analysis

To validate the practical industrial deployment value of our end to end architecture, we compare the two step distillation variant, which still relies on the heavy teacher model for multimodal feature decoupling during online inference with our lightweight end to end model **FreDRec**. Analysis reveals that compared to the two step variant, **FreDRec** achieves an order of magnitude reduction in both inference latency and memory footprint, while maintaining highly superior recommendation accuracy. Furthermore, under various missing ratios up to 90%, while all conventional baselines suffer a drastic performance drop, FreDRec exhibits a remarkably graceful decline. This strongly highlights that our model can still maintain consistently competitive recommendation accuracy even under extreme feature sparsity conditions. A detailed quantitative comparison regarding memory consumption, inference time, and recommendation performance is provided in Appendix C.1.

5.5 Frequency Decoupling Effects

To verify our physical-level decoupling, we visualize the feature reconstruction process. As Figure 3 shows, the FFT module (b) adaptively denoises raw signals. Subsequently, IDWT reconstruction demonstrates that the LF component (d) preserves stable core semantics, while the HF component (c) explicitly isolates specific details. This proves FreDRec achieves interpretable disentanglement rather than relying on black-box mapping. Furthermore, t-SNE visualization of 500 randomly sampled instances reveals that the extracted LF and HF features form distinctly separated clusters in the latent space, confirming robust feature decoupling. Detailed t-SNE results are provided in Appendix C.2.

6 Conclusion

We propose FreDRec, a novel, lightweight and highly efficient framework simultaneously resolving semantic hallucination and noise entanglement in incomplete multimodal recommendation. By fundamentally integrating physical domain frequency decoupling with global collaborative priors and multilevel knowledge distillation, our approach explicitly isolates core semantics from raw signals to achieve high fidelity feature completion. Extensive experiments across three large scale real world datasets demonstrate that FreDRec significantly outperforms all current leading baselines. Furthermore, comprehensive analyses confirm its superior ranking quality, exceptional deployment efficiency, and remarkably robust performance, successfully maintaining competitive accuracy even under extreme sparsity conditions with up to ninety percent missing modalities.

Acknowledgments

This work was partially supported by the General Project for Basic Scientific Research of Higher Education Institutions in Jiangsu Province, under Grant No. 25KJB520028, and the National Natural Science Foundation of China, under Grant No. 62602280 and No. 62372242.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9163–9171, 2019.

[2] Haoyue Bai, Le Wu, Min Hou, Miaomiao Cai, Zhuangzhuang He, Yuyang Zhou, Richang Hong, and Meng Wang. Multimodality invariant learning for multimedia-based new item recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 677–686, 2024.

[3] Huilin Chen, Miaomiao Cai, Fan Liu, Zhiyong Cheng, Richang Hong, and Meng Wang. I3-MRec: Invariant learning with information bottleneck for incomplete modality recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6133–6142, 2025.

[4] Jae Won Cho, Dong-Jin Kim, Jinsoo Choi, Yunjae Jung, and In So Kweon. Dealing with missing modalities in the visual question answer-difference prediction task through knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1592–1601, 2021.

[5] Christian Ganhör, Marta Moscati, Anna Hausberger, Shah Nawaz, and Markus Schedl. A multimodal single-branch embedding network for recommendation in cold-start and missing modality scenarios. In *Proceedings of the 18th ACM conference on recommender systems*, pages 380–390, 2024.

[6] Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. LGMRec: Local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 8454–8462, 2024.

[7] 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.

[8] 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.

[9] Michael Heideman, Don Johnson, and Charles Burrus. Gauss and the history of the fast fourier transform. *IEEE Assp Magazine*, 1(4):14–21, 1984.

[10] SeongKu Kang, Junyoung Hwang, Wonbin Kweon, and Hwanjo Yu. DE-RRD: A knowledge distillation framework for recommender system. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 605–614, 2020.

[11] Jiwan Kim, Hongseok Kang, Sein Kim, Kibum Kim, and Chanyoung Park. Disentangling and generating modalities for recommendation in missing modality scenarios. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1820–1829, 2025.

[12] Jin Li, Shoujin Wang, Qi Zhang, Shui Yu, and Fang Chen. Generating with fairness: A modality-diffused counterfactual framework for incomplete multimodal recommendations. In *Proceedings of the ACM on Web Conference 2025*, pages 2787–2798, 2025.

[13] Guojiao Lin, Meng Zhen, Dongjie Wang, Qingqing Long, Yuanchun Zhou, and Meng Xiao. GUME: Graphs and user modalities enhancement for long-tail multimodal recommendation. In *Proceedings of the 33rd ACM international conference on information and knowledge management*, pages 1400–1409, 2024.

[14] Zhenghong Lin, Yanchao Tan, Yunfei Zhan, Weiming Liu, Fan Wang, Chaochao Chen, Shiping Wang, and Carl Yang. Contrastive intra-and inter-modality generation for enhancing incomplete multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6234–6242, 2023.

[15] Fan Lu, Xiaolong Xu, Haolong Xiang, Lianyong Qi, Xiaokang Zhou, Fei Dai, and Wanchun Dou. Cllmrec: Contrastive learning with llms-based view augmentation for sequential recommendation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, 2025.

[16] Haokai Ma, Zhuang Qi, Xinxin Dong, Xiangxian Li, Yuze Zheng, Xiangxu Meng, and Lei Meng. Cross-modal content inference and feature enrichment for cold-start recommendation. In *2023 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2023.

[17] Rongqing Kenneth Ong and Andy W. H. 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*, 2025.

[18] 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.

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

[20] 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.

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

[22] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

[23] Cheng Wang, Mathias Niepert, and Hui Li. LRMM: Learning to recommend with missing modalities. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3360–3370, 2018.

[24] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd international ACM SIGIR conference on Research and development in Information Retrieval*, pages 165–174, 2019.

[25] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM web conference 2023*, pages 790–800, 2023.

[26] 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.

[27] 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.

[28] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 726–735, 2021.

[29] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

[30] Xiaolong Xu, Hongsheng Dong, Lianyong Qi, Xuyun Zhang, Haolong Xiang, Xiaoyu Xia, Yanwei Xu, and Wanchun Dou. Cmclrec: Cross-modal contrastive learning for user cold-start sequential recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1589–1598, 2024.

[31] Xiaolong Xu, Hongsheng Dong, Haolong Xiang, Xiyuan Hu, Xiaoyong Li, Xiaoyu Xia, Xuyun Zhang, Lianyong Qi, and Wanchun Dou. C2lrec: Causal contrastive learning for user cold-start recommendation with social variables. *ACM Transactions on Information Systems*, 43(6):1–28, 2025.

[32] Xiaolong Xu, Xudong Zhao, Haolong Xiang, Xuyun Zhang, Wei Shen, Hongsheng Hu, and Lianyong Qi. Hpserec: A hierarchical partitioning and stepwise enhancement framework for long-tailed sequential recommendation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[33] 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*, pages 3645–3656, 2024.

[34] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1294–1303, 2022.

[35] 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.

[36] Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. Mind individual information! principal graph learning for multimedia recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 13096–13105, 2025.

[37] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3872–3880, 2021.

[38] Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. Tmdc: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 2281–2289, 2026.

A: Theoretical Analysis of Physical-Level Frequency Decoupling

This section establishes a representation-error comparison between source-level frequency filtering with teacher-guided distillation and feature-level mappings satisfying inverse Lipschitz stability. We first characterize the propagation of input noise through a nonlinear encoder. We then derive a Lipschitz-based lower bound for feature-level processing and an upper bound for frequency filtering and distillation. The resulting theorem specifies sufficient conditions for a strict reduction in expected representation error.

A.1 Signal Model and Frequency Projection

Let the observed signal be $\mathbf{s} = \mathbf{s}_c + \mathbf{n} \in \mathbb{R}^D$, where $\mathbf{s}_c$ denotes a fixed clean signal and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_D)$ denotes additive noise. All expectations in this section are conditional on $\mathbf{s}_c$, with the learned model parameters held fixed.

Let $\mathbf{P} \in \mathbb{R}^{D \times D}$ denote an orthogonal projector onto the retained low-frequency subspace:

$$\mathbf{P}^T = \mathbf{P}, \qquad \mathbf{P}^2 = \mathbf{P}. \tag{14}$$

For an orthogonal wavelet transform $\mathbf{W}$, the projector is represented as $\mathbf{P} = \mathbf{W}^T\mathbf{M}_{LF}\mathbf{W}$, where $\mathbf{M}_{LF}$ selects the low-frequency coefficients. This operator includes both coefficient selection and reconstruction.

We assume that the target core signal lies in the retained subspace:

$$\mathbf{P}\mathbf{s}_c = \mathbf{s}_c. \tag{15}$$

This condition specifies the semantic preservation requirement of the projection analysis. It does not assert that all task-relevant information in a multimodal signal is low-frequency.

Let $\phi : \mathbb{R}^D \to \mathbb{R}^d$ be a differentiable encoder and define $\mathbf{z}^\star = \phi(\mathbf{s}_c)$ and $\mathbf{J} = \nabla\phi(\mathbf{s}_c)$. Write the encoder response as

$$\phi(\mathbf{s}_c + \mathbf{h}) = \mathbf{z}^\star + \mathbf{J}\mathbf{h} + \mathbf{r}(\mathbf{h}), \tag{16}$$

where $\mathbf{r}(\mathbf{h})$ is the exact remainder after the first-order term. Assume that the following second moments are finite:

$$\begin{aligned} R_{raw} &= \left(\mathbb{E}\|\mathbf{r}(\mathbf{n})\|_2^2\right)^{1/2}, \\ R_{LF} &= \left(\mathbb{E}\|\mathbf{r}(\mathbf{P}\mathbf{n})\|_2^2\right)^{1/2}. \end{aligned} \tag{17}$$

Define the first-order noise energy

$$B = \mathrm{Tr}\left(\mathbf{J}\Sigma_n\mathbf{J}^T\right) = \sigma_n^2\|\mathbf{J}\|_F^2 > 0, \tag{18}$$

where $\Sigma_n = \sigma_n^2\mathbf{I}_D$. The corresponding retained energy is

$$B_{LF} = \sigma_n^2\|\mathbf{J}\mathbf{P}\|_F^2. \tag{19}$$

Since $\mathbf{P}$ and $\mathbf{I}_D - \mathbf{P}$ are complementary orthogonal projectors,

$$B - B_{LF} = \sigma_n^2\|\mathbf{J}(\mathbf{I}_D - \mathbf{P})\|_F^2 \geq 0. \tag{20}$$

Consequently, the retained perturbation ratio

$$\rho_J = \frac{B_{LF}}{B} = \frac{\|\mathbf{J}\mathbf{P}\|_F^2}{\|\mathbf{J}\|_F^2} \in [0, 1] \tag{21}$$

characterizes the interaction between frequency selection and encoder sensitivity. In particular, $\rho_J < 1$ whenever $\mathbf{J}(\mathbf{I}_D - \mathbf{P}) \neq \mathbf{0}$.

The reverse triangle inequality and Minkowski's inequality yield

$$\left(\mathbb{E}\|\phi(\mathbf{s}) - \mathbf{z}^\star\|_2^2\right)^{1/2} \geq \left(\sqrt{B} - R_{raw}\right)_+, \tag{22}$$

and

$$\left(\mathbb{E}\|\phi(\mathbf{P}\mathbf{s}) - \mathbf{z}^\star\|_2^2\right)^{1/2} \leq \sqrt{\rho_J B} + R_{LF}, \tag{23}$$

where $(a)_+ = \max\{a, 0\}$. These inequalities retain the nonlinear remainder instead of identifying the first-order approximation with the complete representation error.

A.2 A Lipschitz-Based Lower Bound for Feature-Level Processing

Consider a feature-level mapping $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^d$ applied to the pre-extracted representation $\mathbf{x} = \phi(\mathbf{s})$. The analysis includes mappings trained with mutual information regularization when they satisfy the following stability condition.

Inverse Lipschitz stability. Assume that $\mathcal{F}$ is injective on the analyzed representation set, which contains $\mathbf{z}^\star$ and the support of $\mathbf{x}$, and that its inverse is $L_{\mathcal{F}^{-1}}$-Lipschitz on the corresponding image:

$$\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \le L_{\mathcal{F}^{-1}} \|\mathcal{F}(\mathbf{x}_1) - \mathcal{F}(\mathbf{x}_2)\|_2. \tag{24}$$

Here, $L_{\mathcal{F}^{-1}} > 0$ is the Lipschitz constant of the inverse mapping, rather than the forward mapping. This condition excludes complete collapse of the analyzed input distinctions.

For example, a residual mapping $\mathcal{F}(\mathbf{x}) = \mathbf{x} + g(\mathbf{x})$ with $\mathrm{Lip}(g) \le \kappa < 1$ satisfies

$$\|\mathcal{F}(\mathbf{x}_1) - \mathcal{F}(\mathbf{x}_2)\|_2 \ge (1 - \kappa)\|\mathbf{x}_1 - \mathbf{x}_2\|_2. \tag{25}$$

Its inverse therefore admits the Lipschitz bound $L_{\mathcal{F}^{-1}} \le (1 - \kappa)^{-1}$.

Define the clean-signal transformation bias

$$b_{\mathcal{F}} = \|\mathcal{F}(\mathbf{z}^\star) - \mathbf{z}^\star\|_2 \tag{26}$$

and the expected representation error

$$\mathcal{E}_{feat} = \mathbb{E}\|\mathcal{F}(\phi(\mathbf{s})) - \mathbf{z}^\star\|_2^2. \tag{27}$$

Theorem 1 (Inverse-Lipschitz Representation Error Lower Bound). Under the assumptions above,

$$\mathcal{E}_{feat} \ge \left[\frac{(\sqrt{B} - R_{raw})_+}{L_{\mathcal{F}^{-1}}} - b_{\mathcal{F}} \right]_+^2. \tag{28}$$

Proof. Applying the reverse triangle inequality in L^2 gives

$$\begin{aligned} \sqrt{\mathcal{E}_{feat}} &\ge \left(\mathbb{E}\|\mathcal{F}(\mathbf{x}) - \mathcal{F}(\mathbf{z}^\star)\|_2^2 \right)^{1/2} - b_{\mathcal{F}} \\ &\ge \frac{1}{L_{\mathcal{F}^{-1}}} \left(\mathbb{E}\|\mathbf{x} - \mathbf{z}^\star\|_2^2 \right)^{1/2} - b_{\mathcal{F}}. \end{aligned} \tag{29}$$

Substituting Eq. (22), taking the positive part, and squaring establish Eq. (28). □

For an affine encoder and a clean-signal-preserving mapping, $R_{raw} = 0$ and $b_{\mathcal{F}} = 0$. The lower bound reduces to

$$\boxed{\mathcal{E}_{feat} \ge \frac{1}{L_{\mathcal{F}^{-1}}^2} \mathrm{Tr}\left(\mathbf{J} \Sigma_n \mathbf{J}^T \right).} \tag{30}$$

Theorem 1 attributes the lower bound to inverse stability of the feature mapping. Mutual information minimization alone does not establish this condition. Accordingly, the theorem applies to the stated class of stable feature mappings and does not impose a universal lower bound on all disentanglement methods.

A.3 Frequency Filtering, Teacher Guidance, and Strict Error Comparison

Define the filtered representation as $\mathbf{z}_{LF} = \phi(\mathbf{P}\mathbf{s})$. Let $\mathcal{T}$ denote a teacher-guidance operator. Assume that its sensitivity along the analyzed perturbations satisfies

$$\|\mathcal{T}(\mathbf{z}_{LF}) - \mathcal{T}(\mathbf{z}^\star)\|_2 \le q\|\mathbf{z}_{LF} - \mathbf{z}^\star\|_2 \quad \text{almost surely}, \tag{31}$$

where $q \ge 0$. This condition concerns perturbations around the fixed clean representation. It does not require global contraction of the entire semantic space.

Let

$$b_T = \|\mathcal{T}(\mathbf{z}^\star) - \mathbf{z}^\star\|_2 \tag{32}$$

measure the teacher's clean-signal bias. For a student representation $\mathbf{z}_{stu} = f_\theta(\mathbf{X}_{obs})$, define its root-mean-square distillation error as

$$\epsilon_D = \left(\mathbb{E}\|\mathbf{z}_{stu} - \mathcal{T}(\mathbf{z}_{LF})\|_2^2 \right)^{1/2}. \tag{33}$$

Here, $\mathbf{X}_{obs}$ contains only the inputs available to the student under the specified missing-modality setting.

Theorem 2 (Frequency-Guided Student Error Upper Bound). The student representation error satisfies

$$\mathcal{E}_{stu} := \mathbb{E}\|\mathbf{z}_{stu} - \mathbf{z}^\star\|_2^2 \le \left[\epsilon_D + b_T + q\left(\sqrt{\rho_J B} + R_{LF}\right)\right]^2. \tag{34}$$

Proof. Decompose the error into

$$\begin{aligned}\mathbf{z}_{stu} - \mathbf{z}^\star = \mathbf{z}_{stu} - \mathcal{T}(\mathbf{z}_{LF}) \\ + \mathcal{T}(\mathbf{z}_{LF}) - \mathcal{T}(\mathbf{z}^\star) \\ + \mathcal{T}(\mathbf{z}^\star) - \mathbf{z}^\star.\end{aligned} \tag{35}$$

Minkowski's inequality and Eq. (31) imply

$$\begin{aligned}\sqrt{\mathcal{E}_{stu}} \le& \epsilon_D + b_T + q\left(\mathbb{E}\|\mathbf{z}_{LF} - \mathbf{z}^\star\|_2^2\right)^{1/2} \\ \le& \epsilon_D + b_T + q\left(\sqrt{\rho_J B} + R_{LF}\right).\end{aligned} \tag{36}$$

Squaring establishes the result. □

Missing-modality uncertainty. The distillation error is a population quantity and includes information absent from the student inputs. Writing $\mathbf{Z}_T = \mathcal{T}(\mathbf{z}_{LF})$, the conditional expectation decomposition gives

$$\begin{aligned}\epsilon_D^2 = \mathbb{E}\|\mathbf{Z}_T - \mathbb{E}[\mathbf{Z}_T \mid \mathbf{X}_{obs}]\|_2^2 \\ + \mathbb{E}\|f_\theta(\mathbf{X}_{obs}) - \mathbb{E}[\mathbf{Z}_T \mid \mathbf{X}_{obs}]\|_2^2.\end{aligned} \tag{37}$$

The first term is the irreducible uncertainty under the available inputs. The second term measures the student's approximation error.

Strict comparison. Define

$$\mathcal{B}_{feat} = \left[\frac{(\sqrt{B} - R_{raw})_+}{L_{\mathcal{F}^{-1}}} - b_{\mathcal{F}}\right]_+ \tag{38}$$

and

$$\mathcal{B}_{stu} = \epsilon_D + b_T + q\left(\sqrt{\rho_J B} + R_{LF}\right). \tag{39}$$

Whenever

$$\boxed{\mathcal{B}_{stu} < \mathcal{B}_{feat}}, \tag{40}$$

Theorems 1 and 2 establish

$$\mathcal{E}_{stu} \le \mathcal{B}_{stu}^2 < \mathcal{B}_{feat}^2 \le \mathcal{E}_{feat}. \tag{41}$$

In the idealized affine setting with $R_{raw} = R_{LF} = 0$, $b_{\mathcal{F}} = b_T = 0$, and $\epsilon_D = 0$, the sufficient condition becomes

$$q^2 \rho_J < \frac{1}{L_{\mathcal{F}^{-1}}^2}. \tag{42}$$

Under this condition,

$$\boxed{\mathcal{E}_{stu} \le q^2 \rho_J \operatorname{Tr}\left(\mathbf{J}\Sigma_n \mathbf{J}^T\right) < \frac{\operatorname{Tr}\left(\mathbf{J}\Sigma_n \mathbf{J}^T\right)}{L_{\mathcal{F}^{-1}}^2} \le \mathcal{E}_{feat}}. \tag{43}$$

Conclusion and scope. The analysis establishes a strict representation-error advantage under the stated semantic preservation, inverse stability, and error-budget conditions. Frequency selection reduces the first-order noise energy, while teacher stability and distillation accuracy determine the remaining student error.

The projection model describes an idealized low-frequency pathway. The complete FreDRec architecture additionally contains high-frequency branches, learned gates, graph propagation, and textual attention. Applying the comparison theorem to a specific implementation requires verification of its assumptions and constants. The theorem concerns representation error; comparative Recall and NDCG performance is evaluated experimentally.

B: Experimental Settings

B.1 Dataset and Preprocessing

We compare FreDRec with baseline models using three real-world datasets from online services: Amazon Clothing, Amazon Sports, and Amazon Baby. For data pre-processing, we follow prior research by filtering out users with fewer than five interactions. Table 3 presents the statistics of the datasets after pre-processing.

Table 3: Statistics of the evaluated datasets.

Datasets	**#Items**	**#Users**	**#Int**	**Avg** $\lvert S_u \rvert$
Baby	7,050	19,445	160,792	8.27
Sports	18,357	35,598	296,337	8.32
Clothing	23,033	39,387	278,677	7.08

Table 4: Efficiency, deployment cost, and performance comparison between the Two-step variant and **FreDRec** on the Clothing dataset. Inference latency is reported as the average time required to process a single item.

Architecture	**Pipeline Strategy**	**Online Serving Costs**		**Model Size**	**Performance**	
		GPU VRAM	**Latency / Item**	**Parameters**	**Recall@20**	**NDCG@20**
Two-step Variant	Heavy Offline	9,685 MB	6.18 ms	3,290 M	0.0858	0.0432
FreDRec	**Lightweight End-to-End**	**1,585 MB**	**1.36 ms**	**584 M**	**0.0837**	**0.0418**

B.2 Baselines

- **CI2MG** develops a contrastive intra- and inter-modality generation module to reconstruct absent features, enhancing multimedia recommendation under incomplete modalities.
- **MILK** introduces a multimodality invariant learning framework that constructs heterogeneous environments via cyclic mixup to capture robust user preferences against missing modalities.
- **SiBraR** utilizes a weight-sharing single-branch embedding network to encode both collaborative and multimodal side information simultaneously, effectively handling missing modality scenarios.
- **DGMRec** disentangles modality representations into general and specific features, generating missing modalities by integrating aligned cross-modal signals with user preferences.

B.3 Implementation Details

All experimental evaluations were carried out on a high-performance Linux server equipped with dual NVIDIA GeForce RTX 5090 GPUs, with the implementation developed in Python 3.11.14 and PyTorch 2.2.0. To balance information propagation against the risk of over-smoothing, the graph propagation depth L is uniformly configured at **2** across both **FreDRec** and all GNN-based baselines. An embedding dimension of $d = 128$ is consistently maintained for every evaluated model to ensure fair comparative analysis. Parameter optimization relies on the Adam algorithm, initialized with a learning rate of 1×10^{-4} and a mini-batch size of **256**. Hyperparameters specific to **FreDRec** were determined through a grid search on the validation set. In particular, the frequency-domain decoupling loss weight λ_{freq} is set to **0.1** to promote effective separation of modality features. The expert constraint weight w_{exp}, which regulates the intensity of multi-level distillation for latent space alignment, is fixed at **6.0**. Finally, the contrastive learning temperature τ is established at **0.1**.

C: More Experimental Results

C.1 Efficiency and Robustness Analysis

Efficiency and Industrial Deployment Analysis. To validate the practical industrial deployment value of our end-to-end architecture, we quantitatively compare the online serving efficiency, deployment costs, and recommendation accuracy between the traditional two-step variant and our lightweight **FreDRec**. As summarized in Table 4, the two-step strategy relies heavily on cumbersome offline

feature extraction, resulting in a massive model size of 3,290 M parameters. Consequently, it consumes an excessive 9,685 MB of GPU VRAM and requires 6.18 ms to process a single item, rendering it highly impractical for large-scale, real-time online serving. In stark contrast, by seamlessly integrating our physical-level frequency decoupling within an end-to-end framework, **FreDRec** achieves a remarkable 82% reduction in model parameters, bringing the total down to 584 M. This lightweight design allows our model to operate within a profoundly fractional memory budget of merely 1,585 MB and accelerates the per-item inference latency to an impressive **1.36 ms**, delivering a 4.5× speedup. More importantly, this extraordinary efficiency leap is achieved with only a marginal and highly acceptable performance trade-off, as the model maintains a highly competitive Recall@20 of 0.0837 and an NDCG@20 of 0.0418. These results perfectly demonstrate that FreDRec strictly preserves high-fidelity multimodal ranking capabilities while thoroughly satisfying the stringent latency and memory constraints of real-world industrial recommendation systems.

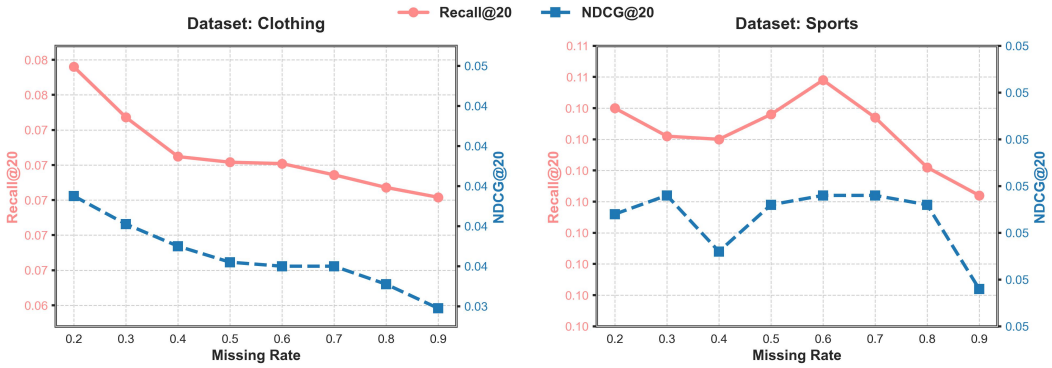


Figure 4: Robustness analysis of FreDRec under varying modality missing ratios on the Clothing and Sports datasets across Recall@20 and NDCG@20.



Figure 5: Visual analysis of frequency decoupling across text and vision modalities using t-SNE, where the distinct separation of high-frequency components confirms effective disentanglement at the source while the clustering of low-frequency features demonstrates cross-modal alignment for the same underlying item.

Robustness Analysis. To evaluate the resilience of FreDRec under varying degrees of data sparsity, we vary the modality missing rate from 20% to 90% and report the performance on the Clothing and Sports datasets. As illustrated in Figure 4, FreDRec exhibits an exceptionally stable performance trajectory. On the Clothing dataset, both Recall@20 and NDCG@20 experience only a marginal and smooth decline despite the severe escalation in missing rates. More notably, on the Sports dataset, the performance remains remarkably steady, with fluctuations kept within an extremely narrow range even under extreme sparsity. We attribute this strong robustness to our physical-level frequency decoupling and teacher-guided distillation mechanisms. Instead of suffering from catastrophic semantic loss like conventional methods, FreDRec effectively filters out environmental noise and anchors on pure

low-frequency skeletons. Consequently, even when up to 90% of the modalities are missing, our model relies on high-fidelity distilled structural knowledge to maintain competitive ranking accuracy, demonstrating a truly graceful performance decline.

C.2 Frequency Decoupling Analysis

To evaluate the effectiveness of our source-level frequency disentanglement, we visualize the decoupled multimodal representations using t-SNE. As illustrated in the Figure 5, our approach successfully decouples the low-frequency and high-frequency information directly from the modality encoders. The disentanglement effect is highly significant, as the high-frequency features from the visual and textual branches diverge into distinct modality-specific spaces. This separation verifies that the high-frequency components successfully capture unique fine-grained details specific to each modality without mutual interference. More importantly, the low-frequency features from both the vision and text modalities cluster tightly together within the joint embedding space. This convergence demonstrates that despite being extracted from entirely different modality sources, the low-frequency signals consistently retain the core shared semantics. This strong semantic alignment proves that the low-frequency information successfully describes the same underlying item, confirming the validity and stability of our source-level frequency decoupling strategy.

D: Limitation

Two potential limitations of this work should be noted. First, the training phase of the proposed framework relies on the offline extraction of high-dimensional teacher features, which requires significant computational resources and storage capacity during the initial distillation stage. While the student model remains highly efficient during online serving, the heavy dependence on teacher knowledge poses a challenge for frequent model retraining in dynamic environments where item features change rapidly. Second, the performance of the frequency-level disentanglement is sensitive to the configuration of filtering thresholds and distillation weights, and identifying the optimal balance between semantic stability and modal details across diverse datasets may involve a non-trivial hyperparameter tuning process. In future research, exploring automated architecture search or adaptive filtering mechanisms may further enhance the scalability and generalizability of the framework.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer:[Yes]

   Justification: The contributions and scope of the paper are included in the abstract and Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: A limitation section is included in the Appendix D.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: For each of our main theorems, we provide a detailed description of the modeling assumptions prior to the theorem. We provide complete proof in Appendix A. Guidelines:

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification:We introduce the details of the experiment, such as the information on hardware and software, in the implementation detail section, i.e., Appendix B.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have attached the data and code used in this paper in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the details of the experimental settings, such as the data split, optimizer, etc., in the experimental setting section (Section 4.1) in the main paper and the implementation detail section (Appendix B.3) in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We did not perform statistical error analysis because the performance improvements achieved by FreDRec over all compared baselines are highly substantial and consistent across all datasets and evaluation metrics, demonstrating its superiority without the need for marginal statistical significance testing.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: We provide the details of compute resources in Appendix B.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have made sure that our paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential positive impacts that our algorithm will bring in the Introduction section, i.e., Section 1.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: There is no risk of misuse of the proposed method and the datasets used in the paper are open-sourced.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have cited the original paper or attached the link to the existing assets used in this paper.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We have attached the introduction of how to run the code and the license in the code repository.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: This paper does not involve crowd sourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification:This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: In this study, we employed a large language model, and the detailed methodology for its use is presented in Section 3.2.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.