Capturing Dynamic User Interests Under Modality Imbalance for Multimodal Sequential Recommendation

Zilong Li[1], Jia Zhu[1*], Chenglei Huang[1], Zhangze Chen[1], Hanghui Guo[2], Guoqing Ma[1], Jianxia Ling[1]

[1]Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University
[2]School of Computer Science and Engineering, Southeast University
{lizilong, jiazhu}@zjnu.edu.cn

Abstract

Multimodal sequential recommender systems leverage diverse modal inputs to enhance the accuracy and relevance of personalized recommendations. However, existing fusion strategies often struggle to capture intricate cross-modal interactions, especially under the evolving dynamics of user intent. Moreover, they frequently neglect modality imbalance issues, leading to suboptimal utilization of multimodal information. To address these challenges, we propose DuAF-MAT, a novel framework for robust multimodal sequential recommendation. Our approach consists of three key components: (1) a Dual-Aware Adaptive Fusion (DuAF) module dynamically calibrates modality contributions by jointly modeling user preferences and temporal information, enabling the extraction of multimodal features aligned with evolving user interests; (2) by integrating Modality Adversarial Training with the Mixture-of-Experts paradigm, MAT-MoE employs an ensemble of expert generators to dynamically reconstruct missing modality representations, effectively mitigating modality imbalance challenges; (3) to address the inherent sparsity of sequential behavior data, we propose a Multi-Supervised Contrastive Learning strategy that integrates cross-modal alignment and virtual sequence augmentation. This approach enhances user interest modeling by leveraging diverse learning signals, resulting in improved model robustness and generalization capability. Extensive experiments on four public datasets demonstrate that DuAF-MAT significantly outperforms state-of-the-art baselines.

Introduction

Sequential Recommendation (SR) (Wang et al. 2019a; Wu et al. 2024; Quadrana, Cremonesi, and Jannach 2018; Wang et al. 2019b) has emerged as a prominent research direction in recent years, aiming to capture users' dynamic interests and predict the next item of interest based on their historical behaviors. Current SR approaches (Chen et al. 2022; Kang and McAuley 2018; Lin et al. 2023; Sun et al. 2019) predominantly rely on user-item interaction IDs as input. This reliance results in performance degradation under data sparsity and cold-start scenarios, ultimately limiting model robustness and scalability.

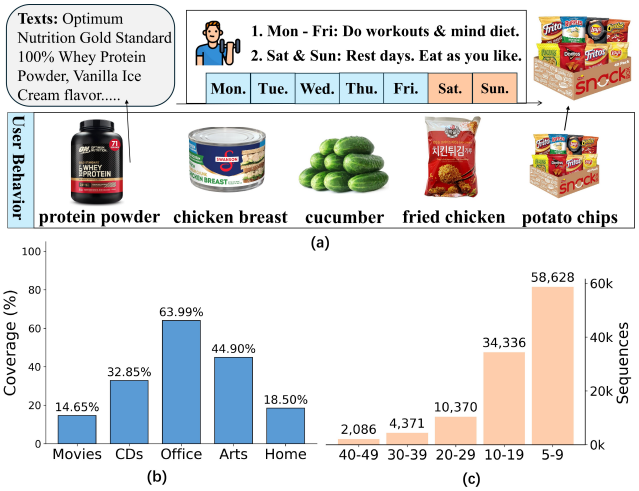

Figure 1: MMSR Challenges: a) Users' attention to different modalities of items may depend on their preferences, and these preferences dynamically evolve over time. b) Numerous real-world datasets exhibit modal incompleteness phenomena. c) The sparsity of sequential data affects recommendation accuracy.

Real-world recommendation systems often involve multimodal information (e.g., text, images), which provides additional semantic enrichment for user interest modeling. In recent years, Multimodal Sequential Recommendation (MMSR) (Bian et al. 2023; Hou et al. 2022; Hu et al. 2023; Liang et al. 2023; Guo et al. 2025b) has gained increasing attention. MMSR integrates multimodal data to enhance user interest modeling, offering richer representations while alleviating data sparsity and cold-start issues.

Existing MMSR methods primarily focus on modality fusion to improve recommendation accuracy. For instance, UniSRec (Hou et al. 2022) maps textual semantics into ID embeddings through an expert mixture mechanism, and MissRec (Wang et al. 2023) designs a lightweight fusion module to model user preferences across modalities adaptively. While these methods have advanced multimodal fusion, three critical challenges remain: (1) Although multimodal data greatly enriches user interest modeling, traditional fusion strategies often fail to effectively capture cross-

*Corresponding author.

modal interactions within the complex dynamics of constantly changing user intentions. This leads to extracted multimodal features that are inconsistent with dynamic user interests. Users may prioritize different modalities in different situations. As shown in Figure 1 a), fitness enthusiasts may focus more on textual information such as nutritional content and calories when selecting food during their dietary discipline periods, while visual attributes like color and appearance in item images may play a more important role during their indulgence phases. (2) In practical applications, recommender systems typically involve multiple heterogeneous data sources, resulting in incomplete modal information for certain items (e.g., missing images or descriptions), as illustrated in Figure 1 b). (3) User interaction data is typically sparse, especially for cold-start users or those with limited historical behaviors, which creates challenges in accurately modeling their interest patterns. As shown in Figure 1 c), interaction sequences with lengths of 5-9 account for the majority of the training set, highlighting the need to extract meaningful signals from these relatively short sequences.

To address the aforementioned challenges, we propose a novel framework for Multimodal Sequential Recommendation, termed DuAF-MAT. Our approach consists of three key components. First, to tackle the dynamic nature of modality contributions, we design a Dual-Aware Adaptive Fusion (DuAF) module that dynamically adjusts modality weights based on user preferences and temporal information, effectively capturing evolving user interests across scenarios. Second, to handle incomplete modality data, we propose a Modality Adversarial Training with Mixture-of-Experts (MAT-MoE), which leverages adversarial learning and specialized expert generators to reconstruct modalities, ensuring robust item representations even under data heterogeneity. Third, to mitigate data sparsity and enhance generalization, we introduce Multi-Supervised Contrastive Learning (MSCL), integrating cross-modal and virtual sequence contrastive objectives to enrich signals and improve preference modeling for sparse sequences. In summary, the contributions of our work are as follows:

- We propose a Dual-Aware Adaptive Fusion module that synergistically combines user-aware and time-aware adaptation mechanisms, enabling fine-grained modality weighting tailored to user preferences and temporal information.
- We propose a Modality Adversarial Training framework enhanced by a Mixture of Experts, which generates high-fidelity synthetic embeddings for missing modalities using modality-guided adversarial learning.
- We propose a Multi-Supervised Contrastive Learning strategy that integrates cross-modal alignment, virtual sequence augmentation to address sparse sequential data issues. This approach enhances user interest modeling by leveraging diverse learning signals, resulting in improved model robustness and generalization capability.
- We conducted extensive experiments on four real-world public datasets, demonstrating that our DuAF-MAT method surpasses state-of-the-art approaches.

Related Work

Sequential Recommendation

Sequential Recommendation (SR) focuses on modeling user preferences based solely on historical interaction sequences, aiming to capture dynamic user interests and predict future interactions. Early methods relied on Markov Chains, such as FPMC (Rendle, Freudenthaler, and Schmidt-Thieme 2010) and HRM (Wang et al. 2015). Inspired by the success of deep learning in sequence modeling, researchers applied various neural network architectures for sequential recommendation, including CNNs (Tang and Wang 2018), RNNs (Hidasi 2015; Tan, Xu, and Liu 2016), and GNNs (Chang et al. 2021; Wu et al. 2019). The introduction of self-attention-based models has significantly advanced SR, with SASRec (Kang and McAuley 2018) leveraging self-attention for long-range dependency modeling and BERT4Rec (Sun et al. 2019) enhancing this with bidirectional encoding for contextualized sequence representation. Among them, some Transformer-based methods and graph attention-based methods have achieved remarkable results (Chen et al. 2022; Kang and McAuley 2018; Xia et al. 2022; Zhang et al. 2022; Zhou et al. 2020, 2023; Zhu et al. 2025). However, these methods primarily rely on user-item interaction IDs, making them susceptible to data sparsity and cold-start issues.

Mutlimodal Sequential Recommendation

Multimodal sequential Recommendation (MMSR) has emerged as a promising direction to incorporate multimodal information into sequential recommendation, significantly enhancing recommendation quality by capturing richer user interests (Ji et al. 2023; Li et al. 2023; Liang et al. 2023; Ye et al. 2024; Yuan et al. 2023; Guo et al. 2025a). Existing research primarily focuses on modality fusion strategies to improve recommendation accuracy. UniSRec (Hou et al. 2022) leverages a Mixture of Experts (MoE) mechanism to facilitate semantic transfer from textual representations to ID embeddings, improving the expressiveness of user representations. Building on these foundations, subsequent studies have introduced adaptive fusion mechanisms to enhance the effectiveness of multimodal integration. For instance, MISSRec (Wang et al. 2023) employs a lightweight fusion module to dynamically model user attention across different modalities, enabling personalized modality weighting. MMSR (Hu et al. 2023) incorporates heterogeneous graph neural networks to adaptively capture the intricate relationships between modalities, allowing for more flexible and robust multimodal fusion. Beyond conventional fusion techniques, TedRec (Xu et al. 2024) introduces a novel frequency-domain fusion approach, leveraging the Fast Fourier Transform (FFT) to process embedding sequences of item IDs and textual content, enabling sequence-level semantic fusion in the frequency spectrum. While these methods have achieved substantial progress in multimodal fusion, challenges remain in effectively modeling modality dynamics, handling missing modalities, and mitigating the impact of data sparsity, highlighting the need for further research in MMSR.

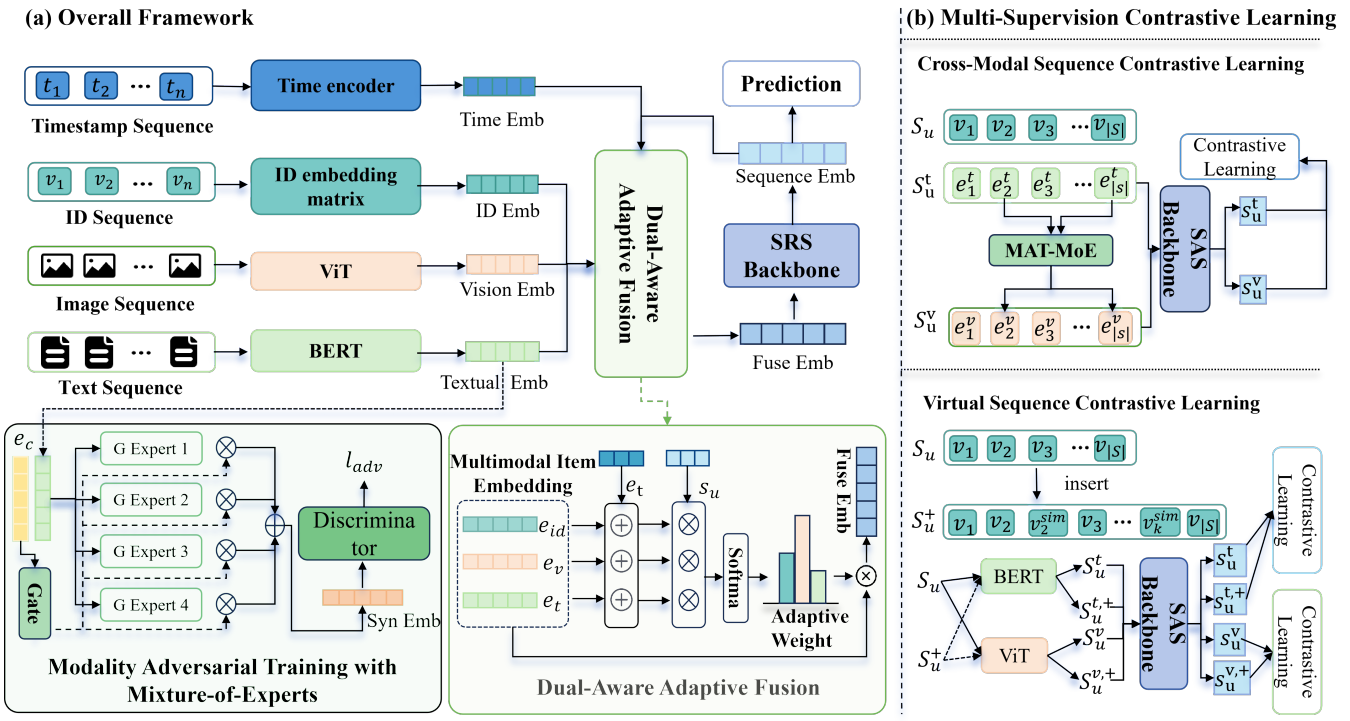

Figure 2: The overall architecture of the proposed DuAF-MAT.

Preliminaries

In our scenario, we focus on multimodal sequential recommendation tasks. Let $U = \{u_1, u_2, \ldots, u_N\}$ represent the set of users, and $V = \{v_1, v_2, \ldots, v_M\}$ represent the set of items. Each item $v \in V$ is represented as $v = \langle i_{\text{id}}, i_{\text{t}}, i_{\text{v}} \rangle$, where i_{id} is the item ID, while i_{t} and i_{v} refer to the associated textual and visual content of item v.

For each user $u \in U$, we define their interaction sequence $S_u = \{r_1, r_2, \ldots, r_{|S|}\}$, where each interaction record r_j corresponds to an interaction between user u and item $v_j \in V$ at time t_j. The corresponding timestamp sequence is denoted as $T = \{t_1, t_2, \ldots, t_{|S|}\}$.

Given a user's interaction sequence S_u, our objective is to predict the next item v_{n+1} the user is likely to interact with at the $(n+1)$-th time step. This can be formulated as finding the item v that maximizes the conditional probability given the interaction sequence:

$$v_{n+1} = \arg\max_{v \in V} P(v|S_u). \tag{1}$$

Proposed Method

Multimodel Item Representation

Multimodal information provides comprehensive item characterization essential for effective recommendation systems. We integrate four modalities for initial item embeddings: ID, textual, visual, and structural features.

For ID-based representation, we define an embedding matrix where each item v is associated with an ID embedding x_{id} of dimension d. The textual representation leverages BERT (Devlin et al. 2019) to extract semantic features from an item's textual content, using the special [CLS] token to capture sentence-level information. This produces a textual feature vector f_t that we transform to the embedding space through a linear projection to obtain x_t.

Similarly, for visual representation, we employ ViT (Dosovitskiy et al. 2020) to process images by segmenting them into patches and extracting visual features. The resulting visual feature vector f_v is also projected to the common embedding space to obtain x_v. These linear transformations involve trainable weight matrices and bias terms, while the original feature vectors remain frozen during training.

Finally considering the significance of sequential order in recommendation tasks, we incorporate positional embeddings into all modality representations:

$$\mathbf{e}_m = \mathbf{x}_m + \mathbf{p}, \tag{2}$$

where $m \in \{id, textual, visual\}$ represents the modality, and $\mathbf{p} \in \mathbb{R}^d$ denotes the positional embedding.

Dual-Aware Adaptive Fusion

Traditional fusion strategies, such as those employed in MissRec, typically rely on static or user-aware weighting schemes, which overlook the temporal dynamics of user intent. As a result, they often fail to capture nuanced cross-modal interactions that evolve over time, leading to misalignment between extracted multimodal features and users' real-time preferences. To address this limitation, we propose the Dual-Aware Adaptive Fusion (DuAF) module, which dynamically adjusts the contribution of each modality based on user preferences and temporal information.

In more specific terms, given a user sequence $S_u = \{v_1, v_2, \ldots, v_{|S|}\}$ for user u, and the corresponding timestamp sequence $T = \{t_1, t_2, \ldots, t_{|S|}\}$, we calculate the time interval sequence as follows:

$$\{\Delta t_1, \Delta t_2, \ldots, \Delta t_{|S|}\} = \{0, t_2 - t_1, \ldots, t_{|S|} - t_{|S-1|}\}, \tag{3}$$

where Δt_i denotes the corresponding interval of the item $v_i \in S_u$. Than we calculate the time interval embedding as follows:

$$e_i^{\text{inter}} = W_1 \cdot \log(\Delta t_i + 1) + b_1, \tag{4}$$

where W and b are learnable parameters in our method. Different time intervals e_i^{inter} indicate different user interest transitions. Meanwhile inspired by the positional embedding in Transformer, we calculate the timestamp embedding as follows:

$$e^{\text{ts}}[2k] = \sin(\alpha_k t + \beta_k), \quad e^{\text{ts}}[2k+1] = \cos(\alpha_k t + \beta_k), \tag{5}$$

where α_k and β_k are learnable parameters. then we combine both representations to get the final time embedding:

$$e_i^{\text{time}} = W_2[e_i^{\text{inter}}; e_i^{\text{ts}}] + b_2, \tag{6}$$

the adaptive weights for each modality m in the sequence are computed as follows:

$$\omega_i^m(u, t_i) = \frac{\exp\left(\left\langle s_u, \tanh(e_i^m + e_i^{\text{time}})\right\rangle\right)}{\sum_{n \in \mathcal{M}} \exp\left(\left\langle s_u, \tanh(e_i^n + e_i^{\text{time}})\right\rangle\right)}, \tag{7}$$

where $\langle \cdot, \cdot \rangle$ is the inner product, and $\tanh(\cdot)$ is the hyperbolic tangent function. e_i^m is the item embedding of the modality m, calculated from Equation (2), and $\mathbf{s}_u$ represents the dynamic preference of the user u, calculated by Equation (17). Finally, the fused embedding of the item v_i is aggregated as:

$$h_i = \sum_{m \in M \cup \{S\}} \omega_i^m(u, t) \mathbf{e}_i^m, \tag{8}$$

This weighted aggregation ensures that the fused embedding extracts the modal information most relevant to dynamic user interests, leading to more personalized item representations.

Modality Adversarial Training with Mixture-of-Experts

Traditional multimodal recommendation models suffer from performance degradation when facing missing modalities, particularly images. Existing methods often overlook these gaps or rely on naive random initialization, failing to capture complex cross-modal relationships and resulting in suboptimal representations. To address this, we propose a Modality Adversarial Training with Mixture-of-Experts (MAT-MoE) module that actively reconstructs missing modality embeddings through adversarial learning, enhancing the robustness of item representations.

Drawing from Wasserstein GAN (WGAN) (Arjovsky, Chintala, and Bottou 2017), we formulate a min-max game between the discriminator D and generator G:

$$\max_D \min_G \mathbb{E}_{x \sim p_{\text{real}}}[D(x)] - \mathbb{E}_{x' \sim G}[D(x')], \tag{9}$$

where x is the data sample. In this min-max game, the discriminator D is assigned to discriminate the input data sample x with a score while the generator G aims to generate a synthetic data sample x'. With the adversarial training, the generator G adapts to the actual distribution of the data x and the discriminator D can learn to judge the plausibility of the input data.

The Design of Generator In the design of MAT-MoE, We implement each generator expert G_i with a two-layer feed-forward network (FFN), which could be denoted as:

$$G_i(\mathbf{e}_{\text{t}}, z) = W_{2,i} \cdot \delta(W_{1,i}[\mathbf{e}_{\text{t}}; z] + b_{1,i}) + b_{2,i}, \tag{10}$$

where δ is the LeakyReLU activation function, $z \sim \mathcal{N}(0, I)$ is the random noise, and $[;]$ represents the concatenation operation. e_t represents the text embedding of the item e. Different item categories often exhibit distinct visual characteristics and semantic patterns. To facilitate more fine-grained expert selection conditioned on item categories, we maintain a learnable category embedding matrix $\mathbf{E}_c \in \mathbb{R}^{C \times d}$, where C is the total number of categories. Then we compute the routing weights:

$$\mathbf{g} = \text{softmax}(W_g \mathbf{e}_{\text{c}} + \mathbf{b}_g), \tag{11}$$

where $\mathbf{e}_{\text{c}}$ is the category embedding retrieved from $\mathbf{E}_c$. $\mathbf{g} = [g_1, g_2, \ldots, g_N] \in \mathbb{R}^N$ and N is the number of generator experts. The final synthetic item embedding $\mathbf{e}_{\text{syn}}$ is obtained by aggregating the outputs of all experts according to their respective routing weights:

$$\mathbf{e}_{\text{syn}} = G(\mathbf{e}_{\text{t}}, z) = \sum_{i=1}^{N} g_i \cdot G_i(\mathbf{e}_{\text{t}}, z), \tag{12}$$

where $\mathbf{e}_{\text{syn}} \in \mathbb{R}^d$ represents the generated visual embedding of item e.

The Design of Discriminator Discriminator D serves as a classifier designed to determine whether a pair of textual embedding e_t and visual embedding e_v are compatible, which is a binary classifier. The existing textual-visual embedding pair (e_t^i, e_v^i) for $e_i \in E_c$ are positive feature pairs with label 1, while the generated pair $(e_t^i, \mathbf{e}_{\text{syn}}^i)$ are viewed as negative embedding pairs with ground-truth label 0. In practice, D is a two-layer network denoted as:

$$D(e_t, e_v) = W_4[\delta(W_3 e_t + b_3); e_v] + b_4, \tag{13}$$

where W_3, W_4, b_3, b_4 are the learnable parameters of the network.

During training, we apply binary cross-entropy as the loss function to optimize the models:

$$\begin{aligned} \mathcal{L}_{\text{adv}} = &- \frac{1}{|E|} \sum_{e^i \in E} \log(1 - D(e_t^i, e_{syn}^i)) \\ &+ \frac{1}{|E_c|} \sum_{e^i \in E_c} \log(D(e_t^i, e_v^i)). \end{aligned} \tag{14}$$

In the adversarial context, the generator G aims to generate convincing visual embedding and fool the discriminator

D, while D is designed to make robust predictions to recognize the generated embeddings. Thus, G and D play a minimax game and optimize their parameters in an adversarial manner, denoted as:

$$\min_G \max_D \mathcal{L}_{\text{adv}} - \lambda_{\text{gp}} \mathcal{L}_{\text{gp}}, \tag{15}$$

where λ_{gp} is a hyperparameter controlling the intensity of the gradient penalty. $\mathcal{L}_{\text{gp}}$ is the gradient penalty (Gulrajani et al. 2017) , commonly used for stable WGAN training, denoted as:

$$\mathcal{L}_{\text{gp}} = \lambda_{\text{gp}} \mathbb{E}_{\hat{e}_v \sim \mathbb{P}_{\hat{e}_v}} \left[(\|\nabla_{\hat{e}_v} D(e_t, \hat{e}_v)\|_2 - 1)^2 \right], \tag{16}$$

where $\hat{e}_v = \alpha e_v + (1-\alpha)\mathbf{e}_{\text{syn}}$ are interpolated samples. $\alpha \sim \mathcal{U}(0,1)$ is an interpolation coefficient randomly sampled from a uniform distribution.

Prediction and Optimization

Based on the sequence $S_u = \{i_1, i_2 ..., i_{|S_u|}\}$, our objective is to predict the next item i_t that the user u is likely to interact with at the t-th step. Following the above design of MAT-MoE, for items without visual embeddings, we first generate visual embedding e_{syn} by G and assess their compatibility with the text embedding e_t using D. Then, we apply mean pooling to the valid visual embedding e_{syn} to obtain the final visual embedding e'_v. This process can be denoted as $e'_v = \frac{\sum_{j=1}^{K} y_{i,j} \cdot e_{syn}}{\sum_{j=1}^{K} y_{i,j}}$:where $y_{i,j} \in \{0,1\}$ is the prediction result of (e_t, e_{syn}) made by D. Then we use Equation (8) to get the fused item embedding h. After this, we got user sequence $S_u = \{h_1, h_2 ..., h_{|S_u|}\}$, we use SeqEncoder to capture the user's evolving interests, which can be denoted as:

$$s_u = \text{SeqEncoder}(S_u), \tag{17}$$

where s_u denotes the representation of sequence S_u. SeqEncoder is a sequence encoder, which can be implemented with the Attention (Vaswani et al. 2017) or other neural architectures (Sherstinsky 2020; Zhang et al. 2019a). In this work, we adopt SASRec(Kang and McAuley 2018) as our sequence encoder. After generating the above representations, we can obtain the final prediction $\hat{y} \in \{0,1\}$ at time t with dot product, where each element $\hat{y}_{ui}$ indicates how likely the item i should be recommended to the target user u. Finally, the model is trained with binary cross-entropy loss as follows:

$$\mathcal{L}_{main} = -\sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} y_{ui} \log \hat{y}_{ui} + (1 - y_{ui}) \log(1 - \hat{y}_{ui}). \tag{18}$$

Multi-Supervised Contrastive Learning

The inherent sparsity of sequential behavioral data often fails to provide sufficient signal fidelity for robustly modeling users' evolving preferences. This fundamental limitation leads traditional recommendation models to overfit spurious patterns, thereby degrading the quality of learned representations. To address this challenge, we propose Multi-Supervised Contrastive Learning (MSCL). This novel framework integrates CSCL and VSCL to generate complementary, multifaceted supervisory signals, thereby effectively enhancing model training.

Cross-Modal Sequence Contrastive Learning In order to better adapt the model to the relevant semantic information of different modalities, especially to extract the global shared behavioral preferences of users by coarse-grained alignment of semantic information between different modalities, we designed Cross-modal Sequence Contrastive Learning (CSCL).

Specifically, given a text sequence $S_u^t = \{e_1^t, e_2^t ..., e_{|S_u|}^t\}$ and the image sequence $S_u^v = \{e_1^v, e_2^v ..., e_{|S_u|}^v\}$ (include generated visual embedding e'_v) of the same user. We use SeqEncoder to capture user preferences s_u^t and s_u^v in different modalities and consider them as positives, while other in-batch ones are considered as negatives (N_u). The cross-modal sequence contrastive loss can be formally presented as follows:

$$L_{CSCL} = -\log \frac{\exp(\mathbf{s}_u^t \cdot \mathbf{s}_u^v)}{\exp(\mathbf{s}_u^t \cdot \mathbf{s}_u^v) + \sum_{u^- \in N_u} \exp(\mathbf{s}_u^t \cdot \mathbf{s}_u^{v^-})}. \tag{19}$$

By the contrastive learning of sequences in different modalities, coarse-grained alignment of cross-modal semantic information can be achieved, thereby enhancing the modeling of user preferences and improving the stability of the training process of MAT-MoE.

Virtual Sequence Contrastive Learning To further enhance the robustness of sequence representations, we propose a sequence augmentation contrastive learning method called Virtual Sequence Contrastive Learning (VSCL). VSCL generates semantically consistent virtual sequences through strategic similar-item insertion across dual modalities.

Given a user sequence $\mathcal{S}_u = \{i_1, i_2, ..., i_n\}$, we construct an virtual sequence $\mathcal{S}_u^{m,+}$ for modality m by inserting similar but not interacted items at semantic transition points:

$$\begin{aligned} P(i_k^{\text{sim}} = j \mid i_k, i_{k+1}) \propto \exp(\frac{1}{\lambda}(\gamma \cdot \text{sim}(e_k^m, e_j^m) \\ +(1-\gamma) \cdot \text{sim}(e_{k+1}^m, e_j^m))), \end{aligned} \tag{20}$$

where γ balances the contribution from adjacent items, and λ controls sampling diversity. The number of insertions follows a Poisson distribution with a mean proportional to the sequence length.

After encoding the original and virtual sequences into representations $\mathbf{s}_u^m$ and $\mathbf{s}_u^{m,+}$, we optimize the contrastive objective:

$$\mathcal{L}_{VSCL}^m = -\log \frac{\exp(s(\mathbf{s}_u^m, \mathbf{s}_u^{m,+})/\tau)}{\sum_{j=1}^{|\mathcal{B}|} \exp(s(\mathbf{s}_u^m, \mathbf{s}_j^m)/\tau)}. \tag{21}$$

The final VSCL loss combines both modalities:

$$\mathcal{L}_{VSCL} = (\mathcal{L}_{VSCL}^v + \mathcal{L}_{VSCL}^t)/2. \tag{22}$$

This dual-modality approach effectively enriches sparse sequences with plausible interactions from complementary information sources, helping the model learn more robust sequential patterns while maintaining semantic coherence.

Method	Toys				Games				Beauty				Home			
	NDCG		MRR		NDCG		MRR		NDCG		MRR		NDCG		MRR	
	@5	@10	@5	@10	@5	@10	@5	@10	@5	@10	@5	@10	@5	@10	@5	@10
GRU4Rec	0.0236	0.0289	0.0201	0.0222	0.0385	0.0514	0.0311	0.0365	0.0271	0.0337	0.0223	0.0260	0.0067	0.0087	0.0056	0.0064
SASRec	0.0348	0.0411	0.0257	0.0295	0.0391	0.0546	0.0286	0.0349	0.0325	0.0415	0.0246	0.0283	0.0118	0.0148	0.0089	0.0100
LRURec	0.0362	0.0446	0.0278	0.0312	0.0410	0.0557	0.0315	0.0376	0.0323	0.0412	0.0250	0.0286	0.0112	0.0141	0.0084	0.0096
FDSA	0.0255	0.0307	0.0214	0.0235	0.044	0.0576	0.0372	0.0423	0.0298	0.0365	0.0254	0.0282	0.0115	0.0138	0.0098	0.0108
UniSRec	0.0276	0.0384	0.0194	0.0239	0.0386	0.0535	0.0291	0.0353	0.0287	0.0398	0.0212	0.0288	0.0121	0.0160	0.0092	0.0108
TedRec	0.0318	0.0397	0.0265	0.0297	0.0468	0.0604	0.0384	0.0439	0.0330	0.0419	0.0275	0.0311	0.0113	0.0140	0.0094	0.0105
NOVA	0.0381	0.0478	0.0288	0.0328	0.0417	0.0576	0.0303	0.0368	0.0347	0.0442	0.0273	0.0312	0.0120	0.0147	0.0092	0.0102
DIF-SR	0.0359	0.0444	0.0284	0.0318	0.0416	0.0575	0.0311	0.0375	0.0340	0.0436	0.0262	0.0302	0.0092	0.0112	0.0078	0.0086
MISSRec	0.0323	0.0414	0.0235	0.0270	0.0398	0.0506	0.0312	0.0353	0.0315	0.0397	0.0240	0.0271	0.0140	0.0174	0.0107	0.0119
M3SRec	0.0416	0.0486	0.0363	0.0391	0.0493	0.0643	0.0404	0.0464	0.0345	0.0428	0.0294	0.0328	0.0122	0.0144	0.0105	0.0115
IISAN	0.0395	0.0489	0.0354	0.0393	0.0525	0.0671	0.0432	0.0491	0.0372	0.0464	0.0309	0.0347	0.0152	0.0187	0.0129	0.0142
DuAF-MAT	**0.0483**	**0.0553**	**0.0405**	**0.0437**	**0.0534**	**0.0682**	**0.0457**	**0.0511**	**0.0403**	**0.0475**	**0.0338**	**0.0374**	**0.0162**	**0.0189**	**0.0139**	**0.0156**
Improv.	+16.11%	+13.09%	+11.57%	+11.20%	+1.71%	+1.64%	+5.79%	+4.07%	+8.33%	+2.37%	+9.39%	+7.78%	+6.58%	+1.07%	+7.75%	+9.86%

Table 1: Performance comparison of different recommendation methods across various categories.

Joint Training Objective Finally, we integrate the proposed multi-supervised contrastive learning objectives with the main task to form a unified training framework:

$$\mathcal{L} = \mathcal{L}_{main} + \lambda_1 \mathcal{L}_{CSCL} + \lambda_2 \mathcal{L}_{VSCL}, \tag{23}$$

where λ_1 and λ_2 are hyperparameters controlling the contribution of CSCL and VSCL.

Experiments

Experimental Setup

Datasets We adopt four domains, including Toys and Games" (Toys), "Video Games" (Games), "Beauty" and "Home and Kitchen" (Home), from the standard benchmark dataset, Amazon Review (He and McAuley 2016). The statistics of the processed datasets are shown in the supplementary material. Following previous studies (Hou et al. 2022; Lin et al. 2024; Wang et al. 2024; Xie, Zhou, and Kim 2022), for each dataset, duplicated interactions are removed, and the interactions of each user are sorted by timestamps chronologically to build behavior sequences.

Compared Methods To verify the effectiveness of our method, we select the following representative and competitive baselines for sequential recommendation from three categories: (1) For pure ID-based methods, including GRU4Rec (Hidasi 2015), SASRec (Kang and McAuley 2018), and LRURec (Yue et al. 2024). (2) For text-based methods, including FDSA (Zhang et al. 2019b), UniSRec (Hou et al. 2022), and TedRec (Xu et al. 2024). (3) For multimodal methods, including NOVA (Liu et al. 2021), DIF-SR (Xie, Zhou, and Kim 2022), MISSRec (Wang et al. 2023), M3SRec (Bian et al. 2023), and IISAN (Fu et al. 2024).

Evaluation Settings Following previous works (Hou et al. 2022; Zhou et al. 2020), we adopt two standard metrics, namely Normalized Discounted Cumulative Gain (NDCG@K) and Mean Reciprocal Rank (MRR@K). We set K to 5 and 10 for showcases. We adopt the leave-one-out evaluation strategy to conduct the experiments. The ranking scores are computed on the whole item set without sampling.

Implementation Details We implement DuAF-MAT using a popular open-source recommendation library RecBole (Zhao et al. 2021). To ensure a fair comparison, we optimize all the methods with Adam optimizer and carefully search the hyper-parameters of all the compared methods. The batch size is set to 2,048. We adopt early stopping with the patience of 10 epochs to prevent overfitting, and NDCG@10 is set as the indicator. We tune the learning rate in $\{0.0003, 0.001, 0.003, 0.01\}$ and the embedding dimension in $\{64, 128, 300\}$. The expert number for MAT-MoE is selected from $\{2, 4, 6, 8, 10\}$, τ is searched within $\{0.1, 0.2, 0.3, 0.5, 0.7, 1.0\}$. the coefficient λ_{gp} are tuned from $\{1e^{-2}, 1e^{-3}, 1e^{-4}, 1e^{-5}\}$. The loss weight λ_1 and λ_2 are tuned from $\{0.2, 0.4, 0.6, 0.8, 1.0\}$. We conduct each experiment on a Linux server with Ubuntu 20.04.5 operating system and NVIDIA GeForce RTX 3090 GPUs.

Overall Comparison

We present the comprehensive experimental results in Table 1, which yields several significant findings. First, DuAF-MAT consistently outperforms both conventional and multimodal baselines across all evaluation metrics. Specifically, compared to the best-performing multimodal method, DuAF-MAT achieves performance improvements ranging from 1.07% to 16.11%. When compared against the strongest text-based model, our approach demonstrates more remarkable gains of 12.91% to 52.83%. Second, text-based models consistently outperform conventional recommendation methods across most experimental settings. This gain highlights how textual information provides valuable semantic context that enriches item representations, leading to improved recommendation quality. Third, multimodal approaches demonstrate clear advantages over traditional and text-based models. This finding confirms that integrating diverse modal information captures complementary item characteristics. Notably, DuAF-MAT significantly outperforms existing multimodal methods through its dual adaptive fusion that dynamically calibrates modal contributions, extracting multimodal features aligned with dynamic user interests.

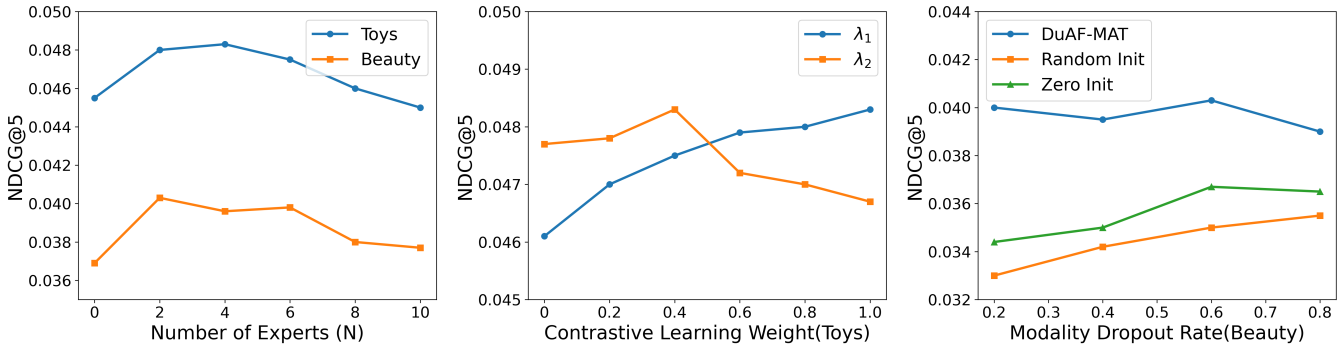

Figure 3: Evaluation of DuAF-MAT performance across various hyper-parameter configurations.

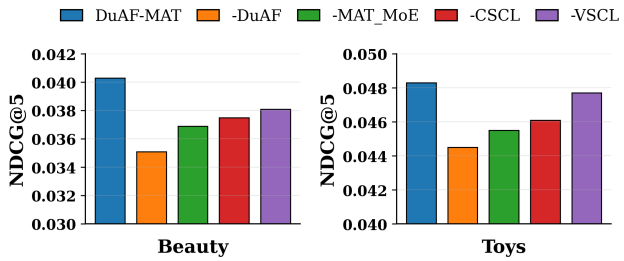

Figure 4: Ablation study of DuAF-MAT variants on "Beauty" and "Toys"

Ablation Study

To validate the effectiveness of each component in our proposed framework, we conduct ablation experiments on all datasets. Figure 4 presents the experimental results for different variants of our model, including (1) −DuAF that replaces the Dual-Aware Adaptive Fusion with average fusion. (2) −MAT_MoE removes the Modality Adversarial Training with MoE module. (3) −CSCL removes the cross-modal sequence contrastive learning task. (4) −VSCL removes the virtual sequence contrastive learning task. First, DuAF demonstrates the importance of our adaptive fusion mechanism in dynamically adjusting modality contributions based on user preferences and temporal information, which is especially valuable in domains with heterogeneous user interests. MAT-MoE confirms that our adversarial approach with multiple expert generators effectively reconstructs missing modality information, enhancing robustness against modality incompleteness that commonly occurs in real-world recommendation scenarios. Among the two contrastive learning components, CSCL demonstrates the most significant influence. This suggests that aligning semantic information across different modalities substantially improves the model's ability to capture comprehensive user preferences.

Hyper-Parameter Study

To comprehensively evaluate the sensitivity of the proposed DuAF-MAT framework to key hyperparameters, we conducted extensive experiments by varying three critical parameters: the number of experts (N), the contrastive learning weights (λ_1, λ_2), and the modality dropout rate. Figure 3 illustrates our findings on two datasets: Toys and Beauty. First, the model achieves peak performance when $N = 2$ or 4, confirming that the Mixture-of-Experts (MoE) mechanism effectively handles modal heterogeneity. However, an excessive number of experts may destabilize the adversarial training process. Second, our proposed DuAF-MAT consistently outperforms random initialization and zero initialization across all dropout rates, achieving optimal performance at a dropout rate of 0.6. This demonstrates the robustness of our model even under modality-missing scenarios. Interestingly, as the modality dropout rate increases, random initialization does not exhibit a strong negative correlation with performance. This suggests that DuAF may appropriately reduce the weight of visual modality contributions to mitigate its impact.

Conclusion

In this paper, we present DuAF-MAT, a novel multimodal sequential recommendation framework that effectively addresses key challenges in real-world recommender systems. Our approach integrates three innovative components: (1) Dual-Aware Adaptive Fusion (DuAF) module that dynamically calibrates modal contributions based on user preferences and temporal information to extract multimodal features consistent with dynamic user interests; (2) Modality Adversarial Training with Mixture of Experts (MAT-MoE) that leverages specialized expert generators to reconstruct missing modal features, effectively mitigating modality imbalance; and (3) Multi-Supervised Contrastive Learning (MSCL) strategy that combats data sparsity through complementary supervision signals derived from cross-modal alignment and virtual sequence augmentation. Comprehensive experiments across four diverse real-world datasets demonstrate our approach's substantial performance gains over state-of-the-art methods, validating the effectiveness of our proposed architecture in addressing the multifaceted challenges of multimodal sequential recommendation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62577050), the Zhejiang Provincial Philosophy and Social Sciences Program (Grant

No. 24NDJC191YB), and the Natural Science Foundation of Zhejiang Province (Grant No. LY23F020010).

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. arXiv:1701.07875.

Bian, S.; Pan, X.; Zhao, W. X.; Wang, J.; Wang, C.; and Wen, J.-R. 2023. Multi-modal mixture of experts represetation learning for sequential recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 110–119.

Chang, J.; Gao, C.; Zheng, Y.; Hui, Y.; Niu, Y.; Song, Y.; Jin, D.; and Li, Y. 2021. Sequential recommendation with graph neural networks. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 378–387.

Chen, Y.; Liu, Z.; Li, J.; McAuley, J.; and Xiong, C. 2022. Intent contrastive learning for sequential recommendation. In *Proceedings of the ACM web conference 2022*, 2172–2182.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fu, J.; Ge, X.; Xin, X.; Karatzoglou, A.; Arapakis, I.; Wang, J.; and Jose, J. M. 2024. IISAN: Efficiently adapting multimodal representation for sequential recommendation with decoupled PEFT. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 687–697.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.

Guo, H.; Shi, W.; Li, M.; Li, J.; Chen, H.; Cui, Y.; Xu, J.; Zhu, J.; Shen, J.; Chen, Z.; et al. 2025a. Consistent and Invariant Generalization Learning for Short-video Misinformation Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2254–2263.

Guo, H.; Zhu, J.; Di, S.; Shi, W.; Chen, Z.; and Xu, J. 2025b. DioR: Adaptive Cognitive Detection and Contextual Retrieval Optimization for Dynamic Retrieval-Augmented Generation. *arXiv preprint arXiv:2504.10198*.

He, R.; and McAuley, J. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, 507–517.

Hidasi, B. 2015. Session-based Recommendations with Recurrent Neural Networks. *arXiv preprint arXiv:1511.06939*.

Hou, Y.; Mu, S.; Zhao, W. X.; Li, Y.; Ding, B.; and Wen, J.-R. 2022. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 585–593.

Hu, H.; Guo, W.; Liu, Y.; and Kan, M.-Y. 2023. Adaptive multi-modalities fusion in sequential recommendation systems. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 843–853.

Ji, W.; Liu, X.; Zhang, A.; Wei, Y.; Ni, Y.; and Wang, X. 2023. Online distillation-enhanced multi-modal transformer for sequential recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 955–965.

Kang, W.-C.; and McAuley, J. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, 197–206. IEEE.

Li, J.; Wang, M.; Li, J.; Fu, J.; Shen, X.; Shang, J.; and McAuley, J. 2023. Text is all you need: Learning language representations for sequential recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1258–1267.

Liang, J.; Zhao, X.; Li, M.; Zhang, Z.; Wang, W.; Liu, H.; and Liu, Z. 2023. Mmmlp: Multi-modal multilayer perceptron for sequential recommendations. In *Proceedings of the ACM Web Conference 2023*, 1109–1117.

Lin, G.; Gao, C.; Zheng, Y.; Chang, J.; Niu, Y.; Song, Y.; Li, Z.; Jin, D.; and Li, Y. 2023. Dual-interest factorization-heads attention for sequential recommendation. In *Proceedings of the ACM Web Conference 2023*, 917–927.

Lin, X.; Luo, J.; Pan, J.; Pan, W.; Ming, Z.; Liu, X.; Huang, S.; and Jiang, J. 2024. Multi-sequence attentive user representation learning for side-information integrated sequential recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 414–423.

Liu, C.; Li, X.; Cai, G.; Dong, Z.; Zhu, H.; and Shang, L. 2021. Noninvasive self-attention for side information fusion in sequential recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 4249–4256.

Quadrana, M.; Cremonesi, P.; and Jannach, D. 2018. Sequence-Aware Recommender Systems. arXiv:1802.08452.

Rendle, S.; Freudenthaler, C.; and Schmidt-Thieme, L. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, 811–820.

Sherstinsky, A. 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404: 132306.

Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; and Jiang, P. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, 1441–1450.

Tan, Y. K.; Xu, X.; and Liu, Y. 2016. Improved recurrent neural networks for session-based recommendations. In *Proceedings of the 1st workshop on deep learning for recommender systems*, 17–22.

Tang, J.; and Wang, K. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 565–573.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Zeng, Z.; Wang, Y.; Wang, Y.; Lu, X.; Li, T.; Yuan, J.; Zhang, R.; Zheng, H.-T.; and Xia, S.-T. 2023. Missrec: Pre-training and transferring multi-modal interest-aware sequence representation for recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6548–6557.

Wang, P.; Guo, J.; Lan, Y.; Xu, J.; Wan, S.; and Cheng, X. 2015. Learning hierarchical representation model for nextbasket recommendation. In *Proceedings of the 38th International ACM SIGIR conference on Research and Development in Information Retrieval*, 403–412.

Wang, S.; Hu, L.; Wang, Y.; Cao, L.; Sheng, Q. Z.; and Orgun, M. 2019a. Sequential recommender systems: challenges, progress and prospects. *arXiv preprint arXiv:2001.04830*.

Wang, S.; Hu, L.; Wang, Y.; Cao, L.; Sheng, Q. Z.; and Orgun, M. 2019b. Sequential Recommender Systems: Challenges, Progress and Prospects. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI-2019, 6332–6338. International Joint Conferences on Artificial Intelligence Organization.

Wang, S.; Shen, B.; Min, X.; He, Y.; Zhang, X.; Zhang, L.; Zhou, J.; and Mo, L. 2024. Aligned side information fusion method for sequential recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, 112–120.

Wu, S.; Tang, Y.; Zhu, Y.; Wang, L.; Xie, X.; and Tan, T. 2019. Session-Based Recommendation with Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 346–353.

Wu, W.; Wang, C.; Shen, D.; Qin, C.; Chen, L.; and Xiong, H. 2024. Afdgcf: Adaptive feature de-correlation graph collaborative filtering for recommendations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1242–1252.

Xia, L.; Huang, C.; Xu, Y.; and Pei, J. 2022. Multi-behavior sequential recommendation with temporal graph transformer. *IEEE Transactions on Knowledge and Data Engineering*, 35(6): 6099–6112.

Xie, Y.; Zhou, P.; and Kim, S. 2022. Decoupled side information fusion for sequential recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 1611–1621.

Xu, L.; Tian, Z.; Li, B.; Zhang, J.; Wang, D.; Wang, H.; Wang, J.; Chen, S.; and Zhao, W. X. 2024. Sequence-level Semantic Representation Fusion for Recommender Systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 5015–5022.

Ye, Y.; Zheng, Z.; Shen, Y.; Wang, T.; Zhang, H.; Zhu, P.; Yu, R.; Zhang, K.; and Xiong, H. 2024. Harnessing multimodal large language models for multimodal sequential recommendation. *arXiv preprint arXiv:2408.09698*.

Yuan, Z.; Yuan, F.; Song, Y.; Li, Y.; Fu, J.; Yang, F.; Pan, Y.; and Ni, Y. 2023. Where to go next for recommender systems? id-vs. modality-based recommender models revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2639–2649.

Yue, Z.; Wang, Y.; He, Z.; Zeng, H.; McAuley, J.; and Wang, D. 2024. Linear recurrent units for sequential recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining*, 930–938.

Zhang, M.; Wu, S.; Yu, X.; Liu, Q.; and Wang, L. 2022. Dynamic graph neural networks for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 4741–4753.

Zhang, S.; Tong, H.; Xu, J.; and Maciejewski, R. 2019a. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1): 1–23.

Zhang, T.; Zhao, P.; Liu, Y.; Sheng, V. S.; Xu, J.; Wang, D.; Liu, G.; Zhou, X.; et al. 2019b. Feature-level deeper self-attention network for sequential recommendation. In *IJCAI*, 4320–4326.

Zhao, W. X.; Mu, S.; Hou, Y.; Lin, Z.; Chen, Y.; Pan, X.; Li, K.; Lu, Y.; Wang, H.; Tian, C.; Min, Y.; Feng, Z.; Fan, X.; Chen, X.; Wang, P.; Ji, W.; Li, Y.; Wang, X.; and Wen, J.-R. 2021. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. arXiv:2011.01731.

Zhou, K.; Wang, H.; Zhao, W. X.; Zhu, Y.; Wang, S.; Zhang, F.; Wang, Z.; and Wen, J.-R. 2020. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 1893–1902.

Zhou, P.; Ye, Q.; Xie, Y.; Gao, J.; Wang, S.; Kim, J. B.; You, C.; and Kim, S. 2023. Attention calibration for transformer-based sequential recommendation. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, 3595–3605.

Zhu, J.; Guo, H.; Shi, W.; Chen, Z.; and De Meo, P. 2025. Radio: Real-time hallucination detection with contextual index optimized query formulation for dynamic retrieval augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 26129–26137.