Aligning Large Multimodal Model with Sequential Recommendation via Content-Behavior Guidance

Zihao Wu
Department of Computer Science and Technology, Tsinghua University
Beijing, China
wuzh22@mails.tsinghua.edu.cn

Xin Wang*
Department of Computer Science and Technology, BNRist, Tsinghua University
Beijing, China
xin_wang@tsinghua.edu.cn

Heng Chang
Department of Computer Science and Technology, Tsinghua University
Beijing, China
changh17@tsinghua.org.cn

Hong Chen
Department of Computer Science and Technology, Tsinghua University
Beijing, China
h-chen20@mails.tsinghua.edu.cn

Lifeng Sun
Department of Computer Science and Technology, Tsinghua University
Beijing, China
sunlf@tsinghua.edu.cn

Wenwu Zhu*
Department of Computer Science and Technology, BNRist, Tsinghua University
Beijing, China
wwzhu@tsinghua.edu.cn

Abstract

Large language models (LLMs) have significantly influenced advancements in sequential recommendation. Nevertheless, the integration and alignment of LLMs with sequence recommenders is often underexploited in current research. Existing LLM-based sequential recommenders mostly rely on textual descriptions, neglecting user visual preferences and suffering from LLM hallucination, which can result in suboptimal recommendations. To address these challenges, we propose AlmostRec, a novel framework that incorporates multimodal information, including historical interaction IDs, textual descriptions, and images of items into large foundation models, to facilitate controllable predictions. Instead of employing a textual LLM, AlmostRec utilizes a large multimodal model (LMM) as a backbone, complemented by a content-behavior guidance module to align multimodal information. The framework's ID prediction objective, enhanced via the parameter-efficient LoRA approach, ensures a principal alignment with the sequential recommendation and is not swayed by hallucinations. AlmostRec effectively bridges the gap between large vision-language models with sequential recommenders, offering contextually relevant predictions in multimodal scenarios. Experimental results on real-world datasets demonstrate the superior performance of AlmostRec compared to both traditional and recent LLM-based recommendation approaches.

CCS Concepts

• **Information systems** → *Multimedia and multimodal retrieval.*

Keywords

large language models, sequential recommendation, multimodal machine learning

ACM Reference Format:
Zihao Wu, Xin Wang, Heng Chang, Hong Chen, Lifeng Sun, and Wenwu Zhu. 2025. Aligning Large Multimodal Model with Sequential Recommendation via Content-Behavior Guidance. In *Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR '25), June 30-July 3, 2025, Chicago, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3731715.3733273

1 Introduction

In the realm of online services, recommender systems play a crucial role by analyzing a series of user actions, such as product browsing, link clicking, and video watching. The sequence of actions has sparked a growing interest in sequential recommendation approaches, aiming to anticipate future user activities based on their historical interactions with items. Traditional approaches have utilized Markov Chains to analyze temporal sequences of interactions [42, 46]. The evolution of deep learning has led to significant advancements in this field. Recent research efforts have successfully integrated diverse neural network architectures as sequence encoders [13–15, 21, 24, 43, 44], enhancing the ability to detect and interpret the complex patterns inherent in user behavior dynamics.

The emergence of large language models (LLMs) has prompted recent studies [2, 27, 51] to explore their utilization for sequential recommendation systems. However, these efforts suffer from several problems: (i) existing works predominantly rely on predefined text prompts to recommend new items, ignoring the multimodal preferences of users, which can result in sub-optimal recommendation performance; (ii) the output of existing LLM-based recommenders are typically plain textual description, which may not intrinsically align with recommendation objectives and are vulnerable to hallucinations and biases in prior data distributions [38]. We show these pitfalls in directly employing large foundation models for recommendation tasks with examples in Figure 1.

*Corresponding Authors. BNRist is the abbreviation for Beijing National Research Center for Information Science and Technology.

ICMR '25, Chicago, IL, USA

ACM ISBN 979-8-4007-1877-9/2025/06
https://doi.org/10.1145/3731715.3733273

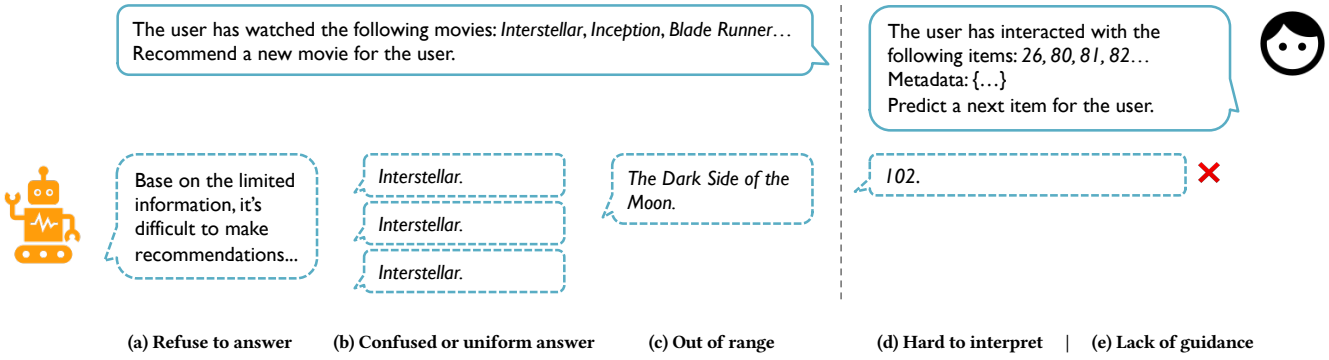


Figure 1: The pitfalls in employing large foundation models for recommendation tasks. Firstly, the challenges are rooted in the gap between the task formulation and the internal objective of large models and recommender systems (e.g., cases (a), (b), and (c)). Secondly, the defect of interpretation of the predominant ID-based paradigm and the lack of substantial content and behavior guidance leads to the dilemma of large models (e.g., cases (c) and (d)).

Specifically, the pre-training of large language models seldom imparts a genuine comprehension of the specific objectives associated with sequential recommendation tasks, which essentially require the modeling of user behaviors and application scenarios. Furthermore, conventional recommendations including collaborative filtering and sequential recommendation predominantly utilize the ID-based paradigm, a concept that large models find challenging to directly interpret and leverage the internal collaborative information [27, 39]. The literature inadequately explores the difficulties arising from LLMs' incapacity to effectively model ID-based representations in conventional sequential recommendation tasks. This gap highlights significant challenges that need to be overcome to harness the full potential of large models in the domain of recommender systems.

To tackle the problems, in this work, we propose AlmostRec, a novel LMM-based framework that can make controllable predictions based on multimodal information from the user's historical behaviors. As demonstrated in Figure 2, AlmostRec jointly exploits the ID information, textual description, and image of the historical items for the multimodal recommendation, considering the user's preferences in different modalities. To adequately and universally tackle the multimodal information, instead of using a textual LLM, we resort to the large multimodal model (LMM) as the backbone. Based on the LMM architecture, we propose the content-behavior guidance module to align the ID, text description, and image of the item in the same representation space. Furthermore, to align the multimodal information with the user's preference, we adopt the ID prediction training objective to optimize the LMM model, which is updated in the parameter-efficient manner of LoRA. The ID prediction objective is more suitable for the recommendation scenario and guarantees the output is within the candidate item list. AlmostRec bridges the gap between large multimodal models and sequential recommenders. With the help of the pre-trained knowledge of large multimodal models, it enables the generation of controllable, contextually relevant predictions under multimodal scenarios. Extensive experiments on different datasets show that our proposed AlmostRec significantly outperforms both traditional recommendation works and recent LLM-based works.

To summarize, this work makes the following contributions.

- To the best of our knowledge, this is the first work to investigate the problem of incorporating collaborative guidance and aligning LMM for sequential recommendation.
- We propose the simple but effective AlmostRec framework, which involves the content-behavior guidance module and the ID prediction objective to align the multimodal information with the user preference, ensuring reliable recommendations of the large multimodal model.
- We conduct extensive experiments on several real-world multimodal datasets to demonstrate not only the superiority of the proposed AlmostRec but also the effectiveness of our proposed components.

We review the related work in Sec. 2 and the preliminaries in Sec. 3, describe the proposed framework in Sec. 4 and then present the experimental results and further discussions in Sec. 5 and Sec. 6.

2 Related Work

2.1 Sequential Recommendation

Sequential recommendation systems suggest items to users by analyzing their previous interactions. The approaches differ from traditional recommendation methods by utilizing the historical behavior of users to better predict their current preferences. Initially, Markov chains were employed to model the transitions between items. With the advent of deep learning, more complex approaches have been developed, i.e., encoding historical sequences into hidden vectors using deep neural networks such as GRU [15], CNN [44], and their variants, which have proven highly effective in sequential recommendation tasks. Recent research has further enhanced these models by incorporating self-attention mechanisms, which assess the interplay among past interactions and significantly improve performance [21, 43, 58]. Additionally, graph neural networks have shown promise in capturing high-order relationships within sequences through information propagation and aggregation, extending their application to sequential recommendation [7, 9, 24, 57]. Recent emerging studies in sequential recommendation explore novel training and augmentation strategies [36, 50]. Wang et al. [48]

propose curriculum co-disentanglement model with a curriculum weighting strategy to learn the disentangled user representation across consuming and social environments. Jiang et al. [20] propose a graph diffusion recommendation framework DiffKG to address the problem of noisy and misleading information through knowledge-aware data augmentation and graph convolution mechanism.

2.2 Large Multimodal Models

With the swift advancement of large language models (LLMs), an increasing number of researchers have started to integrate LLMs into multimodal tasks, leading to the creation of large multimodal models (LMMs) [1, 23, 31, 59], among which the visual-language model has emerged as the most prominent. This trend underscores a burgeoning interest in harnessing the synergy between textual and visual modality, suggesting a novel approach to enriching downstream multimodal tasks [16, 19, 29, 30, 34]. LLaVa [31] proposes to connect a vision encoder and LLM through visual instruction tuning to build general-purpose large multimodel models, which exhibit advanced multimodel perception and inference abilities. InternVL [3] scales up the vision foundation model and progressively aligns it with the LLM backbone, achieving advanced performance on generic visual-linguistic benchmarks including visual perception, vision-language, and multi-modal dialogue tasks. CogAgent [16] further possesses agent capabilities including processing GUI images on top of the strength of the base visual expert models.

2.3 Foundation Models for Recommendation

Since the great success of LLMs, progress has been made to facilitate recommender systems with the strength of large foundation models [6, 28]. Early attempts formalize the recommendation task into language processing tasks that large models excel at, and verify the feasibility of the large foundation model in the recommendation field, especially for the interactive recommendation, review summary, recommendation interpretation, and other semantically oriented tasks [8, 10, 35, 47, 56].

Under the discrepancy between conventional ID-based recommendation and LLM-powerd prompting-based recommendation, research efforts have been devoted to investigating the extensible ways for recommendation foundation models [18, 25, 27, 33, 39, 52, 54, 55]. Yang et al. [51] examines the capacity of LLMs to decipher the representation space of sequential recommenders and designed prompting approaches to guide LLMs to understand hidden representations from sequential recommenders. Qiu et al. [39] propose to achieve the alignment and integration between language model and personalized recommendation through a contrastive prompt learning framework. VIP5 [11] managed to unify various modalities and recommendation tasks in a shared architecture through fine-tuning extra adapters on the top of the pre-trained P5 [10] model.

Primarily, most of the existing work tries to achieve a specific phase of a recommender system, e.g., item ranking or preference prediction, by calling the LLMs with a specifically designed prompt and generating a response as the recommended result. Nevertheless, the instructions for the objective of a recommender are implicit in the process, and large models do not necessarily obtain the intrinsic ability to make recommendations and generalize to new instances. By incorporating multimodal information and aligning the objective of LMMs and sequential recommendation, AlmostRec holds the potential to enhance the semantic relevance and accuracy of recommendations, offering a more nuanced and comprehensive understanding of user preferences.

3 Preliminary

3.1 Sequential Recommendation

Let $\mathcal{U}$ and $\mathcal{I}$ denote the user and item set, respectively. For each user $u \in \mathcal{U}$, a sequentially ordered interaction sequence $S_N = [i_1, i_2, ..., i_N]$ is provided, where each element $i_j \in \mathcal{I}$ is an item that was interacted, and N is the length of the sequence. The goal of sequential recommendation is to predict a list of items that the user may be interested in, based on its historical sequence S_N leading up to the target time step N.

3.2 Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) [26, 32] represents a pivotal advancement in the optimization of large foundation models. This technique modifies a relatively small subset of a pre-trained model's parameters, enabling significant improvements in performance without the extensive computational cost typically associated with training large models from scratch. This also overcomes the issues of catastrophic forgetting [12], a behavior observed during the full finetuning of LLMs. LoRA [17] introduces a novel PEFT technique to strategically adapt pre-trained models by applying low-rank updates to the weight matrices within the transformer layers. Specifically, the adaptation is achieved by decomposing the update to the original weight matrix W' into a product of two low-rank matrices A and B where $W' = AB$, and introduces minimal set of parameters Θ. LoRA allows for significant modifications to the model's behavior with minimal adjustments to the parameters, effectively balancing performance improvements with computational efficiency.

4 AlmostRec: The Proposed Framework

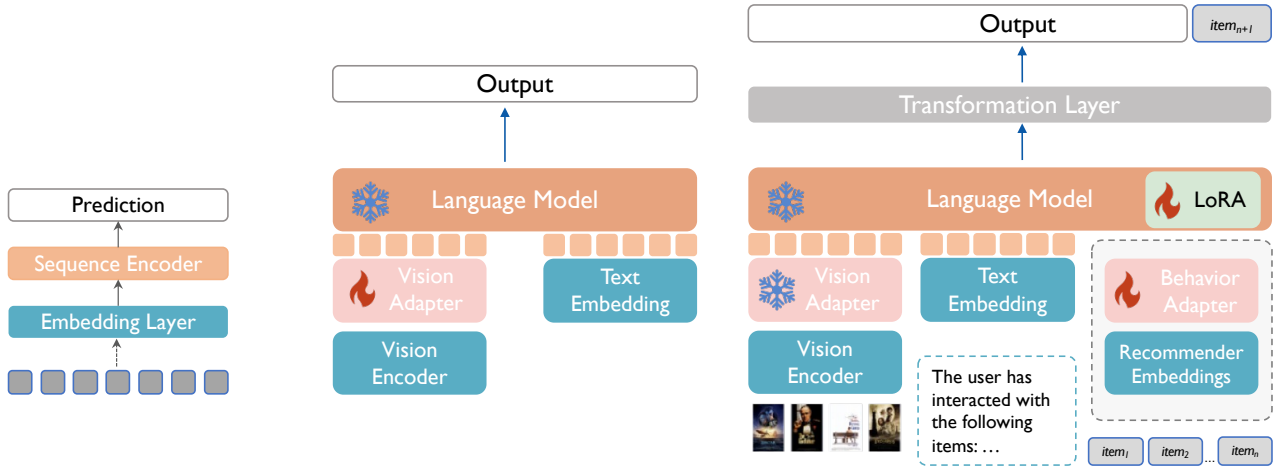
In this section, we start by introducing the design of each component of our framework individually and then combine them to formulate our overall architecture.

4.1 Large Multimodal Models as Backbones

To reasonably tackle the challenges of employing foundation models for recommendation and extend the capability of large foundation models to the multimodal domain, we choose to leverage Large Multimodal Models (LMMs) to adequately perceive the multimodal content information as well as perform the instruction-following to facilitate the sequential recommendation potentially.

As in Figure 2 (b), LMMs typically consist of a visual encoder for feature extraction and a language model for text decoding, linked via a trainable connection module. The training of LMMs involves large-scale image-text pairs, with auto-regressive loss applied to the output text tokens.

Specifically, we employ the LLaVa-1.5-7B [29] model as the large multimodal model backbone, which demonstrates a simple yet potent architecture by integrating a CLIP [40] vision encoder with the



(a) Pre-trained sequential recommender **(b) Pre-trained large multimodal model** **(c) Aligning LMM with sequence recommendation**

Figure 2: The overall illustration of the proposed AlmostRec framework. (a) demonstrates a representative structure of a sequential recommender model, which is pre-trained for integrating into the AlmostRec backbone. (b) depicts the common practice for building a large multimodal model, where the visual encoder and language model can be connected by a trainable connection adapter. (c) is the framework of AlmostRec. The pre-trained embedding from the sequential recommender is integrated with the LMM backbone, whilst the LMM is adapted for recommendation prediction objectives via an elaborated content behavior guidance module and LoRA tuning.

Vicuna [4] language model through an MLP adapter. The efficacy of LLaVa-1.5 is significantly enhanced through visual instruction tuning, which involves training on diverse types of instruction-following data, including detailed descriptions, complex reasoning tasks, and visual question answering (visual-QA). This visual instruction tuning plays a pivotal role in advancing the capabilities of LMMs beyond their initial pre-training [29, 31]. Consequently, LLaVa-1.5 is adept at processing multimodal content and executing instruction-following tasks, thereby paving the way for the sequential recommendation process.

4.2 Content-Behavior Guidance

While the foundation model possesses the ability to perceive and reason in multimodal scenarios, it is still difficult to serve as practicable recommenders without bridging the internal gap between natural language generation and user preference prediction tasks, which lies in depth in the representation and optimization of recommender systems. To this end, we propose content-behavior guidance to effectively guide the LMM to interpret user behavior patterns in multiple modalities and seamlessly integrate the ID-based item embeddings from sequential recommenders with the LMM backbone.

4.2.1 Content-Behavior Adapter. The dense item embeddings from the pre-trained sequential recommender are encapsulated with rich semantics and collaborative signals, which is vital for performing sequential recommendations. Specifically, to make use of the item embeddings, AlmostRec first employs the embedding layer in the pre-trained sequential recommender, and performs a look-up operation to retrieve the vector representation associated with a particular item input from the pre-trained embedding matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{I}| \times d_i}$, where $|\mathcal{I}|$ denotes the size of item set, d_i the dimension of the item embeddings. Consequently, the ID-based item embeddings are delivered to the LMM backbone through a connection module, *i.e.*, the content-behavior adapter. To make it simple and effective, the content-behavior adapter is essentially a learnable weight matrix $\mathbf{W}^B \in \mathbb{R}^{d_i \times d_m}$, where d_m is the dimension of the large multimodal model embeddings. The elaboration of the content-behavior adapter bridges the intrinsic embeddings from the sequential recommender to the LMM, which immits the requisite latent information for the recommendation.

4.2.2 LoRA Tuning. On the basis of content-behavior information, AlmostRec introduces a LoRA [17] adapter to appropriately guide the LMM to generate high-quality and robust recommendations. The LoRA adapter with a small set of parameters Θ modifies the behavior of the LMM backbone to align the output with proper recommendation predictions and is updated in the forward process during training.

The content-behavior adapter and LoRA adapter jointly encourage the LMM to effectively absorb the rich semantics and collaborative signals contained in the ID embeddings and visual input, thus achieving content-behavior guidance while ensuring the simple and minimal trainable parameters plugged in.

4.3 Overall Architecture

Combining the components introduced above, the overall architecture of our proposed AlmostRec is composed of an LMM backbone, a content-behavior guidance module, and a transformation layer.

To begin with, AlmostRec receives the input of different modalities including visual images, text prompts, and item sequence IDs simultaneously, denoted as a triplet $(\mathcal{V}_N, \mathcal{T}, \mathcal{S}_N)$. As designed in the content-behavior adapter, the item sequence $\mathcal{S}_N$ is converted into pre-trained embeddings, and the rest of the modalities in the input are projected to visual and text embeddings as LMMs typically do. The final inputs are the concatenation of the visual embeddings, word embeddings, and item embeddings:

$$\text{Input} = \textbf{Concatenate}(\mathbf{V}^{d_m}, \mathbf{T}^{d_m}, \mathbf{S}^{d_m}) \tag{1}$$

In the following of the forward process of the LMM backbone, the transformation layer is designed to map the generated embeddings of the language model to recommendation predictions through a learnable linear projection $\mathbf{W}^T \in \mathbb{R}^{d_m \times |\mathcal{I}|}$, where $|\mathcal{I}|$ denotes the size of candidate item set. Given the user interaction sequence $S_N = [i_1, i_2, ..., i_N]$, the probability distribution of next item prediction is obtained by

$$q(i_{N+1}|i_1, \cdots, i_N) = \text{Softmax}(\mathbf{F}^{d_m}\mathbf{W}^T), \tag{2}$$

where $\mathbf{F}^{d_m}$ is the generated output from the backbone model in a forward process.

Additionally, the transformation layer features a mapping table utilized upon the generated prediction which is to convert the predicted item ID to its corresponding multimodal content for intuitive and interactive recommendation output.

4.4 Optimization of AlmostRec

Benefiting from its transformation layer design, AlmostRec can predict across all candidates in each forward pass, which is adequate for conventional sequential recommendation systems. To optimize the model, we resort to the essence of sequential recommendation and utilize the standard cross-entropy loss as its learning objective,

$$\mathcal{L} = -\sum_{i=1}^{|\mathcal{I}|} p_i \log q_i, \tag{3}$$

where q_i denotes the output predictions and p_i denotes the ground-truth subsequent items. The details of AlmostRec are shown in Algorithm 1.

Algorithm 1 AlmostRec pipeline for multi-modal sequential recommendation

Input: Input of different modalities $(\mathcal{V}_N, \mathcal{T}, \mathcal{S}_N)$, pre-trained item embedding matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{I}| \times d_i}$, adapted LMM $\mathcal{F}$; target i_{N+1}
Output: Predicted next item i

repeat
  $\mathbf{V}^{d_m} \leftarrow \mathbf{W}^V(\textbf{Encoder}_\mathbf{v}(\mathcal{V}_N))$
  $\mathbf{T}^{d_m} \leftarrow \textbf{Emb}_\mathbf{t}(\mathcal{T})$
  $\mathbf{S}^{d_m} \leftarrow \mathbf{W}^B(\mathbf{M}(\mathcal{S}_N))$
  $\mathbf{x} \leftarrow \textbf{Concatenate}(\mathbf{V}^{d_m}, \mathbf{T}^{d_m}, \mathbf{S}^{d_m})$
  $\mathbf{F}^{d_m} \leftarrow \mathcal{F}(\mathbf{x}; \Theta_{\text{LoRA}})$ ▷ forward pass of LMM with LoRA
  $q(i_{N+1}|i_1, \cdots, i_N) \leftarrow \text{Softmax}(\mathbf{F}^{d_m}\mathbf{W}^T)$
  $\mathcal{L} \leftarrow -\sum_{i=1}^{|\mathcal{I}|} p(i_{N+1}) \log q(i_{N+1})$
  Update $\Theta_{\text{LoRA}}, \mathbf{W}^B, \mathbf{W}^T$
until *converged*

4.5 Discussions

Alignment of Modalities. Despite the rich content and user preference information contained within different modalities, it is crucial to seamlessly align multiple modalities and effectively utilize them for sequential recommendation predictions. We obtain user collaborative signals and robust multimodal representations with the pre-trained sequential recommender and the LMM. These elements are then seamlessly integrated through the Content-Behavior Guidance Module. Building on this foundation, AlmostRec employs an inherent training objective and the parameter-efficient LoRA to effectively leverage the aligned multimodal representations to achieve sequential recommendations.

Capture of User Intentions. While sequential recommendation approaches often suffer from the insufficient interpretation of latent intentions of user behavior sequence, AlmostRec is able to naturally capture the user intentions and relations of interacted items, benefiting from the transformer encoders within the LMM backbone. Thus AlmostRec is promising to perform effective sequential recommendations with a simple and flexible architecture.

5 Experiments

In this section, we will evaluate the proposed AlmostRec on several real-world datasets and compare it with several baselines, including traditional sequential recommendation models and LLM-based models. To demonstrate the superiority of our framework, we will showcase it to answer the following research questions:

- RQ1: How does AlmostRec perform compared with other multimodal sequential recommendation approaches and large model-based approaches?
- RQ2: How does the content-behavior guidance module and the pre-trained sequential embeddings affect the performance of AlmostRec?
- RQ3: How does AlmostRec perform qualitatively on real-world scenarios?

5.1 Experimental Setup

5.1.1 Datasets. To properly validate the effectiveness of user preference modeling in multimodal scenarios, we conduct extensive experiments with the proposed approach on several multimodal recommendation datasets, including:

- Netflix: The Netflix Prize Data on Kaggle[1] is a comprehensive collection of data that provides insights into user interactions and viewing habits on the platform. This dataset typically includes variables such as user IDs, movie IDs, ratings, and timestamps. Here we adopt the multi-modal version collected by [49], where they crawl and integrate the multimodal information, including movie poster images, into the original dataset structure.
- MicroLens-100K: The MicroLens [37] is a multimodal content-driven recommendation dataset consisting of a large number of users, micro-videos, and user-item interaction behaviors. It also features various modality information including video titles, cover images, and raw videos. We selectively use the

[1] https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data

original ID-based user-item interactions and the visual images to conduct recommendation experiments.

Table 1: Statistics of datasets.

Dataset	# Users	# Items	# Actions	Sparsity
Netflix	13,187	17,366	68,933	99.97%
MicroLens-100K	100,000	19,738	719,405	99.96%

Typically, the reviews or ratings are treated as implicit feedback, representing a user-item interaction, and organized chronologically by their associated timestamps. Additionally, users with fewer than five related actions are removed. Table 1 depicts the statistics of datasets.

5.1.2 Evaluation Settings. The performance evaluation is conducted by the widely used *leave-one-out* strategy [43, 58] where the most recent interaction is kept as the test data, the penultimate interaction is for validation, and all earlier interactions are for training. We employ top-k Hit Ratio (HR@k) and top-k Normalized Discounted Cumulative Gain (NDCG@k) to evaluate the recommendation performance, which are widely used as common practice [21, 36]. HR@k measures the average number of positive items retrieved in the top-N recommendation list generated for each user. NDCG@k further considers the position of retrieved positive items in the top-N list. MRR measures the ranking performance over the entire ranking list. We report results on HR@{10, 20} and NDCG@{10, 20}, and employ the all-ranking strategy to avoid potential biases from negative sampling [22].

5.1.3 Comparison Baselines. We compare AlmostRec against a series of representative baselines, including conventional recommendation models, typical sequential models, and recently proposed LLM-based novel approaches.

- **BPR** [41] is a matrix factorization variant of the classic Bayesian personalized ranking algorithm.
- **NCF** [13] is one of the most representative collaborative filtering methods based on neural networks.
- **GRU4Rec** [15] firstly employs the GRU network with a ranking-based loss for the sequential recommendation.
- **NextItNet** [53] is a CNN-based network capable of learning high-level representation from both short- and long-range item dependencies.
- **SASRec** [21] utilizes self-attention [45] to exploit the long-term mutual influence between historical interactions.
- **SASRec**$_{\text{ID+V}}$ and **YouTube**$_{\text{ID+V}}$ [5] are revisions of the corresponding sequential models that utilize pre-extracted video features as side information [37].
- **TALLRec** [2] conducts instruction tunings to align LLM with recommendation tasks.
- **LLaRA** [27] proposes to align LLMs with sequential recommenders by adapting LLMs to interpret text-ID-hybrid representations.
- **E4SRec** [25] enables LLMs to handle efficient and extensible sequential recommendation through parameter-efficient fine-tuning.
- **VIP5** [11] unifies various modalities and recommendation tasks in a shared architecture via fine-tuning extra adapters on the top of a pretrained LLM.

For fair comparisons, we implement both ID and visual encoding for the multi-modal sequential settings, following the implementations of [37]. We amend the output target of the TALLRec [2] model to meet the sequential recommendation setting instead of its original binary prediction of like/dislike.

5.2 Overall Performance

We first present the overall comparison between AlmostRec and baseline models in Table 2, from which we can summarize the observations as follows.

(1) AlmostRec shows consistent superior performance over all the baseline models across two real-world datasets in terms of all metrics.
(2) In particular, compared with the different paradigms of sequential recommendation approaches, including visual-fused and LLM-based baselines, AlmostRec achieves overall improvements on each dataset.

Such improvements suggest that 1) AlmostRec is capable of interpreting and modeling intricate patterns of historical user-item interactions. 2) AlmostRec effectively perceives various latent intentions of users across multiple modalities and scenarios. 3) The content-behavior guidance module effectively aligns the sequential behaviors with textual and visual information and encourages the model to predict user preferences in multimodal contexts to yield better recommendations.

Alignment Notably, it is argued that the world knowledge embedded in LMMs leads to a positive impact on making recommendation predictions, particularly in common domains such as books and movies [27]. Nevertheless, as for the MicroLens dataset, which features more diverse and less typical video content for LMMs, our proposed AlmostRec demonstrates greater improvements compared to the common Netflix dataset. This finding highlights that the interior design of AlmostRec framework facilitates superior multimodal alignment and robust, generalized sequential recommendation performance, extending beyond the reliance on the generic knowledge encoded in LMMs.

Efficiency As described above, AlmostRec introduces a minimal set of pluggable parameters to the backbone LMM, leaving the overall forward pass largely unchanged. Consequently, the model's size, complexity, and inference time closely align with those of the backbone model, LLaVa-1.5. This design allows a single shared backbone LMM to support multiple independent pluggable components tailored to specific datasets. As a result, AlmostRec offers a cost-effective solution for real-world applications.

5.3 The Behavior-Content Guidance

The rich content and user preferences embedded in multimodal information are crucial for predicting user sequential behavior. However, it is observed in Table 2 that in the compared baselines, the *Visual + Sequential Recommendation* setting does not significantly outperform their naive ID-based sequential models. Practically, they even perform worse than the ID-based models. This implies it is

Table 2: Performance comparison on different real-world datasets. The best results are in bold, second in underline.

Model	Netflix				MicroLens-100K			
	HR@10	HR@20	NDCG@10	NDCG@20	HR@10	HR@20	NDCG@10	NDCG@20
Conventional								
BPR	0.0291	0.0592	0.0139	0.0205	0.0291	0.0443	0.0131	0.0207
NCF	0.0309	0.0616	0.0141	0.0226	0.0297	0.0451	0.0135	0.0211
Sequential								
GRU4Rec	0.0771	0.1568	0.0436	0.0554	0.0780	0.1137	0.0419	0.0508
NextItNet	0.0794	0.1640	0.0467	0.0602	0.0818	0.1192	0.0439	0.0543
SASRec	0.0867	0.1782	0.0559	0.0682	0.0900	0.1264	0.0519	0.0603
Visual + Sequential								
$\text{YouTube}_{\text{ID+V}}$	0.0368	0.0907	0.0202	0.0290	0.0385	0.0638	0.0192	0.0256
$\text{SASRec}_{\text{ID+V}}$	0.0769	0.1734	0.0434	0.0585	0.0807	0.1234	0.0410	0.0514
Text + Sequential								
TALLRec	0.0845	0.1731	0.0561	0.0672	0.0896	0.1236	0.0506	0.0592
LLaRA	0.0846	0.1762	0.0567	0.0682	0.0911	0.1265	0.0525	0.0595
E4SRec	0.0956	0.1957	0.0644	0.0750	0.1018	0.1357	0.0606	0.0687
Text + Visual + Sequential								
VIP5	0.0932	0.1850	0.0568	0.0653	0.0920	0.1261	0.0519	0.0597
AlmostRec	**0.1128**	**0.2175**	**0.0737**	**0.0801**	**0.1249**	**0.1618**	**0.0617**	**0.0760**
Improv.	18.05%	11.14%	14.44%	6.87%	22.68%	19.22%	10.74%	10.62%

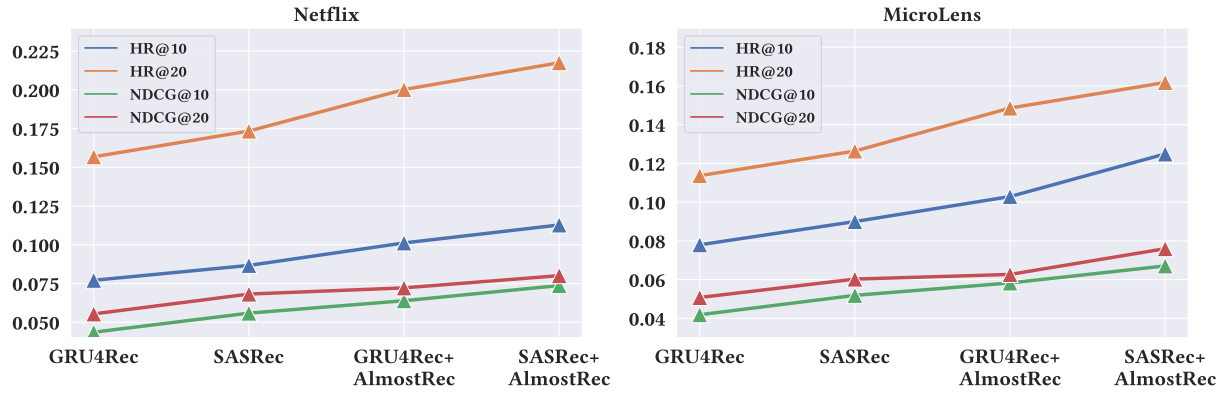


Figure 3: Ablation study of the pre-trained sequential recommender of AlmostRec

challenging to rationally align and harness the synergy of successful sequential recommendations. The failure may be due to the fact that these models simply inject pre-extracted video features into the ID embeddings without essential fusion and alignment. In contrast, AlmostRec merely uses raw images and item IDs yet achieves highly accurate and relevant recommendations based on the interpretation of multimodal user behaviors. This demonstrates that our designed Content-Behavior Guidance module not only incorporates rich multimodal information but also effectively aligns multiple modalities, enhancing the model's understanding of user sequential behavior with minimal modifications to the LMM parameters.

5.4 The Sequential Embeddings

To further investigate the effect of the pre-trained sequential embeddings of AlmostRec, we conduct ablation studies on the sequential recommenders, and the results are shown in Figure 3. It could be observed that

(1) with the incorporation of pre-trained sequential embeddings, AlmostRec consistently outperforms the original sequential recommenders, which validates the effectiveness of the proposed framework and the capability of LMMs in sequential recommendations.

Figure 4: The showcased qualitative evaluation between AlmostRec, SASRec, and LLaRA.

(2) While GRU4Rec performs worse than SASRec, GRU4Rec + AlmostRec noticeably outperforms SASRec over multiple domains and metrics, which implies that the sequential and collaborative signals in the pre-trained sequential embeddings serve as a significant ingredient for LMMs to perform sequential recommendations.
(3) Under the content-behavior guidance in AlmostRec, sequential embeddings of finer quality enables superior recommendation performance.

The results demonstrate that the incorporation of pre-trained sequential embeddings in AlmostRec significantly enhances the performance of sequential recommenders by leveraging both sequential and collaborative signals. This improvement underscores the effectiveness of the proposed framework and highlights the critical role of high-quality embeddings in enabling LMMs to excel in sequential recommendation tasks across diverse domains and metrics.

5.5 Qualitative Evaluation

In order to validate the practical application of AlmostRec, we conducted a qualitative evaluation of AlmostRec, SASRec, and LLaRA. Texts, images, and item IDs are provided as inputs, while the models' input and output capabilities vary. The results are visualized in Figure 4. In the task of predicting a user's next item of interest based on his historical viewing sequence, AlmostRec and SASRec accurately forecasted the results, while LLaRA struggled, likely due to its insufficient multimodal guidance and ID prediction capabilities. Furthermore, while SASRec and LLaRA were limited to producing single outputs, AlmostRec, through its optimized objectives and design of the transformation layer, is capable of simultaneously providing outputs for item ID, text, and visual image. This demonstrates that AlmostRec not only possesses enhanced multimodal understanding and user interest perception capabilities but also delivers recommendations in an intuitive, interactive, and user-centric manner.

6 Limitations

While AlmostRec has shown remarkable ability to model user preferences with the elaborated components, there is a few limitations remaining underexplored:

(1) The LMM backbone of AlmostRec is specifically chosen as LLaVA-1.5 [29] because of its simple yet potent property. The effect of different LMM backbones is to be explored, and more powerful LMMs could promisingly advance the performance of AlmostRec.
(2) Although there are some implications, the generalization ability of AlmostRec may not be sufficiently investigated. Experiments under cold-start and cross-domain settings may be helpful to verify the effectiveness and robustness of the LMM-based approach compared with conventional approaches.

7 Conclusion

In this paper, we propose AlmostRec to effectively align LMMs with sequential recommendations, addressing the challenges of interpreting user preferences and mitigating hallucinations based on the synergy of textual, visual, and behavioral sequence modalities. Experiments demonstrate that our proposed AlmostRec can effectively bring significant improvements to the sequential recommendation. Through AlmostRec, we offer a promising solution for developing simple yet intuitive, efficient, and user-centric multimodal recommendation systems, thus laying the groundwork for future research in this domain.

Acknowledgments

This work is supported by the National Key Research and Development Program of China under Grant No.2022ZD0115903, National Natural Science Foundation of China No.62222209, Beijing National Research Center for Information Science and Technology under Grant No.BNR2023TD03006.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.

[2] Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. TALLRec: An effective and efficient tuning framework to align large language model with recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1007–1014.

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238* (2023).

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[5] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*. 191–198.

[6] Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *arXiv preprint arXiv:2307.02046* (2023).

[7] Ziwei Fan, Zhiwei Liu, Jiawei Zhang, Yun Xiong, Lei Zheng, and Philip S Yu. 2021. Continuous-time sequential recommendation with temporal graph collaborative transformer. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 433–442.

[8] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524* (2023).

[9] Chendi Ge, Xin Wang, Ziwei Zhang, Yijian Qin, Hong Chen, Haiyang Wu, Yang Zhang, Yuekui Yang, and Wenwu Zhu. 2025. Behavior Importance-Aware Graph Neural Architecture Search for Cross-Domain Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 11708–11716.

[10] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*. 299–315.

[11] Shijie Geng, Juntao Tan, Shuchang Liu, Zuohui Fu, and Yongfeng Zhang. 2023. VIP5: Towards Multimodal Foundation Models for Recommendation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[12] Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. 2014. An Empirical Investigation of Catastrophic Forgeting in Gradient-Based Neural Networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

[13] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[14] Balázs Hidasi and Alexandros Karatzoglou. 2018. Recurrent neural networks with top-k gains for session-based recommendations. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 843–852.

[15] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[16] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. CogAgent: A Visual Language Model for GUI Agents. arXiv:2312.08914 [cs.CV]

[17] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[18] Wenyue Hua, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2023. How to index item ids for recommendation foundation models. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 195–204.

[19] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14271–14280.

[20] Yangqin Jiang, Yuhao Yang, Lianghao Xia, and Chao Huang. 2024. Diffkg: Knowledge graph diffusion model for recommendation. In *Proceedings of the 17th ACM international conference on web search and data mining*. 313–321.

[21] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 197–206.

[22] Walid Krichene and Steffen Rendle. 2020. On sampled metrics for item recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1748–1757.

[23] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020* (2023).

[24] Haoyang Li, Xin Wang, Ziwei Zhang, Jianxin Ma, Peng Cui, and Wenwu Zhu. 2021. Intention-aware sequential recommendation with structured intent transition. *IEEE Transactions on Knowledge and Data Engineering* 34, 11 (2021), 5403–5414.

[25] Xinhang Li, Chong Chen, Xiangyu Zhao, Yong Zhang, and Chunxiao Xing. 2023. E4SRec: An elegant effective efficient extensible solution of large language models for sequential recommendation. *arXiv preprint arXiv:2312.02443* (2023).

[26] Vladislav Lialin, Vijeta Deshpande, and Anna Rumshisky. 2023. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647* (2023).

[27] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. 2023. LLaRA: Aligning large language models with sequential recommenders. *arXiv preprint arXiv:2312.02445* (2023).

[28] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, et al. 2023. How can recommender systems benefit from large language models: A survey. *arXiv preprint arXiv:2306.05817* (2023).

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning.

[30] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., 34892–34916.

[32] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. GPT understands, too. *AI Open* (2023).

[33] Yuqing Liu, Yu Wang, Lichao Sun, and Philip S Yu. 2024. Rec-GPT4V: Multimodal Recommendation with Large Vision-Language Models. *arXiv preprint arXiv:2402.08670* (2024).

[34] Chaochao Lu, Chen Qian, Guodong Zheng, Hongxing Fan, Hongzhi Gao, Jie Zhang, Jing Shao, Jingyi Deng, Jinlan Fu, Kexin Huang, et al. 2024. From GPT-4 to Gemini and Beyond: Assessing the Landscape of MLLMs on Generalizability, Trustworthiness and Causality through Four Modalities. *arXiv preprint arXiv:2401.15071* (2024).

[35] Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, and Jiebo Luo. 2023. Llm-rec: Personalized recommendation via prompting large language models. *arXiv preprint arXiv:2307.15780* (2023).

[36] Jianxin Ma, Chang Zhou, Hongxia Yang, Peng Cui, Xin Wang, and Wenwu Zhu. 2020. Disentangled self-supervision in sequential recommenders. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 483–491.

[37] Yongxin Ni, Yu Cheng, Xiangyan Liu, Junchen Fu, Youhua Li, Xiangnan He, Yongfeng Zhang, and Fajie Yuan. 2023. A content-driven micro-video recommendation dataset at scale. *arXiv preprint arXiv:2309.15379* (2023).

[38] Gustavo Penha, Ali Vardasbi, Enrico Palumbo, Marco De Nadai, and Hugues Bouchard. 2024. Bridging Search and Recommendation in Generative Retrieval: Does One Task Help the Other?. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 340–349.

[39] Junyan Qiu, Haitao Wang, Zhaolin Hong, Yiping Yang, Qiang Liu, and Xingxing Wang. 2023. ControlRec: Bridging the semantic gap between language model and personalized recommendation. *arXiv preprint arXiv:2311.16441* (2023).

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[41] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 452–461.

[42] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.

[43] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 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.

[44] Jiaxi Tang and Ke Wang. 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.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[46] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, Shengxian Wan, and Xueqi Cheng. 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.

[47] Xin Wang, Hong Chen, Zirui Pan, Yuwei Zhou, Chaoyu Guan, Lifeng Sun, and Wenwu Zhu. 2025. Automated disentangled sequential recommendation with large language models. *ACM Transactions on Information Systems* 43, 2 (2025), 1–29.

[48] Xin Wang, Zirui Pan, Yuwei Zhou, Hong Chen, Chendi Ge, and Wenwu Zhu. 2023. Curriculum co-disentangled representation learning across multiple environments for social recommendation. In *International Conference on Machine Learning*. PMLR, 36174–36192.

[49] Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Llmrec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*. 806–815.

[50] Zihao Wu, Xin Wang, Hong Chen, Kaidong Li, Yi Han, Lifeng Sun, and Wenwu Zhu. 2023. Diff4Rec: Sequential Recommendation with Curriculum-scheduled Diffusion Augmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9329–9335.

[51] Zhengyi Yang, Jiancan Wu, Yanchen Luo, Jizhi Zhang, Yancheng Yuan, An Zhang, Xiang Wang, and Xiangnan He. 2023. Large language model can interpret latent space of sequential recommender. *arXiv preprint arXiv:2310.20487* (2023).

[52] Yuyang Ye, Zhi Zheng, Yishan Shen, Tianshu Wang, Hengruo Zhang, Peijun Zhu, Runlong Yu, Kai Zhang, and Hui Xiong. 2025. Harnessing multimodal large language models for multimodal sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 13069–13077.

[53] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M Jose, and Xiangnan He. 2019. A simple convolutional generative network for next item recommendation. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 582–590.

[54] Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, et al. 2024. Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations. *arXiv preprint arXiv:2402.17152* (2024).

[55] Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Yan Gao, Yao Hu, and Enhong Chen. 2024. NoteLLM-2: Multimodal Large Representation Models for Recommendation. *arXiv preprint arXiv:2405.16789* (2024).

[56] Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Recommendation as instruction following: A large language model empowered recommendation approach. *arXiv preprint arXiv:2305.07001* (2023).

[57] Zeyang Zhang, Xin Wang, Haibo Chen, Haoyang Li, and Wenwu Zhu. 2024. Disentangled Dynamic Graph Attention Network for Out-of-Distribution Sequential Recommendation. *ACM Transactions on Information Systems* 43, 1 (2024), 1–42.

[58] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 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.

[59] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=1tZbq88f27