Harnessing Multimodal Large Language Models for Multimodal Sequential Recommendation

Yuyang Ye[1], Zhi Zheng[2], Yishan Shen[3], Tianshu Wang[4], Hengruo Zhang[4], Peijun Zhu[5], Runlong Yu[6], Kai Zhang[2], Hui Xiong[7]*

[1]Department of Management Science and Information Systems, Rutgers University
[2]School of Data Science, University of Science and Technology of China
[3]Department of Applied Mathematics and Computational Science, University of Pennsylvania
[4]Bytedance Inc.
[5]School of Computer Science, Georgia Institute of Technology
[6]Department of Computer Science, University of Pittsburgh
[7]Thrust of Artificial Intelligence, The Hong Kong University of Science and Technology (Guangzhou)
yuyang.ye@rutgers.edu, zhengzhi97@mail.ustc.edu, yishans@sas.upenn.edu, {ustcwtsphy, hengruoz}@gmail.com, zpj@gatech.edu, ruy59@pitt.edu, kkzhang08@ustc.edu.cn, xionghui@ust.hk

Abstract

Recent advances in Large Language Models (LLMs) have demonstrated significant potential in the field of Recommendation Systems (RSs). Most existing studies have focused on converting user behavior logs into textual prompts and leveraging techniques such as prompt tuning to enable LLMs for recommendation tasks. Meanwhile, research interest has recently grown in multimodal recommendation systems that integrate data from images, text, and other sources using modality fusion techniques. This introduces new challenges to the existing LLM-based recommendation paradigm which relies solely on text modality information. Moreover, although Multimodal Large Language Models (MLLMs) capable of processing multi-modal inputs have emerged, how to equip MLLMs with multi-modal recommendation capabilities remains largely unexplored. To this end, in this paper, we propose the Multimodal Large Language Model-enhanced Sequential Multimodal Recommendation (MLLM-MSR) model. To capture the dynamic user preference, we design a two-stage user preference summarization method. Specifically, we first utilize an MLLM-based item-summarizer to extract image feature given an item and convert the image into text. Then, we employ a recurrent user preference summarization generation paradigm to capture the dynamic changes in user preferences based on an LLM-based user-summarizer. Finally, to enable the MLLM for multi-modal recommendation task, we propose to fine-tune a MLLM-based recommender using Supervised Fine-Tuning (SFT) techniques. Extensive evaluations across various datasets validate the effectiveness of MLLM-MSR, showcasing its superior ability to capture and adapt to the evolving dynamics of user preferences.

Code — https://github.com/YuyangYe/MLLM-MSR

Introduction

The development of Large Language Models (LLMs) has significantly enhanced the capacity for natural language understanding (Floridi and Chiriatti 2020; Achiam et al. 2023; Touvron et al. 2023), which has been instrumental in advancing recommendation systems (RSs). LLMs have demonstrated remarkable improvements in processing complex user preferences due to its strong semantic understanding and summarization ability. These attributes significantly enhance personalization and accuracy in recommendations(Wu et al. 2023; Zhang et al. 2024; Ren et al. 2024), particularly in Sequential Recommendations (SRs) where extracting long historical preferences is crucial (Hou et al. 2023; Zheng et al. 2024; Zhai et al. 2023; Li et al. 2023).

Simultaneously, beyond solely modeling textual information, there has been a growing interest in leveraging multimodal information (Liu et al. 2023; Chen et al. 2024). Techniques such as multimodal fusion and gated multimodal units have been utilized to integrate data from various sources—images, videos, and audio—enriching the context for recommendations. This offers a deeper understanding of user-item interactions and naturally leads to the exploration of Multimodal Large Language Models (MLLMs) for enhancing multimodal recommendation systems (Liu et al. 2021b; Zhou and Miao 2024; Kim et al. 2024). MLLMs merge multimodal information into a unified textual semantic space, enhancing the system's ability to understand complex data inputs, thereby can significantly improving recommendation accuracy (Liu et al. 2024d; Zhang et al. 2024). The application of MLLMs in sequential recommender systems presents a promising avenue for dynamically adapting to user preferences and handling the complexity of multimodal data, which holds considerable untapped potential.

However, integrating MLLMs into multimodal sequential recommendation systems introduces a set of notable challenges. First, the inherent complexity and computational demands of processing sequential multimodal data, particularly with multiple ordered image inputs, significantly constrain the scalability and efficiency of these systems (Yue et al. 2024; Koh, Fried, and Salakhutdinov 2024; Ye et al. 2024). Moreover, conventional MLLMs often exhibit limitations in comprehending the temporal dynamics of user in-

*Corresponding Author

teractions and preferences, particularly in the context of sequential multimodal interactions (Gao et al. 2021; Liu et al. 2024d). This critical limitation undermines the systems' capacity to accurately capture and reflect the evolving nature of user interests over time. Furthermore, fine-tuning MLLMs for specific recommendation scenarios while avoiding overfitting and preserving the generalizability gained during pretraining presents a significant challenge (Borisov et al. 2022; Yin et al. 2023; Li, Zhang, and Chen 2023). These hurdles underscore the need for innovative approaches that can navigate the complexities of multimodal sequential data, ensuring that MLLMs can be effectively leveraged to enhance recommendation systems.

To address these challenges, this paper introduces the Multimodal Large Language Model-enhanced Multimodal Sequential Recommendation (MLLM-MSR), a pioneering approach that leverages the capabilities of MLLMs to enhance and integrate multimodal item data effectively. Specifically, we introduce a Multimodal User Preferences Inference approach, which merges traditional multimodal fusion with sequence modeling techniques with MLLMs. Initially, we employ MLLMs to transform visual and textual data of each item into a cohesive textual description, preserving the integrity of the information as demonstrated by a preliminary study. Subsequently, utilizing the enriched item information processed through the MLLM, we develop an innovative LLM-based recurrent method to infer user preferences, capturing the temporal dynamics of these preferences. This method addresses the above mentioned challenges in processing sequential image inputs by harnessing superior text process capabilities of LLMs and improves the interpretability of recommendation compared with traditional representation based approaches, by providing detailed user preference. Further, we fine-tune an MLLM to function as a recommender, utilizing a carefully designed set of prompts that integrate this enriched item data, inferred user preferences, and the ground-truth of user-item interactions. This process of Supervised Fine-Tuning (SFT) on an open-source MLLM, equips the model with the ability to accurately match user preferences with potential items, thereby enhancing the personalization and accuracy of recommendations. To validate the effectiveness of MLLM-MSR, we conduct extensive experiments across three publicly available datasets from various domains, which confirm the superior performance of our approach. The major contributions of this paper are summarized as follows:

- To best of our knowledge, our work is the first attempt to fine-tune multimodal large models to address the challenges of sequential multimodal recommendation, where our fine-tune strategies achieving significant improvements in recommendation performance.
- We introduce a novel image summarizing method based on MLLMs to recurrently summarize user preferences on multi modality, facilitating a deeper understanding of user interactions and interests over time.
- Our approach is extensively validated across various datasets, demonstrating its effectiveness in enhancing the accuracy and interpretability of recommendations.

Related Work

Multimodal Sequential Recommendation

Sequential Recommenders (SRs) have evolved from matrix-based models to advanced neural architectures. FPMC combined matrix factorization with Markov chains for sequential modeling (Rendle et al. 2010; Yu et al. 2023). Neural models emerged with GRU4Rec, leveraging gated recurrent units for session-based recommendations (Hidasi et al. 2015). SASRec introduced self-attention for capturing long-term dependencies (Kang and McAuley 2018), while BERT4Rec employed transformers for bidirectional training, significantly improving performance (Sun et al. 2019).

Multimodal-enhanced SRs utilize additional contextual data to refine recommendations. Fusion methods are categorized into early, late, and hybrid approaches (Hu et al. 2023). Early fusion integrates modalities at the input level, enhancing feature representation via concatenation and gating (Tang and Wang 2018; Sun et al. 2019; Lei, Ji, and Li 2019), while non-invasive methods use attention mechanisms for attribute merging (Rendle et al. 2019; Liu et al. 2021a). Late fusion combines features from separate modules before the final stage (Zhang et al. 2019; Ji et al. 2020; Du et al. 2023). Hybrid fusion flexibly combines modality fusion and sequential modeling, leveraging inter-modality relationships for enhanced performance (Zhao, Lee, and Wu 2020; Hu et al. 2023).

LLM for Recommendation

Large Language Models (LLMs) have significantly impacted recommendation systems, building on foundational models like BERT (Devlin et al. 2018) and GPT-3 (Brown et al. 2020) to process vast textual data for deeper user behavior insights. BERT4Rec (Sun et al. 2019) and RLMRec (Ren et al. 2024) extend these capabilities to generate personalized, context-aware recommendations.

LLM applications in recommendation systems fall into embeddings-based, token-based, and direct model approaches (Wu et al. 2023; Cao et al. 2024; Guo et al. 2024). Embeddings-based methods utilize LLMs for feature extraction from user and item data (Cui et al. 2022; Liu et al. 2024c). Token-based approaches generate tokens encapsulating semantic meanings and user preferences, integrating them into recommendation logic (Zhai et al. 2023). Lastly, direct model applications use LLMs as end-to-end recommenders, generating recommendations directly from user queries and profiles (Hou et al. 2024; Geng et al. 2022). Additionally, MLLM-based frameworks extend capabilities to multimodal data, incorporating images, text, and video for improved system accuracy and user experience (Liu et al. 2024d; Zhang et al. 2024).

Preliminary

In this section, we will give the definition of our research problem and conduct a preliminary study to discuss the effectiveness of image summarizing approach.

Problem Definition

We first introduce the problem formulation of the Sequential Multimodal Recommendation problem. The dataset used in this work contains the interaction records between users and items. Given a user u, let us first define the historical user behavior sequence of u as $S_u = [I_u^1, \ldots, I_u^n]$, where I^i represents the i-th item with which the user has interacted, through actions such as clicking, purchasing, or watching, and n denotes the length of the user behavior sequence. In addition, each item corresponds to a textual description $\mathcal{W}$ and a image $\mathcal{I}$ (e.g., product diagram, video cover). Consequently, our problem can be formulated as follows.

Definition 1 (Sequential Multimodal Recommendation) *Given a user u with the corresponding historical behavior sequence S^u, including both textual and visual data, and a candidate item I_c, the objective of the Sequential Multimodal Recommendation is to predict the probability of next interacted item I_u^{n+1} (for example, the probability of clicking) with the candidate item I_u^c for the user u, denoted as $g_u : I_c \to \mathbb{R}$.*

Effectiveness of Multiple Images Summary

As highlighted in the Introduction, current multimodal large language models (MLLMs) face challenges in processing multiple image inputs, limiting their effectiveness for sequential multimodal analysis. To overcome this problem, we introduce an image summary approach that leverages MLLMs to convert and summarize image content. The efficacy of this technique is evaluated using the basic sequential recommender, GRU4Rec, on real-world datasets (detailed in the Experiment section). In our approach, we employed simple prompts like "Please summarize the image" with LLaVA (Liu et al. 2024b,a) to generate image summaries. These summaries were transformed into latent vectors using BERT (Devlin et al. 2018), which then fed into the GRU4Rec model. This method is benchmarked against direct image representations from VGG19 (Simonyan 2014), assessing performance via the AUC metric.

Dataset	Microlens	Baby	Games
Image Summary	0.7281	0.7318	0.7451
VGG19 Features	0.7154	0.7383	0.7532

Table 1: Performance of GRU4Rec with Different Inputs

The performance is detailed in Table 1. Results show that using image summaries allows the GRU4Rec model to perform comparably to direct processing with VGG19, confirming that our image summary approach preserves necessary semantic information in sequential modeling. This preliminary validation underscores the effectiveness of our method in addressing the challenges associated with processing multiple ordered images.

Technical Details

This section will introduce the technical details of our proposed MLLM-MSR framework, which contains two main components as Multimodal User Preferences Inference and Tuning MLLM based Recommender, illustrated as Figure 1.

Multimodal User Preferences Inference

In the context of sequential recommendation, a common approach is to learn user representations and predict future interactions with candidate items via calculating affinity scores. Unlike traditional methods that utilize embeddings, LLMs typically analyze user preferences and interaction probabilities directly at token level. This section will detail how our method employs Multimodal Large Language Models (MLLMs) to specifically address challenges associated with multimodal recommendation scenarios.

Multimodal Item Summarization To effectively predict user preferences, it is crucial to analyze historical item sequences. In multimodal recommendation scenarios, handling multiple image inputs presents a significant challenge for MLLMs, especially in maintaining the sequence of these inputs and aligning textual information with corresponding images. To overcome these issues, we propose a Multimodal Item Summarization approach, which simplifies the processing by summarizing multimodal information of images into unified textual descriptions by designing effective prompts to integrate the multimodal data of items.

Our prompt design adheres to foundational methods of multimodal information fusion. Item information can be separated into textual descriptions and image. Hence, In the initial phase, distinct prompts (i.e., text summarization and image description prompt) are used to guide MLLMs to process these modalities independently, to ensure a more thorough comprehension and detailed feature extraction from each modality, ensuring nuanced characteristics often missed in unified analyses are captured. To ensure both modalities contribute equally to item modeling, the outputs of text summarization and image description are calibrated to similar lengths.

After independently analyzing each modality, our design integrates insights from both textual and visual information using a fusion prompt. This approach aligns with traditional multimodal recommendation strategies that emphasize synthesizing diverse data types to create a comprehensive item profile, enhances the multifaceted understanding of the item (Baltrušaitis, Ahuja, and Morency 2018; Huang, Xiao, and Yu 2019; Gao et al. 2020).

Recurrent User Preference Inference In the Sequential Multimodal Recommendation framework, achieving detailed personalization relies on an accurate understanding of user preferences. The advent of Multimodal Large Language Models (MLLMs) marks a significant advancement in understanding multimodal information. However, as we introduced above, they are struggle in dealing with sequential multimodal data. Although our multimodal item summarization method effectively integrates multimodal information into a unified item summary, this complexity still leads to unstable and random outputs when the historical sequence becomes long, leading to excessively long prompts. Consequently, this results in suboptimal performance in sequential recommendation systems.

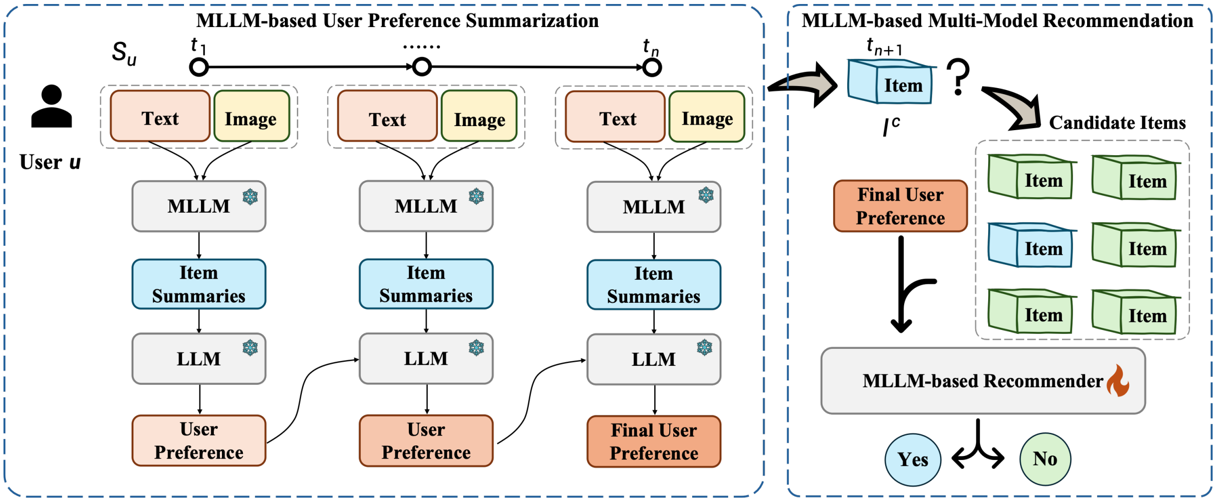

Figure 1: The schematic framework of MLLM-MSR

To address these challenges, our method, inspired by Recurrent Neural Networks (RNNs), employs prompted sequence modeling to iteratively capture user preferences through interaction sequences. In RNNs, each output is influenced by both the current input and the previous state, facilitating contextual awareness across sequences. We segment item interactions into several blocks, each covering interactions within a defined session, converting long multimodal sequences into concise textual narratives that sequentially represent the user's historical interactions. This segmentation enables our approach to dynamically represent user preferences, effectively overcoming the limitations of MLLMs in processing sequential and multimodal data. By incorporating prompt-driven modules in each session, our method integrates insights from previous interactions to refine the understanding of current user preferences continuously. This iterative process is essential for accurately capturing the dynamic of user preferences and offers more interpretable descriptions than traditional representation-based models, enhancing the potential for detailed case studies.

User:
Below is a chronological summary list of videos recently watched by User:
- In this engaging episode, titled "The Cook's Special Request: Inviting the Owner's Mother to Dinner," the video cover sets the scene with a lively cooking demonstration…
- In this captivating episode of the "Gourmet Fun Stomach Project," …

[User's Previous Preference]
The user has preference for content related to food, cooking, and cuisine, particularly Chinese cuisine…
Based on the content and visual style of videos listed above as well as the known user's previous preference, please summarize the user's preferences.

Assistant:
The user has preference for food-related content, particularly Chinese cuisine, and enjoys watching cooking tutorials, recipes, and food-related stories. They also seem to be interested in the behind-the-scenes aspect of food preparation, such as featuring the owner's wife or boss mother in the kitchen. Additionally, the user may recently appreciate humor and entertainment in their food content, suggesting a desire for engaging and lighthearted content.

Figure 2: An example of recurrent user preference inference.

Specifically, as we have initially generated item multimodal summaries for each item, we pair these summaries with prompts that guide the LLMs to infer user preferences based on a sequential narrative. For example, the initial prompt in the first block is designed to summarize the user's initial interests from a chronological list of item interactions at the first timestamp. Subsequently, at following session, the prompt for updating the summarized preference is showed as Figure 2. Our approach uses prompted sequence modeling to iteratively understand user preferences through detailed analysis of each interaction, thus effectively manage challenges with long, multimodal sequences.

Tuning MLLM based Recommender

After gathering user preferences using the methods described above, we can propose a supervised fine-tuning of an open-sourced MLLM, such as LLaVA[1], which excels at understanding images. This model would be used to build a multimodal recommender system through SFT. In line with the definition of sequential recommendation, given a user-item interaction, the MLLM-based recommender system utilizes the prompt contained the obtained user preferences, the textual description and image of the given item and the designed system instruction prompt to predict the probability that the user will interact with the candidate item. Specifically, the prompt designed for the tuned MLLM recommender module is illustrated in Figure 3, where we restrict the output to only include 'yes' or 'no' to avoid irrelevant information about the predicted label. Thus, the probability of item interaction can be calculated from the probability score of the predicted first new token as follows:

$$p = \frac{p(\text{'yes'})}{p(\text{'yes'}) + p(\text{'no'})} \tag{1}$$

To construct a multimodal sequential recommender system based on MLLMs, we implement supervised fine-tuning

[1]https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

USER:

<image> As a vision-llm, you will be given the cover image and the title of a video and the summarized preference of a user, and your task is to predict whether the user would interact with the video. Please only response 'yes' or 'no' based on your judgment.

Based on the previous interaction history, the user's preference can be summarized as: "The user appears to have a preference for content related to anime, manga, and video games, particularly Japanese and Chinese series, with a focus on action, adventure, and fantasy genres, as well as a visual style characterized by vibrant colors, detailed character designs…"

Please predict whether this user would interact with the video at the next opportunity. The video's title is "after 20 years, the squirrels of the Ice Age finally ate acorns, the end of an era! # Glacierage # Photosynthetic Creators Conference 2022 # Animated Movie Push Save", and the given image is this video's cover.

Assistant: YES

Figure 3: An example of MLLM-based sequential recommendation

to optimize the model parameters. This fine-tuning process involves adjusting the model to minimize the discrepancy between predicted and actual user interactions. Our dataset construction strategy employs negative sampling, a commonly used training technique in recommendation systems, wherein each positive user-item interaction is coupled with multiple negative samples representing items with which the user did not interact (Yu et al. 2018). This methodology aids the model in distinguishing between relevant and irrelevant items through contrastive learning, thereby improving its predictive accuracy.

The model is trained on a dataset comprising sequences of user-item interactions, with each interaction encapsulated as a sequence of user preferences, item descriptions, and images. The fine-tuning leverages the next token prediction paradigm, training the model to predict the subsequent token in a sequence based on preceding tokens. This ensures the generation of coherent and contextually pertinent outputs from the input sequences. The supervised fine-tuning loss function is defined as:

$$L = -\sum_{i=1}^{L} \log P(v_i | v_{<i}, \mathcal{I}), \tag{2}$$

where v_i represents the i-th token of the prompt text, L denotes the prompt length and $\mathcal{I}$ is the given image. The probability $P(v_i|v_{<i}, \mathcal{I})$ is calculated using MLLMs within the next token prediction framework, which maximizes the likelihood of the ground truth tokens given the prompt. This ensures the model learns to accurately predict the subsequent token based on the provided context, which is critical for generating precise and contextually aware recommendations. Specifically, we employ LoRA (Hu et al. 2021) to adhere to a parameter-efficient fine-tuning framework (PEFT), which accelerates the training process.

Experiments

In this section, we detail the comprehensive experiment to validate the effectiveness of our proposed Multimodal Large Language Model for Sequential Multimodal Recommendation (MLLM-MSR).

Dataset	Microlens	Amazon-Baby	Amazon-Game
#User	25411	41081	38808
#Item	20276	14393	13379
#Interaction	223263	400876	352136
#Avg Seqlen	11.35	13.65	13.23
Sparsity	99.96%	99.93%	99.93%

Table 2: The Statistics of Datasets

Experimental Setup

Dataset Description Our experimental evaluation utilized three open-source, real-world datasets from diverse recommendation system domains. These datasets include the *Microlens Dataset* (Ni et al. 2023), featuring user-item interactions, video introductions, and video cover images; the *Amazon-Baby Dataset*; and the *Amazon-Game Dataset* (He and McAuley 2016; McAuley et al. 2015), all of them contain user-item interactions, product descriptions, and images. These selections enabled a thorough analysis across different recommendation systems. We preprocessed each dataset by removing infrequent users and items to ensure user history sequences met our minimum length criteria. Additionally, we implemented a 1:1 ratio for negative sampling during training and a 1:20 ratio for evaluation. Further details on these datasets are provided in Table 2.

Baseline Methods To evaluate the effectiveness of our proposed LMM-MSR method, we selected some compared methods which can be categorized into following groups:

- **Basic SR Models**: These models use item attributes including IDs and textual information. We selectively integrate the most effective information from these attributes to achieve optimal performance. *GRU4Rec (Hidasi et al. 2015)*: Utilizes Gated Recurrent Units (GRU) to model the sequential dependencies between items. *SASRec (Kang and McAuley 2018)*: Employs a self-attention mechanism to capture long-term dependencies.
- **Multimodal recommendation model**: *MMGCN (Wei et al. 2019)*: Integrates multimodal features into a graph-based framework using a message-passing scheme.

	Microlens			Video-Games			Amazon-Baby		
	AUC	HR@5	MRR@5	AUC	HR@5	MRR@5	AUC	HR@5	MRR@5
GRU4Rec	72.55	53.32	33.36	73.45	59.32	34.56	73.68	58.21	35.81
SASRec	74.02	58.88	36.57	71.06	43.81	23.94	80.50	70.09	52.34
MGAT	71.23	42.37	30.84	72.21	48.71	32.17	73.19	46.42	31.92
MMGCN	73.35	54.47	34.14	74.91	55.63	35.19	74.88	57.13	36.13
GRU4Rec_F	75.31	60.28	33.25	75.79	62.38	36.39	76.41	63.84	35.77
SASRec_F	77.69	63.03	37.42	77.51	66.64	43.55	81.94	72.08	58.32
MMSR	78.87	69.85	55.21	79.01	71.95	56.52	81.19	70.34	56.38
Trans2D	70.23	51.91	31.21	72.20	56.70	39.28	66.60	28.94	45.91
LLaVa	52.75	28.21	17.33	53.57	33.24	18.73	55.86	34.82	20.96
TALLREC	81.25	71.23	58.22	81.79	72.36	57.38	82.16	74.31	60.15
MLLM-MSR	**83.17**	**78.23**	**60.38**	**84.39**	**77.42**	**63.25**	**85.69**	**79.58**	**63.77**

Table 3: The performance of different methods.

MGAT (Monti, Bronstein, and Bresson 2017): Employs a graph attention network to disentangle personal interests by modality.

- **Multimodal feature enhanced SR models**: *$GRU4Rec_F$*, *$SASRec_F$*: Adaptations of GRU4Rec and SASRec with multimodal feature enhancements. *Trans2D (Zhao, Lee, and Wu 2020)*: Utilizes holistic fusion to integrate features across different dimensions. *MMSR (Hu et al. 2023)*: Depolys a graph-based approach for adaptive fusion of multi-modal features, which dynamically adjusts the fusion order of modalities based on their sequential relationships.
- **LLM based SR models**: *TALLREC (Bao et al. 2023)*: Uses LLMs for sequence recommendation through SFT, exclusively processing textual inputs. *LLaVA w/o SFT*: Utilizes LLaVA as the recommender without a specific fine-tuning for recommendation.

Metrics To assess the performance of baseline methods and our proposed MLLM-MSR for multimodal sequential recommendations, we employed AUC, HR@5, and MRR@5 as evaluation metrics (Yu et al. 2020). To ensure a fair comparison, we standardized the size of the candidate item sets across all baseline methods and our approach.

Implementation Details Our experiments were performed on a Linux server equipped with eight A800 80GB GPUs. We utilized Llava-v1.6-mistral-7b for image description and recommendation tasks, and Llama3-8b-instruct [2] for summarizing user preferences. For the Supervised Fine-Tuning (SFT) process, we employed the PyTorch Lightning library, using LoRA with a rank of 8. The optimization was handled by the AdamW optimizer with a learning rate of 2e-5 and a batch size of 1, setting gradient accumulation steps at 8 and epochs at 10. For distributed training, we implemented Deepspeed [28] with ZeRO stage 2. Additionally, we set the maximum token length for MLLMs at 512 and the number of items per block in recurrent preference inference at 3.

[2]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Performance Analysis

The performance of the compared methods and our MLLM-MSR is presented in Table 3, where all results were obtained using 5-fold cross-validation and various random seeds, and achieved a 95% confidence level. It is evident that, in our evaluation, MLLM-MSR consistently outperforms all other metrics in terms of both classification and ranking, underscoring the personalization accuracy of our recommendation system. We also observed additional insights: Firstly, compared to basic sequential recommendation (SR) models, our adaptations that incorporate multimodal inputs, particularly SASRec, show significantly better results. This emphasizes the critical role of multimodal integration within the sequential recommendation framework and confirms the efficacy of self-attention-based models in handling both multi-modality and sequential inputs. Moreover, MMSR distinguishes itself from other non-LLM baselines, highlighting the importance of integrating multimodal fusion modules with sequential modeling components in SR tasks, thereby indirectly supporting our prompt design idea for user preference inference. Conversely, purely multimodal recommendation models such as MMGCN and MGAT exhibit lower performance due to their lack of dedicated sequential modeling components. This indicates that for optimal effectiveness in SR, the integration of both multimodal and sequential processing capabilities is essential. Lastly, within the realm of large language model (LLM)-based SR models, our approach significantly outperforms LLaVA without specific fine-tuning. This success validates the effectiveness of our strategically designed prompts for SR tasks. Additionally, our method outperforms TALLREC, demonstrating our success in integrating multimodal information and unlocking the potential of large multimodal models compared to other LLM-based approaches using only textual information. This comparative advantage underscores the integration of advanced MLLM training techniques and the strategic application of multimodal data processing in enhancing sequential recommendation systems.

Ablation Study

To evaluate the individual contributions of certain components within our MLLM-MSR framework, we developed

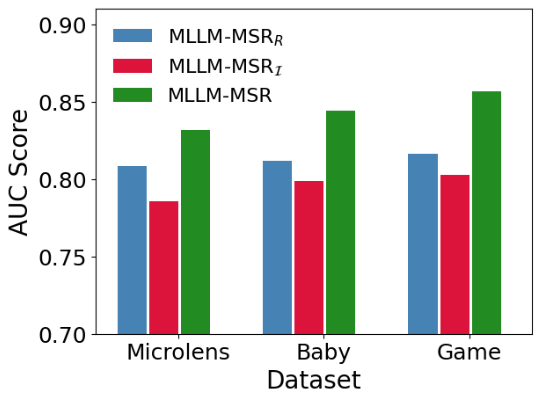

Figure 4: The performance of MLLM-MSR and its variants.

several variants of MLLM-MSR, each described as follows:

- **MLLM-MSR$_\mathrm{R}$**: This variant employs direct user preference inference instead of the recurrent method. It uses the entire chronological order of historical interaction data to infer user preferences.
- **MLLM-MSR$_\mathcal{I}$**: In this version, we omit the image summarization component in item summarization. It relies solely on textual data for user preference inference while still incorporating image information in the recommender module, leveraging capabilities that are readily achievable with current MLLMs.

As Figure 4 shows, in the evaluations across three diverse datasets, our primary model, MLLM-MSR, consistently outperformed its variants, MLLM-MSR$_\mathrm{R}$ and MLLM-MSR$_\mathcal{I}$, demonstrating the essential roles of its key components. The MLLM-MSR$_\mathrm{R}$ variant, which employed direct user preference inference, achieved suboptimal performance. This result validates the importance of our model's recurrent method in capturing the dynamic evolution of user preferences, indicates our methods can reflect current interests more accurately and reduce the negative impact of lengthy prompts. Besides, the worse performance of MLLM-MSR$_\mathcal{I}$ variant, which excluded image summaries and depended solely on textual data for user preference inference, illustrated the significance of integrating multimodal data. This integration is crucial to understand user preferences across different modalities, thereby significantly compensating for the incompleteness of textual information.

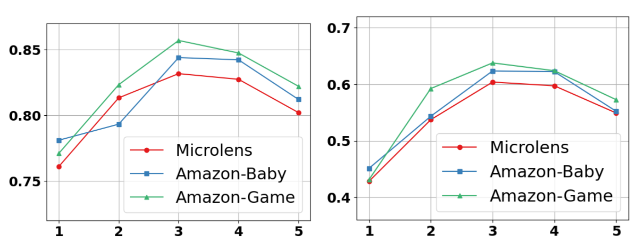

Figure 5: Performance of MLLM-MSR under different block size. (Left: AUC, Right: MRR@5)

Parameter Analysis

In this section, we first analyze the optimal block size for the recurrent user preference inference component of our MLLM-MSR model. As Figure 5 shows, striking the right balance on the block size is crucial; Too small a block size simplifies the approach to direct inference, potentially missing the dynamic evolution of user preferences due to the limited contextual span. Conversely, too large a block size leads to long prompts, increasing computational load and reducing the number of blocks available to effectively capture temporal dynamics, thereby diminishing the system's adaptive capabilities. Optimal block sizing ensures the model processes sequential data efficiently and adapts dynamically to changes in user behavior.

Additionally, we evaluate the impact of context length on the predictive performance of our model. By fixing the output length during user preference generation, we assess how different context lengths affect recommendation outcomes. The results are shown in Figure 6. We found short context lengths cause a loss of information, resulting in suboptimal predictions. However, once the context length reaches a certain threshold, the results stabilize, indicating that the large model has strong summarizing capabilities and can capture all necessary information within a specific optimal range. This demonstrates the importance of selecting a proper context length to maximize information utility without incurring unnecessary computational complexity.

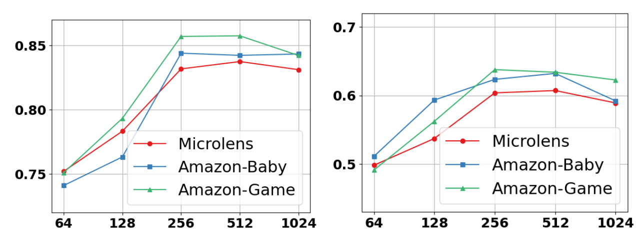

Figure 6: Performance of MLLM-MSR under different user preference summary length. (Left: AUC, Right: MRR@5)

Conclusion

In this study, our proposed model, the Multimodal Large Language Model-enhanced Multimodal Sequential Recommendation (MLLM-MSR), effectively leverages MLLMs for multimodal sequential recommendation. Through a novel two-stage user preference summarization process and the implementation of SFT techniques, MLLM-MSR showcases a robust ability to adapt to and predict dynamic user preferences across various datasets. Our experimental results validate the outstanding performance of MLLM-MSR compared to existing methods, particularly in its adaptability to evolving preferences. This paper introduces a innovative use of MLLMs that enriches the recommendation process by integrating diverse modalities and enhances the personalization and accuracy of the recommendations, and meanwhile providing added interpretability through detailed user preference analysis.

Acknowledgments

This research was partially supported by the National Science Foundation (NSF) via the grant number IIS-2006387 and IIS-2040799.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443.

Bao, K.; Zhang, J.; Zhang, Y.; Wang, W.; Feng, F.; and He, X. 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.

Borisov, V.; Seßler, K.; Leemann, T.; Pawelczyk, M.; and Kasneci, G. 2022. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901.

Cao, Y.; Mehta, N.; Yi, X.; Keshavan, R.; Heldt, L.; Hong, L.; Chi, E. H.; and Sathiamoorthy, M. 2024. Aligning Large Language Models with Recommendation Knowledge. *arXiv preprint arXiv:2404.00245*.

Chen, L.; Sun, Y.; Zhang, S.; Ye, Y.; Wu, W.; and Xiong, H. 2024. Tackling uncertain correspondences for multi-modal entity alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Cui, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. M6-rec: Generative pretrained language models are open-ended recommender systems. *arXiv preprint arXiv:2205.08084*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Du, X.; Yuan, H.; Zhao, P.; Qu, J.; Zhuang, F.; Liu, G.; Liu, Y.; and Sheng, V. S. 2023. Frequency enhanced hybrid attention network for sequential recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 78–88.

Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694.

Gao, J.; Li, P.; Chen, Z.; and Zhang, J. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5): 829–864.

Gao, P.; Zareian, A.; Maji, S.; Darrell, T.; and Rohrbach, M. 2021. CLIP: Connecting Text and Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1–10.

Geng, S.; Liu, S.; Fu, Z.; Ge, Y.; and Zhang, Y. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt, and Predict Paradigm (P5). *Proceedings of the 16th ACM Conference on Recommender Systems*.

Guo, W.; Wang, H.; Zhang, L.; Chin, J. Y.; Liu, Z.; Cheng, K.; Pan, Q.; Lee, Y. Q.; Xue, W.; Shen, T.; et al. 2024. Scaling New Frontiers: Insights into Large Recommendation Models. *arXiv preprint arXiv:2412.00714*.

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.; Karatzoglou, A.; Baltrunas, L.; and Tikk, D. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

Hou, Y.; He, Z.; McAuley, J.; and Zhao, W. X. 2023. Learning vector-quantized item representation for transferable sequential recommenders. In *Proceedings of the ACM Web Conference 2023*, 1162–1171.

Hou, Y.; Zhang, J.; Lin, Z.; Lu, H.; Xie, R.; McAuley, J.; and Zhao, W. X. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, 364–381. Springer.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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.

Huang, Z.; Xiao, W.; and Yu, Y. 2019. A Survey on Deep Learning Based Recommender Systems: From Traditional Methods to Recent Advances in Multimedia Recommendations. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2401–2409. ACM.

Ji, W.; Sun, Y.; Chen, T.; and Wang, X. 2020. Two-stage sequential recommendation via bidirectional attentive behavior embedding and long/short-term integration. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, 449–457. IEEE.

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

Kim, Y.; Kim, T.; Shin, W.-Y.; and Kim, S.-W. 2024. MONET: Modality-Embracing Graph Convolutional Network and Target-Aware Attention for Multimedia Recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 332–340.

Koh, J. Y.; Fried, D.; and Salakhutdinov, R. R. 2024. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36.

Lei, C.; Ji, S.; and Li, Z. 2019. Tissa: A time slice self-attention approach for modeling sequential user behaviors. In *The World Wide Web Conference*, 2964–2970.

Li, L.; Zhang, Y.; and Chen, L. 2023. Prompt distillation for efficient llm-based recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 1348–1357.

Li, X.; Chen, C.; Zhao, X.; Zhang, Y.; and Xing, C. 2023. E4srec: An elegant effective efficient extensible solution of large language models for sequential recommendation. *arXiv preprint arXiv:2312.02443*.

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

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, Q.; Chen, N.; Sakai, T.; and Wu, X.-M. 2024c. Once: Boosting content-based recommendation with both open-and closed-source large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 452–461.

Liu, Q.; Hu, J.; Xiao, Y.; Gao, J.; and Zhao, X. 2023. Multimodal recommender systems: A survey. *arXiv preprint arXiv:2302.03883*.

Liu, Y.; Wang, Y.; Sun, L.; and Yu, P. S. 2024d. Rec-GPT4V: Multimodal Recommendation with Large Vision-Language Models. *arXiv preprint arXiv:2402.08670*.

Liu, Y.; Yang, S.; Lei, C.; Wang, G.; Tang, H.; Zhang, J.; Sun, A.; and Miao, C. 2021b. Pre-training graph transformer with multimodal side information for recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2853–2861.

McAuley, J.; Targett, C.; Shi, Q.; and Van Den Hengel, A. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 43–52.

Monti, F.; Bronstein, M.; and Bresson, X. 2017. Geometric Matrix Completion with Recurrent Multi-Graph Neural Networks. In *Advances in Neural Information Processing Systems*.

Ni, Y.; Cheng, Y.; Liu, X.; Fu, J.; Li, Y.; He, X.; Zhang, Y.; and Yuan, F. 2023. A Content-Driven Micro-Video Recommendation Dataset at Scale. *arXiv preprint arXiv:2309.15379*.

Ren, X.; Wei, W.; Xia, L.; Su, L.; Cheng, S.; Wang, J.; Yin, D.; and Huang, C. 2024. Representation learning with large language models for recommendation. In *Proceedings of the ACM on Web Conference 2024*, 3464–3475.

Rendle, S.; Krichene, W.; Zhang, L.; and Anderson, J. 2019. DIF-SR: Differential Sequential Recommendation. In *Conference on Recommender Systems*.

Rendle, S.; et al. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, 811–820.

Simonyan, K. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1441–1450.

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.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wei, Y.; Zhao, X.; Liu, G.; Zhu, Z.; Zhuang, Y.; Qin, J.; and Wang, J. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1437–1445. ACM.

Wu, L.; Zheng, Z.; Qiu, Z.; Wang, H.; Shen, T.; Qin, C.; Zhu, C.; Zhu, H.; Liu, Q.; et al. 2023. A Survey on Large Language Models for Recommendation. *arXiv preprint arXiv:2305.19860*.

Ye, Y.; Tang, L.-A.; Wang, H.; Yu, R.; Yu, W.; He, E.; Chen, H.; and Xiong, H. 2024. PAIL: Performance based Adversarial Imitation Learning Engine for Carbon Neutral Optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6148–6157.

Yin, B.; Xie, J.; Qin, Y.; Ding, Z.; Feng, Z.; Li, X.; and Lin, W. 2023. Heterogeneous knowledge fusion: A novel approach for personalized recommendation via llm. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 599–601.

Yu, R.; Liu, Q.; Ye, Y.; Cheng, M.; Chen, E.; and Ma, J. 2020. Collaborative list-and-pairwise filtering from implicit feedback. *IEEE Transactions on Knowledge and Data Engineering*, 34(6): 2667–2680.

Yu, R.; Xu, X.; Ye, Y.; Liu, Q.; and Chen, E. 2023. Cognitive Evolutionary Search to Select Feature Interactions for Click-Through Rate Prediction. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3151–3161.

Yu, R.; Zhang, Y.; Ye, Y.; Wu, L.; Wang, C.; Liu, Q.; and Chen, E. 2018. Multiple pairwise ranking with implicit feedback. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 1727–1730.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567.

Zhai, J.; Zheng, X.; Wang, C.-D.; Li, H.; and Tian, Y. 2023. Knowledge prompt-tuning for sequential recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6451–6461.

Zhang, C.; Zhang, H.; Wu, S.; Wu, D.; Xu, T.; Gao, Y.; Hu, Y.; and Chen, E. 2024. NoteLLM-2: Multimodal Large Representation Models for Recommendation. *arXiv preprint arXiv:2405.16789*.

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

Zhao, Z.; Lee, D.; and Wu, X. 2020. Trans2D: A Holistic Fusion Model for Sequential Recommendation. *arXiv preprint arXiv:2006.04563*.

Zheng, Z.; Chao, W.; Qiu, Z.; Zhu, H.; and Xiong, H. 2024. Harnessing Large Language Models for Text-Rich Sequential Recommendation. *arXiv preprint arXiv:2403.13325*.

Zhou, X.; and Miao, C. 2024. Disentangled Graph Variational Auto-Encoder for Multimodal Recommendation With Interpretability. *IEEE Transactions on Multimedia*.