VRAgent-R1: Boosting Video Recommendation with MLLM-based Agents via Reinforcement Learning

Siran Chen[2,3,4*], Boyu Chen[2,3,4*], Yuxiao Luo [2*], Chenyun Yu[1†], Yi Ouyang [4], Lei Cheng[4], Chengxiang Zhuo [4], Zang Li [4], Yali Wang[2, 5†]

[1] Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
[2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[3] University of Chinese Academy of Science, Beijing, China
[4] Tencent, Shenzhen, China
[5] Shanghai Artificial Intelligence Laboratory, Shanghai, China
chensiran17, chenboyu18@mails.ucas.ac.cn, yuchy35@mail.sysu.edu.cn, yl.wang@siat.ac.cn

Abstract

Large language model (LLM) agents have emerged as a promising solution for enhancing recommendation systems via user simulation. However, existing studies predominantly resort to prompt-based simulation using frozen LLMs, which frequently results in suboptimal item modeling and user preference learning, thereby ultimately constraining recommendation performance. To address these challenges, we introduce VRAgent-R1, a novel agent-based paradigm that incorporates human-like intelligence in user simulation. Specifically, VRAgent-R1 comprises two distinct agents: the Item Perception (IP) Agent and the User Simulation (US) Agent, designed for interactive user-item modeling. Firstly, the IP Agent emulates human-like progressive thinking based on MLLMs, effectively capturing hidden recommendation semantics in videos. With a more comprehensive multimodal content understanding provided by the IP Agent, the video recommendation system is equipped to provide higher-quality candidate items. Subsequently, the US Agent refines the recommended video sets based on in-depth chain-of-thought (CoT) reasoning and achieves better alignment with real user preferences through reinforcement learning. Experimental results on a large-scale video recommendation benchmark MicroLens-100k have demonstrated the effectiveness of our proposed VRAgent-R1 method, e.g., the IP Agent achieves a 6.0% improvement in NDCG@10, while the US Agent shows approximately 45.0% higher accuracy in user decision simulation compared to state-of-the-art baselines.

Introduction

With the booming popularity of short video platforms like TikTok and Kuaishou, Multimodal Recommendation Systems (MRS) have attracted considerable attention from both academia and industry. Benefiting from the significant advancements in Multimodal Large Language Models (MLLMs), recent studies tend to utilize them to boost MRS. On one hand, some approaches (Zhang et al. 2024b; Chen et al. 2024b; Luo et al. 2024; Ren et al. 2024; Lee et al. 2024; Jia et al. 2025; Song et al. 2024) leverage the pre-trained MLLMs to directly convert each item's multimodal information into a single embedding. However, fine-tuning MLLMs to achieve semantic alignment for recommendation requires abundant high-quality interaction data and computational resources. On the other hand, some methods (Zhang et al. 2025; Ye et al. 2025; Bao et al. 2023; Zhang et al. 2024c) directly apply LLMs or MLLMs to recommendation tasks, transforming the recommendation into a language generation problem. But these methods also face challenges such as limited input length, computational inefficiency, and hallucinations, making them unsuitable for large-scale recommendations. Recently, growing attention has been paid to using LLM-based agents to enhance recommendation systems' personalization and intelligence (e.g., user profiling, simulating interactions, improving satisfaction). However, existing methods (Zhang et al. 2024a, 2025; Xiang et al. 2024) mostly use frozen LLMs, with the knowledge gap between vertical domains and LLMs/MLLMs limiting their adaptability and effectiveness.

Based on the above discussion, we propose to use LLMs/MLLMs to enhance multimodal content comprehension and simulate user decision-making, which could help human-centric recommendation outcomes, and we are required to address the following challenges. **1) How to discriminately exploit the recommendation-relevant semantics hidden in video items?** Most video MRS heavily rely on understanding video items, which presents two major difficulties. Firstly, previous methods ignore the temporal or casual relation within video frames and fail to identify key information among numerous contents. For example, some methods (He et al. 2016; Ni et al. 2023) directly use randomly selected, pre-processed visual features, while MLLM-MSR (Ye et al. 2025) relies solely on the first frame, underutilizing video dynamics. Secondly, previous methods modalities separately with late fusion, which can lead to modality competition (Shang et al. 2023; Liu et al. 2024), even leading to inferior performance compared to using a single modality (Zhou et al. 2023). The lack of in-depth semantic interaction between raw modalities may also cause a misunderstanding of the high-level semantics, e.g., political topics in Fig 3, expressed in the video. **2) How to**

[*] Equal contribution.
[†] Corresponding authors

Figure 1: **Qualitative examples of VRAgent-R1 for Video Recommendation.** Our VRAgent-R1 stands out from previous supervised fine-tuning of MLLM, which fails to give the correct prediction due to the lack of understanding of video items and deep thinking on user status.

effectively simulate user behavior that mirrors human deep thinking? LLMs show great promise for user simulation in the recommendation systems due to their excellent linguistic understanding capabilities. However, existing LLMs struggle with processing sequential multimodal inputs, which prevents direct modeling of multimodal user behavior sequences. Moreover, most user simulators (Zhang et al. 2024a, 2025, 2024c; Xiang et al. 2024) primarily rely on prompt engineering to instruct frozen LLMs to generate responses without feedback, potentially leading to discrepancies between the agent and real user behavior. Some methods (Bao et al. 2023; Ye et al. 2025) attempt to fine-tune the large model, but simple fine-tuning only yields binary 'Yes' or 'No' judgments without deep analysis of the user's status, limiting the model's generalizability as shown in Fig. 1.

To address the aforementioned challenges, we introduce VRAgent-R1, a novel agent-based paradigm for video recommendation. Unlike prior approaches built on frozen LLMs, VRAgent-R1 effectively mimics the human-like thinking for recommendation by leveraging multimodal collaborative understanding and reinforcement fine-tuning (RFT) on user simulation. Specifically, it consists of two distinct agents: the Item Perception (IP) Agent and the User Simulation (US) Agent. As illustrated in Fig. 2, The IP Agent aims to establish comprehensive multimodal content understanding for videos to improve item modeling through multi-round, in-depth semantic interaction with the MLLM. This process enables it to progressively discover key video content and effectively extract recommendation-relevant semantics from the items. Subsequently, the semantic summarization of videos generated by the IP Agent is used to optimize the fundamental video recommendation model via feature augmentation. This not only enhances the recommendation model but also facilitates the US Agent to capture user preferences and predict the next item based on historical interactions. The US Agent focuses on deep user behavior simulation and provides proxy feedback to refine the candidate set provided by the video recommendation system. To align user simulation with real decision-making, reinforcement learning is leveraged to enable the model to analyze historical user behavior (watched videos and comments) and comprehensively summarize user status with Chain-of-Thought (CoT) reasoning. Additionally, we design a reward mechanism associated with the user's actual final behavior and update the fundamental LLM through policy optimization, i.e., GRPO. Through the collaboration of IP and US agents, VRAgent-R1 effectively boosts video recommendation performance with step-by-step human-like thinking. The main contributions of this work can be summarized as follows:

- We propose VRAgent-R1, a novel agent-based framework designed to assist video recommendations from a user-centric perspective, which exhibits human-like intelligence for interpretable recommendations. By incorporating a user-like understanding, this framework significantly enhances the performance of recommendation system, demonstrating the effectiveness of the pipeline.
- Our IP Agent achieves more comprehensive multimodal understanding of videos by flexibly conducting in-depth semantic interactions between textual and visual contents. Furthermore, our US Agent is the first to use RFT for LLM-based user simulation, achieving more accurate simulation performance through deep thinking on user status with little training data.
- Extensive experiments demonstrate that the IP Agent significantly enhances the performance of existing video recommendation approaches, achieving a 4.3% improvement in HR@10 and a 6.0% improvement in NDCG@10 on the MicroLens-100k dataset. Meanwhile, the US Agent outperforms commercial models such as GPT-4o and SFT methods, and the simulated user feedback can further boost the recommendation accuracy by reranking

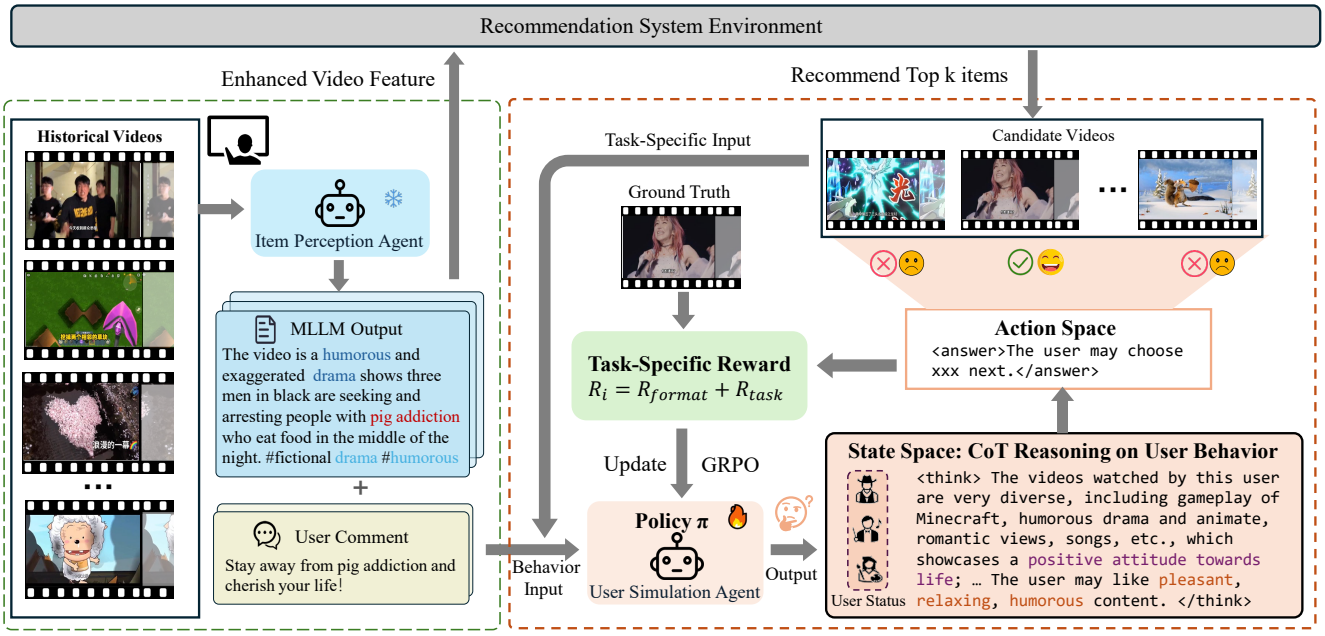

Figure 2: **Overview of our VRAgent-R1 framework.** We propose a framework with two novel agents for better video recommendation. The IP Agent conducts collaborative multimodal understanding to obtain enhanced video features for the recommendation system and the US Agent. Meanwhile, the US Agent simulates user behavior via deep CoT reasoning based on user status. VRAgent-R1 achieves superior simulation performance and helps improve the recommendation accuracy.

the candidate item set generated from the recommendation system.

Related Work

LLMs for Multi-Modal Recommendation. LLMs and MLLMs have performed a profound impact on their integration into current recommendation systems. Existing methods utilizing LLMs can be broadly categorized into implicit and explicit applications. The implicit methods (Sun et al. 2019; Ren et al. 2024; Lee et al. 2024; Jia et al. 2025; Song et al. 2024; Zhang et al. 2024b; Chen et al. 2025, 2024a) directly utilize the pre-trained structure or parameters of large models to convert user and item information into embeddings. For example, NoteLLM-2 (Zhang et al. 2024b) employs an MLLM with end-to-end fine-tuning to fuse multimodal information as the item embedding. Explicit methods (Zhang et al. 2025, 2024c; Hou et al. 2024; Ye et al. 2025; Bao et al. 2023) involve using the reasoning ability of MLLMs to expand item information and analyze user profiles or intentions, ultimately generating textual summarization to aid recommendations. For example, using LLMs as user simulator (Zhang et al. 2024a,c) to predict user behavior. However, research on video recommendation with minute-level visual content is relatively scarce compared to text-based and image-based recommendations, and we are pioneers in using MLLMs for video recommendation.

Training LLMs/MLLMs with RL. With the success of DeepSeek-R1 (Guo et al. 2025), reinforcement learning (RL) has demonstrated its remarkable ability to enhance the logical reasoning capabilities of LLMs with high data efficiency. There have been explorations to improve LLMs' performance in reasoning tasks, such as solving mathematical puzzles (Shao et al. 2024; Yang et al. 2024b; Ying et al. 2024) and coding (Zhang et al. 2024e,d). Furthermore, Visual-RFT (Liu et al. 2025) pioneers the enhancement of reasoning and visual perception in Large Vision Language Models with limited data. In the recommendation scenario, compared with SFT, the RL method requires less data to learn a reasoning strategy with good generalization, making it suitable for cold start scenarios and user simulation. To our knowledge, we are among the first to apply RFT of LLMs for user simulation and video recommendation.

Method

As shown in Fig. 2, our VRAgent-R1 framework mainly consists of two components: the Item Perception Agent (IP Agent) for video modeling and the User Simulation Agent (US Agent) for user modeling.

Item Perception Agent (IP Agent)

Existing video representation learning methods typically process visual and textual information separately by inputting them into distinct encoders for later feature fusion. However, due to the heterogeneity and the imbalance in information volumes between the two modalities, it is prone to modality competition, which in turn leads to a suboptimal semantic space for item representations (Zhou et al. 2023). To accurately localize and extract the high-level semantics of video for more precise recommendations, we propose a progressive approach that utilizes MLLMs to gradually mine

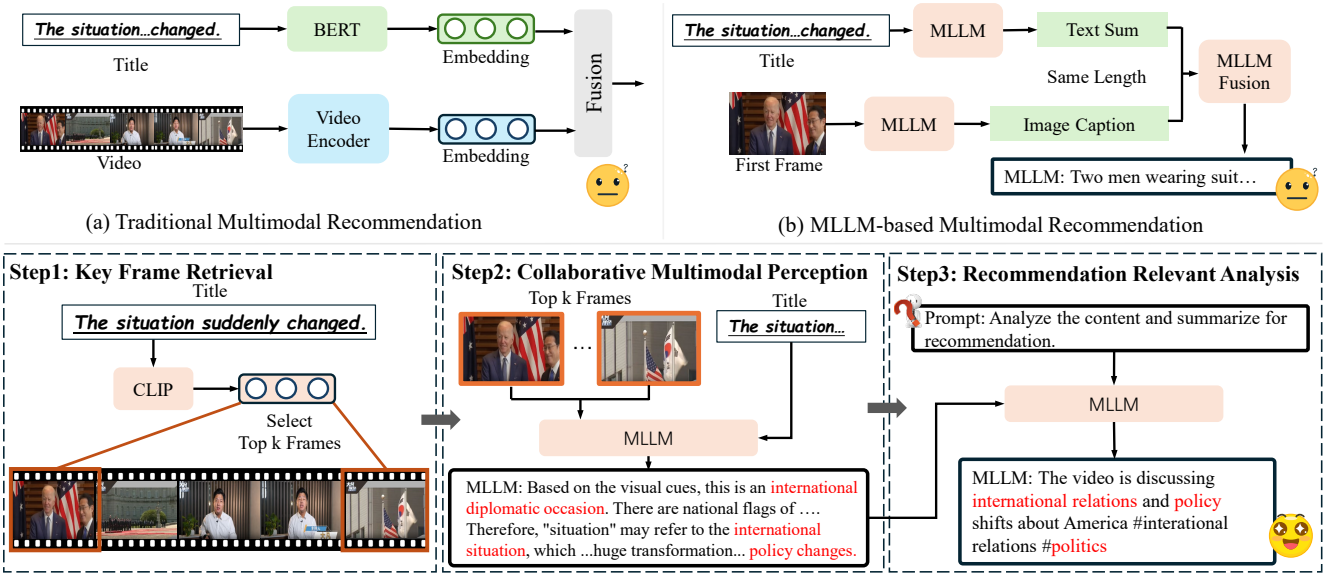

Figure 3: **Video understanding by the IP Agent.** We simulate the human video comprehension process through a progressive approach involving retrieval, collaborative perception, and analysis, so as to obtain a summary of the key video information that is applicable for recommendation.

video information through key frame retrieval, collaborative multimodal perception, and recommendation-relevant analysis, as illustrated in Fig. 3.

Key Frame Retrieval (KFR). Given that videos contain a wealth of visual information, directly utilizing all video frames would introduce substantial redundant information and result in low computational efficiency. To identify the most crucial information in the visual representation while ensuring the algorithm's efficiency, we uniformly sample frames from the video. Subsequently, we employ CLIP (Radford et al. 2021) to compute the visual-text similarity scores between these sampled frames and the video title. The frames with the top 5 highest CLIP scores are then identified as the key visual information for the video's representation.

Collaborative Multimodal Perception (CMP). After obtaining the retrieved frames, the next step is to identify the specific events and high-level semantics conveyed in the video. Our approach stands out from previous methods by fully leveraging the multi-modal understanding capabilities of the MLLM. Specifically, since some titles do not directly reflect the video's topics, we input both the retrieved frames and titles into the MLLM and prompt it to understand the semantic context implied by the titles. During this process, MLLM can provide relevant explanations of the title and offer supplementary information. For instance, in Fig. 3, the term "situation" might pertain to international relations, while "change" could imply shifts in national policy. It is important to note that neither modality's embedding could independently capture such nuanced semantics. Thus, the MLLM can now clearly comprehend the video information, including the main characters, general events, video genre, and the sentiment expressed in the video.

Recommendation Relevant Analysis (RRA). The captions initially generated by MLLM may not be well-suited for the specific recommendation scenario, as they may contain excessive redundant explanations or even hallucinations. To address this issue, we prompt the model to analyze the detailed video content jointly with the characteristics of scenario, as well as focusing on the key information that users are most likely to find interesting. The model then reformulates the video content into a concise and precise caption limited to approximately 35 words, which is close to the average length of the original titles. This process helps filter out unimportant details, resulting in a unified and comprehensive video caption. Note that the reformulated video caption not only can be used to enhance the representation learning for items, but also assist the process of user behavior simulation.

User Simulation Agent (US Agent)

After modeling the video items, we then consider simulating human behavior to refine the recommendation results. Existing LLM-based user agents (Zhang et al. 2025, 2024a; Wang et al. 2025; Xiang et al. 2024) simply prompt frozen LLMs, therefore LLMs can not be optimized and the simulation outcomes may be unrealistic and prone to hallucinations. Moreover, simple supervised fine-tuning only enables models to memorize answers. Due to the lack of in-depth analysis of user behavior, this approach yields limited accuracy and lacks interpretability. To better align the model with the user decision-making process, we innovatively employ reinforcement learning to fine-tune the LLM within a simulated recommendation environment.

Environment and User Modeling. First, we identify the modeling of the recommendation environment and person-

alized user as shown in Fig. 2. The environment aims to simulate a realistic exposure-click recommendation scenario by generating candidate videos for the user, while the US Agent simulates the user to perform specific tasks. More specifically, for a user with N behaviors (including watched videos and corresponding comments), the former N-1 behaviors are used for user profile modeling, and the N-th behavior is regarded as the prediction target. We use SASRec (Ni et al. 2023) as the basis RS to recommend 10 video items based on the user's historical behaviors, simulating a rough recall process. Then *m* items are randomly selected as negative samples, and the real N-th item of user behavior is regarded as the positive one. These *m*+1 items collectively form the candidate videos exposed to the user for future tasks, which we will discuss in the following section. However, processing multiple video and text sequences simultaneously is challenging for the MLLM. To address this issue, the IP Agent converts the relevant multimodal videos into a textual format, enabling the US Agent to process the long text sequence. For a given task in the RFT process, we prompt the US Agent to first thoroughly analyze the user behavior to formulate a unique user status s (e.g., preferences and emotions) through CoT reasoning. Compared to previous methods, the user profile modeling here is dynamically updated based on task rewards, which allows for a learnable and more accurate simulation.

Task and Reward. To align the US Agent with real user preferences, we design two specific tasks for RFT, i.e., User Preference Judgment and Next Video Selection. In the first task, following settings of previous methods (Ye et al. 2025; Zhang et al. 2024a, 2025), the agent is given an item from the candidate list and then prompted to judge whether the user would like the recommended video. The action space $\mathcal{A}$ consists of "Yes" and "No", corresponding to positive items and negative items, respectively. The Reward R_1 for this task comprises two parts: the format reward R_{format} and the judgment reward R_{jud},

$$R_1 = R_{\text{format}} + R_{\text{jud}}. \tag{1}$$

The format reward ensures the model adheres to the required response format, i.e., *<think>the CoT thinking process</think>, <answer>the final answer</answer>*. Additionally, we use a post-processing function f to parse the answer within the *<answer>* tag into a legal action, and check if it matches the ground truth. Here, R_{jud} will be 1 for a correct simulation and -1 for a wrong situation.

In the second task, the agent first reviews all candidate items and selects the video that the user is most likely to watch next, to simulate the exposure-click behavior . The action space $\mathcal{A}$ consists of choosing one item from the *m*+1 candidate videos. The reward R_2 for this task includes the format reward R_{format} and the selection reward R_{sel},

$$R_2 = R_{\text{format}} + R_{\text{sel}}. \tag{2}$$

Given the larger action space of R_2 compared to R_1, this task is more complex, and we assign a score of 2 for correctly selecting the positive video to provide a higher reward.

GRPO Training. We employ Group Relative Policy Optimization (GRPO) (Shao et al. 2024) framework to train the agent, which compares groups of candidate responses directly, without requiring a critic model to evaluate policy performance. Given a problem q for the model π_θ, it samples to generate a group of distinct answers o_i, where $i = 1, 2, \ldots, G$ and G is the sampled number in the group. Each answer involves different CoT reasoning for the user status and final answer, and we compute the corresponding reward r_i. By comparing the relative advantage of the *i*-th answer $\hat{A}_i$,

$$\hat{A}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})} \tag{3}$$

$\mathbf{r} = \{r_1, r_2, \ldots, r_G\}$, GRPO encourages the model to select the answer with higher reward within the group. We initially train the agent with an easy judgment task, then introduce the selection task. Via such a progressive training manner, the agent learns from simpler to more complex tasks, and the CoT process is gradually optimized, which provides thoughtful and interpretable recommendations for the user behavior.

Experiments

Datasets and Metrics. We have conducted extensive experiments on MicroLens-100K (Ni et al. 2023), an open-source, real-world video recommendation dataset that includes 100,000 users, 19,738 items, and 719,405 interactions, with a sparsity level of 99.96%. Notably, MicroLens is the first micro-video recommendation dataset to provide original video content, which enables us to analyze videos from raw frames. The average video duration is 161 seconds, featuring rich visual content. We also conduct user simulation on MovieLens-1M (Harper and Konstan 2015) and Steam (Kang and McAuley 2018). For traditional recommendation metrics across the entire dataset, we report standard recommendation metrics such as HR@10, NDCG@10, HR@20, and NDCG@20. For the evaluation of user behavior simulation, we report binary classification metrics such as accuracy (Acc) and F1 Score for preference judgment, as well as selection accuracy for next video selection.

Implementation Details. We employ Qwen2.5VL-7b for IP Agent and Qwen2.5-7b for US Agent. For video recommendation, we follow the official code of MicroLens-100K (Ni et al. 2023) benchmark for evaluation. In user simulation, we designate the actual last item of user behavior as the positive sample and employ SASRec (Ni et al. 2023) to generate the top 10 recommended items, from which *m* items are randomly chosen as negative samples to more accurately mimic real-world recommendation scenarios. We deploy 4 80G GPUs for the reinforcement fine-tuning of the US Agent using information from 2000 users. The training configuration includes a batch size of 16, 16 sampled policy rollouts, and a KL coefficient of 0.001. We set the maximum number of input and response tokens to 2048 and train the model for 4 epochs, which takes approximately 10 hours. User simulation evaluation is conducted on 1000 randomly selected cold-start users and repeated three times to report the average performance.

Class	Model	HR@10	NDCG@10	HR@20	NDCG@20
IDRec(CF)	DSSM (Huang et al. 2013)	0.0394	0.0193	0.0654	0.0258
	LightGCN (He et al. 2020)	0.0372	0.0177	0.0618	0.0239
	DeepFM (Guo et al. 2017)	0.0350	0.0170	0.0571	0.0225
IDRec(SR)	NextItNet (Yuan et al. 2019)	0.0805	0.0442	0.1175	0.0535
	GRU4Rec (Hidasi et al. 2015)	0.0782	0.0432	0.1147	0.0515
	SASRec (Kang and McAuley 2018)	0.0909	<u>0.0517</u>	0.1278	0.0610
Modality	NextItNet$_V$ (Ni et al. 2023)	0.0862	0.0466	0.1246	0.0562
	MMGCN (Wei et al. 2019)	0.0778	0.0423	0.1138	0.0513
	SASRec$_T$ (Ni et al. 2023)	0.0916	0.0490	0.1343	0.0598
	SASRec$_I$ (Ni et al. 2023)	0.0942	0.0511	0.1358	0.0613
	SASRec$_V$ (Ni et al. 2023)	0.0948	0.0515	<u>0.1364</u>	0.0619
	SASRec$_F$ (Ni et al. 2023)	<u>0.0953</u>	<u>0.0517</u>	0.1362	<u>0.0623</u>
MLLM Enhanced	SASRec+MLLM-MSR (Ye et al. 2025)	0.0606	0.0351	0.0911	0.0446
	NextItNet (Yuan et al. 2019)+Ours	0.0884	0.0478	0.1278	0.0583
	SASRec (Kang and McAuley 2018)+Ours	**0.0994**$_{+4.30\%}$	**0.0548**$_{+6.00\%}$	**0.1418**$_{+3.96\%}$	**0.0655**$_{+5.14\%}$

Table 1: **Comparison results on MicroLens-100K.** The fusion of different modality features is achieved by weighted pooling. The underline *T*, *I*, *V*, *F* correspond to text, image, video, and fusion features, respectively. Our MLLM-enhanced features successfully boost the performance of different traditional sequence recommendation methods.

Method	Acc	Recall	Pre	F1	Acc$_{m=3}$	Acc$_{m=4}$
GPT-4o (Hurst et al. 2024)	0.535	0.480	0.533	0.505	0.307	0.269
DeepSeek-R1 (Guo et al. 2025)	0.528	0.556	0.526	0.541	0.264	0.231
Qwen2.5-7b (Yang et al. 2024a)	0.491	0.512	0.490	0.500	0.245	0.197
LLM_Simulator (Zhang et al. 2025)	0.523	0.539	0.519	0.529	-	-
Agent4Rec (Zhang et al. 2024a)	0.528	0.482	0.52	0.501	-	-
TALLREC (SFT) (Bao et al. 2023)	0.537	**0.913**	0.521	0.663	-	-
MLLM-MSR (SFT) (Ye et al. 2025)	0.585	0.882	0.553	0.679	0.442	0.381
VRAgent-R1	**0.715**	0.760	**0.697**	**0.727**	**0.641**	**0.602**

Table 2: **Evaluation on User Simulation.** Our RFT method outperforms previous prompt and SFT-based simulation, with less training data and higher accuracy. *m* is the number of negative samples, and SFT methods have higher Recall scores since they tend to give positive answers.

Comparison for Video Recommendation

In this section, we evaluate the effectiveness of our IP Agent on the video recommendation baseline, following the experimental protocols in MicroLens (Ni et al. 2023). As shown in Table 1, sequential recommendation models that incorporate multimodal information outperform both Collaborative Filtering (CF) models and simple ID embedding-based models. However, traditional multimodal methods rely on late fusion techniques (e.g., addition or concatenation) of ID, textual, and visual embeddings. Due to the significant differences between modalities, such coarse-grained fusion fails to fully leverage multimodal information. Moreover, the lack of in-depth understanding of video frames means that these fusion methods only achieve similar performance to single-modality approaches. We also experiment with directly using the MLLM to generate a detailed caption for the cover image, as in MLLM-MSR (Ye et al. 2025). However, these captions often focus on unimportant details in the image, failing to capture key information and leading to a significant decline in performance. In contrast, our IP Agent integrates the pre-trained world knowledge of the MLLM to collaboratively understand multimodal content. It summarizes key information in a brief and unified format, which helps to boost HR@10 and NDCG@10 by 4.3% and 6.0%, respectively, compared to previous state-of-the-art methods.

Evaluation on User Simulation

In this subsection, we assess the performance of VRAgent-R1 and other user simulators in accurately modeling user behavior through two tasks: user preference prediction (estimating whether a user will like recommended items) and next video selection (choosing the most likely video a user will click next from candidates). Tab. 2 presents the results comparing our method with commercial models (GPT-4o, etc.) and several user simulation baseline methods. Directly prompting frozen LLMs based on textual inputs and then predict the answer yields poor results. This is because these LLMs rely solely on limited information from video titles and suffer from serious hallucination problems, leading to performance only slightly better than random guessing. For the SFT methods, e.g., MLLM-MSR, training the models to predict user behavior based on summarized user preferences shows improved performance, after fine-tuning with additional information (20k users). MLLM-MSR performs bet-

Method	MovieLens		Steam	
	Acc	F1	Acc	F1
GPT-4o (Hurst et al. 2024)	0.584	0.600	0.634	0.662
RecAgent (Wang et al. 2025)	0.581	0.621	0.627	0.650
Agent4Rec (Zhang et al. 2024a)	0.691	0.698	0.689	0.679
Qwen2.5 (SFT)	0.743	0.732	0.738	0.745
VRAgent-R1	**0.815**	**0.808**	**0.803**	**0.805**

Table 3: Simulation evaluation on other domains.

Method	**All**		**Cold**	
	HR@10	NDCG@10	HR@10	NDCG@10
Original	0.0916	0.0490	0.0586	0.0278
+IP Agent	0.0994	0.0548	0.0663	0.0318
+US dislike	0.0988	0.0543	0.0654	0.0311
+US like	**0.1003**	**0.0554**	**0.0678**	**0.0330**

Table 4: Optimizing RSs with VRAgent-R1.

Method	HR@10	NDCG@10	HR@20	NDCG@20
Baseline	0.0916	0.0490	0.1343	0.0598
w/o KFR	0.0980	0.0542	0.1402	0.0646
w/o CMP	0.0960	0.0519	0.1398	0.0629
w/o RRA	0.0816	0.0466	0.1182	0.0523
VRAgent-R1	**0.0994**	**0.0548**	**0.1418**	**0.0655**

Table 5: Ablation on the IP Agent.

Method	Acc	F1	$\text{Acc}_{m=3}$	$\text{Acc}_{m=4}$
Baseline	0.491	0.500	0.245	0.197
SFT	0.585	0.679	0.442	0.381
w/o CoT Reasoning	0.580	0.592	0.324	0.263
w/o Comment	0.695	0.705	0.618	0.577
w/o IP Agent	0.680	0.686	0.609	0.569
VRAgent-R1	**0.715**	**0.727**	**0.646**	**0.602**

Table 6: Ablation on the US Agent.

ter due to the usage of extra cover image information, but the results are still unsatisfactory, and the final model can only provide binary answers ("Yes" or "No") without a reasoning process. On the contrary, our VRAgent-R1 achieves significantly higher prediction accuracy, e.g., about 45% higher for the next video selection than MLLM-MSR, despite using less training data. Additionally, our model can be applied to different reasoning tasks without losing generality.

To verify the effectiveness of our method across different domains, we conduct user simulation tests following (Zhang et al. 2024a) on the widely used **MovieLens-1M** and **Steam** datasets. And we apply US Agent to predict users' performance toward items, since these datasets only involve textual information. The results in Tab. 3 indicate that in text-dominated movie and game recommendation scenarios, our approach also significantly outperforms previous simulation agents (Wang et al. 2025; Zhang et al. 2024a).

Optimizing RSs with VRAgent-R1 We conduct a preliminary experiment to assess whether VRAgent-R1's simulation could enhance recommendation systems (RSs). Specifically, we randomly select 8,000 cold-start users and provide VRAgent-R1 with the top 10 items recommended by the original RSs. VRAgent-R1 simulates user decisions on which videos they might watch and which they would dislike. This simulated behaviors are then used to supplement the modeling of cold-start users and update the RSs. As shown in Tab. 4, incorporating feedback on user-liked videos improves recommendation performance. In contrast, simulated interactions with user-disliked videos have a negative impact. This outcome effectively demonstrates the feedback-driven recommendation augmentation process.

Ablation on the IP Agent. We ablate the progressive steps and frame number used in our IP Agent. Specifically, "w/o KFR" denotes the absence of the key frame retrieval process, where we replace it with several randomly selected adjacent frames. "w/o CMP" indicates the removal of collaborative multimodal perception, the MLLM is employed to analyze visual frames and titles independently. "w/o RRA" signifies the exclusion of recommendation-relevant analysis, directly utilizing the long textual outputs for video comprehension. Tab. 5 shows both KFR and CMP significantly enhance video modeling. The analysis process is also crucial, as excessive unimportant details can be noisy and degrade recommendation performance.

Ablation on the US Agent. We then conduct ablation studies to assess the impact of each component of VRAgent-R1 on simulation performance, as detailed in Tab. 6. First, we remove the CoT reasoning from the RFT process, and then the LLM directly predicts answers based on the user's historical sequence, bypassing analysis of video contents and user status. This leads to a significant performance drop, underscoring the importance of CoT in reasoning tasks. Next, we evaluate the impact of user comments which aid in more accurate user modeling. Results show that the absence of them causes roughly 4% performance decline. Finally, we ablate the IP Agent, which is responsible for multimodal understanding. We can observe that the IP Agent boosts performance by approximately 6% over the baseline relying on original textual information.

Conclusion

In this paper, we propose a novel VRAgent-R1 framework for user simulation in video recommendation. It first utilizes an MLLM to collaboratively understand the retrieved multimodal content with pre-trained world knowledge, then analyzes the history sequence to establish user status and make final decision through RFT. By exploring different strategies and using real user decisions as verifiable rewards under different tasks, our VRAgent-R1 method achieves significant improvements in user behavior simulation for video recommendation. It outperforms SFT with minimal data and shows strong generalization. As a pioneering work, this study demonstrates the potential of applying RFT to LLMs in recommendation systems.

Acknowledgements

This work was supported by the National Key R&D Program of China(NO.2022ZD0160505). This research was supported by Shenzhen Science and Technology Program, China (Grant No. 202206193000001, 20220817180954005).

References

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.

Chen, B.; Chen, S.; Li, K.; Xu, Q.; Qiao, Y.; and Wang, Y. 2024a. Percept, Chat, and then Adapt: Multimodal Knowledge Transfer of Foundation Models for Open-World Video Recognition. *arXiv preprint arXiv:2402.18951*.

Chen, B.; Chen, S.; Li, K.; Xu, Q.; Qiao, Y.; and Wang, Y. 2025. Super Encoding Network: Recursive Association of Multi-Modal Encoders for Video Understanding. *arXiv preprint arXiv:2506.07576*.

Chen, J.; Chi, L.; Peng, B.; and Yuan, Z. 2024b. Hllm: Enhancing sequential recommendations via hierarchical large language models for item and user modeling. *arXiv preprint arXiv:2409.12740*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Guo, H.; Tang, R.; Ye, Y.; Li, Z.; and He, X. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247*.

Harper, F. M.; and Konstan, J. A. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4): 1–19.

He, R.; Fang, C.; Wang, Z.; and McAuley, J. 2016. Vista: A visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th ACM conference on recommender systems*, 309–316.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 639–648.

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

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.

Huang, P.-S.; He, X.; Gao, J.; Deng, L.; Acero, A.; and Heck, L. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2333–2338.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jia, J.; Wang, Y.; Li, Y.; Chen, H.; Bai, X.; Liu, Z.; Liang, J.; Chen, Q.; Li, H.; Jiang, P.; et al. 2025. LEARN: Knowledge Adaptation from Large Language Model to Recommendation for Practical Industrial Application. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11861–11869.

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

Lee, D.-H.; Kraft, A.; Jin, L.; Mehta, N.; Xu, T.; Hong, L.; Chi, E. H.; and Yi, X. 2024. STAR: A Simple Training-free Approach for Recommendations using Large Language Models. *arXiv preprint arXiv:2410.16458*.

Liu, Q.; Hu, J.; Xiao, Y.; Zhao, X.; Gao, J.; Wang, W.; Li, Q.; and Tang, J. 2024. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 57(2): 1–17.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

Luo, X.; Cao, J.; Sun, T.; Yu, J.; Huang, R.; Yuan, W.; Lin, H.; Zheng, Y.; Wang, S.; Hu, Q.; et al. 2024. QARM: Quantitative Alignment Multi-Modal Recommendation at Kuaishou. *arXiv preprint arXiv:2411.11739*.

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*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

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 Web Conference 2024*, 3464–3475.

Shang, Y.; Gao, C.; Chen, J.; Jin, D.; Ma, H.; and Li, Y. 2023. Enhancing adversarial robustness of multi-modal recommendation via modality balancing. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6274–6282.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Song, C.; Shen, C.; Gu, H.; Wu, Y.; Yi, L.; Wen, J.; and Chen, C. 2024. PRECISE: Pre-training Sequential Recommenders with Collaborative and Semantic Information. *arXiv preprint arXiv:2412.06308*.

Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; and Jiang, P. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Pro-*

ceedings of the 28th ACM international conference on information and knowledge management, 1441–1450.

Wang, L.; Zhang, J.; Yang, H.; Chen, Z.-Y.; Tang, J.; Zhang, Z.; Chen, X.; Lin, Y.; Sun, H.; Song, R.; et al. 2025. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2): 1–37.

Wei, Y.; Wang, X.; Nie, L.; He, X.; Hong, R.; and Chua, T.-S. 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.

Xiang, W.; Zhu, H.; Lou, S.; Chen, X.; Pan, Z.; Jin, Y.; Chen, S.; and Sun, L. 2024. SimUser: Generating Usability Feedback by Simulating Various Users Interacting with Mobile Applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–17.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yang, A.; Zhang, B.; Hui, B.; Gao, B.; Yu, B.; Li, C.; Liu, D.; Tu, J.; Zhou, J.; Lin, J.; et al. 2024b. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Ye, Y.; Zheng, Z.; Shen, Y.; Wang, T.; Zhang, H.; Zhu, P.; Yu, R.; Zhang, K.; and Xiong, H. 2025. Harnessing multimodal large language models for multimodal sequential recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 13069–13077.

Ying, H.; Zhang, S.; Li, L.; Zhou, Z.; Shao, Y.; Fei, Z.; Ma, Y.; Hong, J.; Liu, K.; Wang, Z.; et al. 2024. Internlm-math: Open math large language models toward verifiable reasoning. *arXiv preprint arXiv:2402.06332*.

Yuan, F.; Karatzoglou, A.; Arapakis, I.; Jose, J. M.; and He, X. 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.

Zhang, A.; Chen, Y.; Sheng, L.; Wang, X.; and Chua, T.-S. 2024a. On generative agents in recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*, 1807–1817.

Zhang, C.; Zhang, H.; Wu, S.; Wu, D.; Xu, T.; Zhao, X.; Gao, Y.; Hu, Y.; and Chen, E. 2024b. Notellm-2: Multimodal large representation models for recommendation. *arXiv preprint arXiv:2405.16789*.

Zhang, E.; Wang, X.; Gong, P.; Lin, Y.; and Mao, J. 2024c. Usimagent: Large language models for simulating search users. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2687–2692.

Zhang, K.; Li, G.; Dong, Y.; Xu, J.; Zhang, J.; Su, J.; Liu, Y.; and Jin, Z. 2024d. Codedpo: Aligning code models with self generated and verified source code. *arXiv preprint arXiv:2410.05605*.

Zhang, Y.; Wu, S.; Yang, Y.; Shu, J.; Xiao, J.; Kong, C.; and Sang, J. 2024e. o1-coder: an o1 replication for coding. *arXiv preprint arXiv:2412.00154*.

Zhang, Z.; Liu, S.; Liu, Z.; Zhong, R.; Cai, Q.; Zhao, X.; Zhang, C.; Liu, Q.; and Jiang, P. 2025. Llm-powered user simulator for recommender system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 13339–13347.

Zhou, H.; Zhou, X.; Zeng, Z.; Zhang, L.; and Shen, Z. 2023. A comprehensive survey on multimodal recommender systems: Taxonomy, evaluation, and future directions. *arXiv preprint arXiv:2302.04473*.