A Metric for MLLM Alignment in Large-scale Recommendation

Yubin Zhang*
zhangyubin@xiaohongshu.com
Xiaohongshu Inc.
Shanghai, China

Yanhua Huang*
yanhuahuang@xiaohongshu.com
Xiaohongshu Inc.
Shanghai, China

Haiming Xu
xuhaiming@xiaohongshu.com
Xiaohongshu Inc.
Shanghai, China

Mingliang Qi
mqi@xiaohongshu.com
Xiaohongshu Inc.
Shanghai, China

Chang Wang
wangchang2@xiaohongshu.com
Xiaohongshu Inc.
Shanghai, China

Jiarui Jin
jinjiarui@xiaohongshu.com
Xiaohongshu Inc.
Shanghai, China

Xiangyuan Ren
renxiangyuan@xiaohongshu.com
Xiaohongshu Inc.
Shanghai, China

Xiaodan Wang
xiaodan2@xiaohongshu.com
Xiaohongshu Inc.
Shanghai, China

Ruiwen Xu
ruiwenxu@xiaohongshu.com
Xiaohongshu Inc.
Shanghai, China

Abstract

Multimodal recommendation has emerged as a critical technique in modern recommender systems, leveraging content representations from advanced multimodal large language models (MLLMs). To ensure these representations are well-adapted, alignment with the recommender system is essential. However, evaluating the alignment of MLLMs for recommendation presents significant challenges due to three key issues: (1) static benchmarks are inaccurate because of the dynamism in real-world applications, (2) evaluations with online system, while accurate, are prohibitively expensive at scale, and (3) conventional metrics fail to provide actionable insights when learned representations underperform. To address these challenges, we propose the Leakage Impact Score (LIS), a novel metric for multimodal recommendation. Rather than directly assessing MLLMs, LIS efficiently measures the upper bound of preference data. We also share practical insights on deploying MLLMs with LIS in real-world scenarios. Online A/B tests on both Content Feed and Display Ads of Xiaohongshu's Explore Feed production demonstrate the effectiveness of our proposed method, showing significant improvements in user spent time and advertiser value.

Keywords

Multimodal Recommendation; Information Retrieval; Multimodal Large Language Model;

ACM Reference Format:

Yubin Zhang, Yanhua Huang, Haiming Xu, Mingliang Qi, Chang Wang, Jiarui Jin, Xiangyuan Ren, Xiaodan Wang, and Ruiwen Xu. 2025. A Metric for MLLM Alignment in Large-scale Recommendation. In *Proceedings of The 2nd Workshop on Evaluating and Applying Recommendation Systems with Large Language Models (RecSys '25).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Equal Contribution

RecSys '25, Prague, Czech Republic

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

With increasing user engagement in exploring content feeds, multimodal recommendation techniques play a vital role in modern recommender systems, as they can leverage rich information from content beyond user behaviors. Prior works have shown that combining multimodal content representations with user behavior data leads to substantial gains in several applications [5, 13, 15, 32, 34].

The rapid advancement of multimodal recommendation has been paralleled by remarkable progress in multimodal large language models (MLLMs), such as GPT-4V [1], Gemini [20], and Qwen-VL [2]. One of the key lessons in creating state-of-the-art MLLMs is their alignment with human preferences. Typically, the alignment is assessed using sophisticated benchmarks [22, 28, 33], where a weighted average of scores serves as the evaluation metric [2, 20]. While this metric works well for world knowledge domains, its applicability to recommender systems is inherently limited by the system's dynamism: shifting user interests and continuous algorithmic updates. For example, if the current recommender system already incorporates an MLLM's representations, similar representations will fail to deliver significant performance improvements.

To measure the alignment of an MLLM for the current recommender system, the AUC Improvement Score (AIS), defined as the AUC gain when applying MLLM's representation to the ranking model, is commonly adopted as the metric in practice [12, 18]. While AIS addresses system dynamism by leveraging recent behavior data and the production ranking model, it suffers from the substantial computation costs at scale, primarily due to its dependency on training MLLMs aligned with ranking models and inferring representations on billions of multimodal items. Furthermore, when AIS indicates marginal improvement and further optimization is required, it introduces a diagnostic challenge to distinguish whether the bottleneck lies in (1) suboptimal representation alignment or (2) ineffective utilization of existing representations.

To overcome these challenges, we propose the **Leakage Impact Score** (LIS), a novel metric that evaluates the quality of preference data construction rather than directly assessing MLLMs. LIS quantifies the ranking performance gap between models trained with and without leaked preference information. Unlike AIS, LIS eliminates the need for expensive MLLM training and inference on multimodal data. We argue that **LIS can accurately measure the upper bound of preference data, while the capabilities of MLLMs determine how closely we can approach this bound**. Moreover, we share our practical experience on validating preference data using LIS, resulting in effective preference data construction methods.

We conducted online A/B tests in the Explore Feed of Xiaohongshu (also known as the RedNote) [1], to further study LIS. Specifically, we first align MLLMs with preference data validated by LIS and then feed the aligned representation into the production ranking model. The results of the online A/B test in both recommendation and advertisement scenarios demonstrate significant improvements in core metrics, such as user time spent and advertiser value.

Our contributions are summarized as follows:

- We introduce the Leakage Impact Score (LIS), a novel metric that assesses the upper bound of given preference data, for multimodal recommendation.
- Practical experience is shared, where we introduce two types of preference data and demonstrate how to validate them with LIS efficiently.
- We conduct online A/B tests on two real-world applications, showing the effectiveness of our proposed methods in large-scale scenarios.

2 Preliminary and Motivation

This section outlines the standard pipeline for deploying MLLMs aligned with industrial recommender systems, highlighting the complexities and challenges of multimodal recommendation at scale.

As mentioned above, static benchmarks commonly used for evaluating world knowledge domains are not suitable for recommender systems due to their dynamism nature. Consequently, the AUC improvement score (AIS) on the most recent ranking model is adopted in practice, as it demonstrates consistent correlation between offline improvements and online performance. The standard deployment pipeline consists of three key steps:

Step 1. *Preference Data Construction*: Prior works have established the critical role of alignment and proposed various data construction methods that leverage user behaviors for preference alignment [13, 15, 34]. This step involves significant effort in preference data design, including cleaning and curation.

Step 2. *MLLM Training*: Using the constructed preference data, the next step involves fine-tuning MLLMs to achieve strong validation performance. Prior works have demonstrated the importance of preserving MLLMs' world knowledge while adapting them to recommendation tasks [21, 31, 32]. This step requires substantial computational resources for MLLM refinement.

Step 3. *Production Model Validation*. The final step evaluates the aligned representations by integrating them into the production model. Prior studies indicate that effectiveness depends heavily on how downstream models utilize these representations [5, 18, 26]. Therefore, this step involves both algorithmic exploration of representation application and computational overhead for inferring representations.

Note that the above three steps form a single iteration. When AIS indicates insufficient improvement, we need to diagnose the issue and repeat the iteration. In practice, this iterative process often requires multiple rounds of execution, creating a bottleneck for further applications of multimodal recommendation.

In the aforementioned pipeline, except for the essential work of human design, we observe substantial computational overhead, particularly when the final AIS performance is unsatisfactory. We identify the root cause as the inability of current methods to properly evaluate the relationship between preference data and the ranking model requirements. A metric to pre-validate preference data could allow the pipeline to focus exclusively on MLLM refinement and downstream application, significantly mitigating the overhead from multiple iterations.

3 Leakage Impact Score

We introduce the **Leakage Impact Score** (LIS), a novel metric that leverages the concept of data leakage to measure preference data. Data leakage occurs when information from outside the training is involved in the training procedure. Our work focuses specifically on temporal leakage—for instance, when predicting yesterday's behavior while inadvertently including today's data in training. This example mirrors the real-world constraint where production systems cannot access a user's future interests, making the trained model overestimated [14, 25].

While data leakage is typically avoided, we repurpose this phenomenon constructively. In recommender systems, models trained with leaked data exhibit offline performance that fails to generalize to online deployment. This discrepancy arises because the online system cannot access the leaked data before inference. Note that data unavailable in online deployment may still serve a constructive purpose in offline settings: it provides a mechanism to quantify data importance—irrelevant leaks cause negligible impact, while informative ones lead to significant overestimation.

To this end, we define the LIS as the impact when involving temporally leaked information in the model. The model here refers to the recommender, not the MLLM, thus avoiding the computational overhead associated with MLLMs. In particular, we construct features from leaked data, and the AUC improvement of applying these features to the production ranking model is adopted as the LIS. Note that LIS introduces the upper bound of the preference data, as it is equivalent to a means of accurately predicting future behaviors. If we could validate the effectiveness of an approach to construct the preference data, we believe that MLLMs are able to learn generalized patterns from it.

Here is an example to demonstrate how to calculate LIS in practice. Consider click-through rate prediction, where the production model predicts a user's click probability given their history and a candidate item. If we augment this model by incorporating the

[1] https://www.xiaohongshu.com/explore

next item clicked by the user as an additional feature to the production model, the resulting AUC improvement constitutes the LIS. A high LIS suggests this click behavior contains valuable signal for preference data construction. However, since user behaviors are inherently noisy, effectively distilling this signal into MLLM training remains challenging—a challenge we address in the following section.

4 Practical Experiences

In this section, we introduce two types of preference data and how to validate them with LIS. The first type of preference data is sparse representations learned by the recommender. In particular, without loss of generality, we consider the embedding of item ID in the ranking model. If there are multiple embedding slots representing items, we can choose the most important one through the feature importance techniques. We validate the impact of leaked item ID embedding as follows. Formally, let $\mathcal{M}_T$ denote the production ranking model serving online at date T, i.e., the model have never seen behaviors on date T or thereafter. We replace its item ID embedding with those from $\mathcal{M}_{T+n}$, where $T + n$ denotes the n-th date after date T. As shown in Table 1, this substituion yields LIS values of 0.06 and 0.09 on Xiaohongshu's Explore Feed ranking model. In our settings, where an absolute increase of 0.0010 in AUC is considered significant, these results clearly demonstrate the potential effectiveness of ID embeddings as preference data.

We attribute this phenomenon to the fact that the item ID serves as the unique identifier for an item, causing the recommender to encode the item's distinctive and important information within its ID embedding. The next section will show that MLLMs can learn generalized information from ID embeddings and gain significant improvement in online experiments.

Table 1: LIS by leaked ID embedding using the data from n days later.

	n = 7	n = 30
LIS	+0.06	+0.09

The second type of preference data is from the retrieval perspective. For each item, we identify its 5 most similar items as its side information. This side information is considered as the feature of the target item. Note that if the target item is in the cold-start phase, incorporating similar items with well-learned representations as auxiliary inputs may enhance the cold-start performance. With leaked data, we can accurately identify items similar to a cold-start item with the help of behavior data [29], since we already have knowledge of posterior behaviors.

However, the results show negligible LIS improvement. This suggests the potential existence of analogous information within the current ranking model. This finding is particularly insightful as it reveals which preference data types merit MLLM-based learning versus those that can be effectively handled by existing system components.

5 Related Work

5.1 MLLM Evaluation

The primary design objective for MLLMs is to create intelligent chatbots that can thoroughly address human queries spanning both perceptual understanding and logical reasoning. To evaluate these comprehensive capabilities, researchers have developed numerous specialized benchmarks across world knowledge domains. Early studies proposed evaluating MLLMs with visual understanding tasks [8, 9], extend by following works on multilingual [19, 28], video understanding [6, 33], and mathematics [17, 22]. Beyond general capabilities, researchers have also explored how to evaluate MLLMs for specific downstream tasks [7, 11, 16, 23, 30], focusing more on the mastery of domain knowledge and skills. However, all these benchmarks employ static construction methodologies, making them unable to generalize to recommender system scenarios—which are inherently dynamic systems where user interests constantly evolve and algorithmic upgrades occur continuously.

5.2 Multimodal Recommendation

Multimodal recommendation aims to leverage multimodal representations to improve the recommendation performance, which is critical in modern recommender systems, especially for multi-media content scenarios like TikTok and RedNote. Early works [10, 24] only considered the utilization of multimodal representations, ignoring the alignment between multimodal models and recommender systems. Recent works have presented various sophisticated strategies for constructing perference data to align MLLMs with recommender systems. CB2CF [3] is the first work that considers this alignment by incorporating users' behaviors, where the content encoder learns human preference from collaborative filtering vectors. [13] further addressed the instability issue in the learning procedure of original CB2CF and successfully applied it to the diversified recommendation task in a large-scale scenario. [32] proposed maintaining MLLMS' world knowledge capabilities with auxiliary tasks about content predictions. [18] argued that MLLMs should learn from deep signals in user behaviors such as search and purchase, proposing to mine hard negatives when constructing negative samples. While the above works have made significant progress in multimodal recommendation, the process of deploying MLLMs for recommendation still suffers from challenges in evaluating the alignment. In this paper, we highlight the challenges and complexities within the multimodal recommendation. To address them, we introduce LIS, which measures the upper bound of given preference data, preventing computational overhead from training MLLMs. Moreover, we also introduce a novel approach that aligns MLLMs with the sparse embeddings learned by the recommender. The online A/B tests on two real-world scenarios demonstrated the effectiveness of our proposed method at scale.

6 Experiments

6.1 Implmentation Details

We conduct large-scale online A/B tests using sparse item ID embeddings as preference data, following the validation described in Section 4. Our experimental setup employs InternVL [4] as the base MLLM, with all sparse embeddings extracted from the production

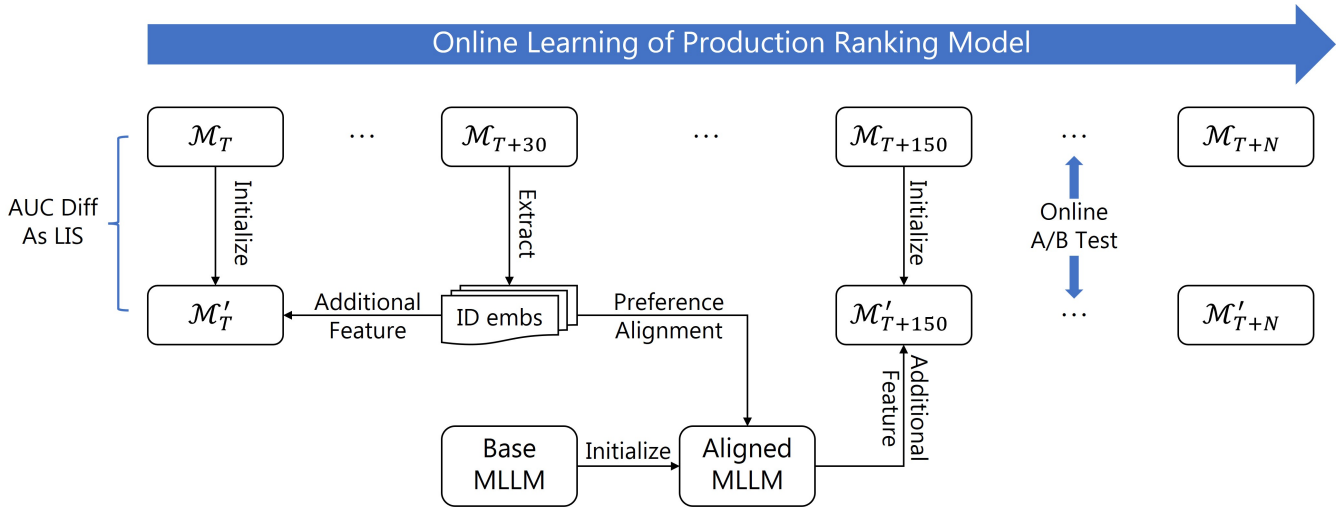

Figure 1: This figure shows how to validate preference data with LIS and how we conduct online A/B tests, where we use ID embeddings as an example of preference data.

ranking model snapshot of May 2024. For data pre-processing, we only retain item embeddings with more than 10000 updates to guarantee statistical reliability. Additionally, we apply data curation inspired by MetaCLIP [27]. During MLLM training, we monitor convergence using the mean recall metric on the validation set.

We evaluate the learned representation in two real-world scenarios, Content Feed and Display Ads, of Xiaohongshu's Explore Feed production. Note that the ranking models for these two scenarios are separate, so we train distinct MLLMs for each scenario. The control group comprises 10% of randomly selected Xiaohongshu users and applies the production ranking model. For the treatment group, we also randomly select 10% of users. Each group contains tens of millions of users, with no overlap between groups. Compared to the control group, the ranking model in the treatment group incorporates learned representations as an additional feature, which is the only difference between models in two groups. The added feature—a dense vector with fewer than 100 dimensions—introduces negligible parameter growth, maintaining experimental validity.

6.2 Online A/B Tests

For the Content Feed scenario, the experiment was conducted in December 2024. We observe statistically significant improvements across all four key performance metrics: time spent, the number of reads, the number of engagements, and APP lifetime over 30 days (LT30), as presented in Table 2.

Table 2: Online A/B test result in the Content Feed scenario of Xiaohongshu's Explore Feed.

	Time	Reads	Engagements	LT30
Improvement	+0.13%	+0.27%	+0.40%	+0.02%

For the Display Ads scenario, the experiment was conducted in November 2024. We observe statistically significant improvements across all four key performance metrics, as shown in Table 3, where Advertiser Value (ADVV) and COST indicate the value of advertisements, while Impression and CTR represent the user experience.

Table 3: Online A/B test result in the Display Ads scenario of Xiaohongshu's Explore Feed.

	Impression	ADVV	COST	CTR
Improvement	+0.32%	+0.86%	+0.79%	+0.43%

In the aforementioned experiments, we utilized the recommender's representations as preference data and aligned the MLLM with the recommender system through this approach, achieving promising results. It's worth noting that in our scenario, an item's life-cycle is typically much shorter than 3 months, while the time interval between collecting preference data and conducting online experiments was maintained at least 4 months. That is, none of the items in the preference data appeared as target items during online experiments. Even when directly using the preference data as input for online experiments, no improvement could be obtained. Therefore, the online improvement primarily stem from the MLLM extracting generalizable patterns from the preference data and successfully applying them to previously unseen items.

7 Limitations

While LIS measures the upper bound of given preference data, two challenges remain. First, how to approach the bound, i.e., how to maximize the MLLM's ability to learn the information contained in the preference data. We argue that, in addition to improving the MLLM's general capabilities, potential approaches may include

hard mining and curriculum learning. The second challenge concerns how to effectively utilize the learned representations in recommender systems. Since the leaked information carries highly predictive information, ranking models can readily extract useful patterns from it. However, in practical deployment scenarios, the preference data that MLLMs learn from contains no leaked information for online recommenders. Consequently, we contend that the representations obtained through alignment require further investigation of their application methodologies.

8 Conclusion

In this paper, we present the leakage impact score (LIS), a novel metric for multimodal recommendation. LIS enables assessment of potential effectiveness of preference data before MLLM alignment. Our approach significantly improves deployment efficiency by providing an early-stage validation mechanism. We further present practical insights on preference data construction, demonstrating that sparse representations learned by ranking models serve as particularly effective preference data for multimodal recommendation. Online A/B tests conducted on two production scenarios, Content Feed and Display Ads in Xiaohongshu's Explore Feed, demonstrate significant improvements, confirming the practical value of our proposed techniques.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[3] Oren Barkan, Noam Koenigstein, Eylon Yogev, and Ori Katz. 2019. CB2CF: a neural multiview content-to-collaborative filtering model for completely cold item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 228–236.

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 24185–24198.

[5] Xiuqi Deng, Lu Xu, Xiyao Li, Jinkai Yu, Erpeng Xue, Zhongyuan Wang, Di Zhang, Zhaojie Liu, Guorui Zhou, Yang Song, et al. 2024. End-to-end training of Multimodal Model and ranking Model. *arXiv preprint arXiv:2404.06078* (2024).

[6] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 24108–24118.

[7] Fan Gao, Hang Jiang, Rui Yang, Qingcheng Zeng, Jinghui Lu, Moritz Blum, Tianwei She, Yuang Jiang, and Irene Li. 2024. Evaluating large language models on wikipedia-style survey generation. In *Findings of the Association for Computational Linguistics ACL 2024*. 5405–5418.

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.

[9] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3608–3617.

[10] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[11] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22170–22183.

[12] Yanhua Huang, Yuqi Chen, Xiong Cao, Rui Yang, Mingliang Qi, Yinghao Zhu, Qingchang Han, Yaowei Liu, Zhaoyu Liu, Xuefeng Yao, et al. 2025. Towards Large-scale Generative Ranking. *arXiv preprint arXiv:2505.04180* (2025).

[13] Yanhua Huang, Weikun Wang, Lei Zhang, and Ruiwen Xu. 2021. Sliding spectrum decomposition for diversified recommendation. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 3041–3049.

[14] Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. 2023. A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems* 41, 3 (2023), 1–27.

[15] Yifan Liu, Kangning Zhang, Xiangyuan Ren, Yanhua Huang, Jiarui Jin, Yingjie Qin, Ruilong Su, Ruiwen Xu, Yong Yu, and Weinan Zhang. 2024. Alignrec: Aligning and training in multimodal recommendations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 1503–1512.

[16] Yuhang Lu, Yichen Yao, Jiadong Tu, Jiangnan Shao, Yuexin Ma, and Xinge Zhu. 2025. Can lvlms obtain a driver's license? a benchmark towards reliable agi for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 5838–5846.

[17] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284* (2024).

[18] Xiang-Rong Sheng, Feifan Yang, Litong Gong, Biao Wang, Zhangming Chan, Yujing Zhang, Yueyao Cheng, Yong-Nan Zhu, Tiezheng Ge, Han Zhu, et al. 2024. Enhancing Taobao Display Advertising with Multimodal Representations: Challenges, Approaches and Insights. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 4858–4865.

[19] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985* (2024).

[20] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[21] Jianling Wang, Yifan Liu, Yinghao Sun, Xuejian Ma, Yueqi Wang, He Ma, Zhengyang Su, Minmin Chen, Mingyan Gao, Onkar Dalal, et al. 2025. User Feedback Alignment for LLM-powered Exploration in Large-scale Recommendation Systems. *arXiv preprint arXiv:2504.05522* (2025).

[22] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems* 37 (2024), 95095–95169.

[23] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. 2024. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19757–19767.

[24] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 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.

[25] Xin Xin, Jiyuan Yang, Hanbing Wang, Jun Ma, Pengjie Ren, Hengliang Luo, Xinlei Shi, Zhumin Chen, and Zhaochun Ren. 2023. On the user behavior leakage from recommender system exposure. *ACM Transactions on Information Systems* 41, 3 (2023), 1–25.

[26] Haibo Xing, Kanefumi Matsuyama, Hao Deng, Jinxin Hu, Yu Zhang, and Xiaoyi Zeng. 2025. ESANS: Effective and Semantic-Aware Negative Sampling for Large-Scale Retrieval Systems. In *Proceedings of the ACM on Web Conference 2025*. 462–471.

[27] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023. Demystifying clip data. *arXiv preprint arXiv:2309.16671* (2023).

[28] Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, et al. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497* (2025).

[29] Xiaoyong Yang, Yadong Zhu, Yi Zhang, Xiaobo Wang, and Quan Yuan. 2020. Large scale product graph construction for recommendation in e-commerce. *arXiv preprint arXiv:2010.05525* (2020).

[30] Kunyu Yu, Rui Yang, Jingchi Liao, Siqi Li, Huitao Li, Irene Li, Yifan Peng, Rishikesan Kamaleswaran, and Nan Liu. 2025. Benchmarking Foundation Models with Multimodal Public Electronic Health Records. *arXiv preprint arXiv:2507.14824* (2025).

[31] Qing Yu, Xiaobei Wang, Shuchang Liu, Yandong Bai, Xiaoyu Yang, Xueliang Wang, Chang Meng, Shanshan Wu, Hailan Yang, Huihui Xiao, et al. 2025. Who You Are Matters: Bridging Topics and Social Roles via LLM-Enhanced Logical Recommendation. *arXiv preprint arXiv:2505.10940* (2025).

[32] Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Xiangyu Zhao, Yan Gao, Yao Hu, and Enhong Chen. 2025. Notellm-2: Multimodal large representation models for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 2815–2826.

[33] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv e-prints* (2024), arXiv–2406.

[34] Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. 2023. Bootstrap latent representations for multimodal recommendation. In *Proceedings of the ACM web conference 2023*. 845–854.