Describe What You See with Multimodal Large Language Models to Enhance Video Recommendations

Marco De Nadai
mdenadai@spotify.com
Spotify
Denmark

Andreas Damianou
andreasd@spotify.com
Spotify
United Kingdom

Mounia Lalmas
mounial@spotify.com
Spotify
United Kingdom

Abstract

Existing video recommender systems rely primarily on user-defined metadata or on low-level visual and acoustic signals extracted by specialised encoders. These low-level features describe what appears on the screen but miss deeper semantics such as intent, humour, and world knowledge that make clips *resonate with viewers*. For example, is a 30-second clip simply a singer on a rooftop, or an ironic parody filmed amid the fairy chimneys of Cappadocia, Turkey? Such distinctions are critical to personalised recommendations yet remain invisible to traditional encoding pipelines. In this paper, we introduce a simple, recommendation system-agnostic zero-finetuning framework that injects high-level semantics into the recommendation pipeline by prompting an off-the-shelf Multimodal Large Language Model (MLLM) to summarise each clip into a rich natural-language description (e.g. "a superhero parody with slapstick fights and orchestral stabs"), bridging the gap between raw content and user intent. We use MLLM output with a state-of-the-art text encoder and feed it into standard collaborative, content-based, and generative recommenders. On the MicroLens-100K dataset, which emulates user interactions with TikTok-style videos, our framework consistently surpasses conventional video, audio, and metadata features in five representative models. Our findings highlight the promise of leveraging MLLMs as on-the-fly knowledge extractors to build more intent-aware video recommenders.

Keywords

Multimodal Large Language Models, LLMs, video recommendation systems, video

ACM Reference Format:
Marco De Nadai, Andreas Damianou, and Mounia Lalmas. 2025. Describe What You See with Multimodal Large Language Models to Enhance Video Recommendations. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3705328.3759303

1 Introduction

The rapid rise of video platforms such as TikTok, Instagram Reels, and YouTube Shorts has highlighted the critical importance of developing effective video recommendation systems [7, 33, 34, 37]. In fact, TikTok's success is often attributed in part to its recommendation algorithm, which is estimated to align with user interests 30% to 50% of the time, alongside other factors such as upload timing and social network effects [24]. This surge in popularity has not only underscored the influence of recommendation algorithms, but also exposed the limitations of conventional approaches in understanding what truly drives user engagement.

Indeed, conventional content-based models distill each clip into low-level visual, audio, and textual embeddings derived from pretrained encoders or sparse metadata [7, 33, 34, 37]. Although these representations capture motion, colour, or keywords, they remain blind to higher-order semantics and cultural references; the very cues that make a clip *resonate with viewers*. For instance, optical flow can reveal that "someone is dancing on a rooftop" but not that the dance parodies a 1990s superhero trope. This lack of semantic understanding limits the ability of traditional systems to fully capture the *why* behind user preferences, particularly for content rich in implicit meaning or tied to world knowledge.

In this paper, we show that Multimodal Large Language Models (MLLMs) [2, 3, 14, 15, 25] constitute a promising solution. MLLMs can jointly analyze video frames, audio signals, and metadata to generate semantically rich captions that describe video content in detail. These captions encapsulate not only observable content, but also intent, style, and context, serving as effective proxies for user preferences. For example, given a gameplay footage from an anime-style fighting game (e.g. Naruto), a MLLM could recognize the characters' signature lines and link them back to the series' world, thereby placing the clip in a context that fans will recognise (see Figure 1).

In more detail, this paper introduces a recommendation system *model-agnostic* framework that plugs MLLM-generated data into existing recommender pipelines without any major change. We run a zero-shot prompting recipe on open weights MLLMs to describe each clip, concatenate video and audio captions, and feed the resulting text to standard recommenders. Using MLLMs to generate multimodal captions, our approach bridges the gap between raw content and user intent, enabling systems to better understand and align with user preferences.

To evaluate the effectiveness of this framework, we conduct extensive experiments on the publicly available MicroLens-100K dataset [17], a large-scale video dataset collected from a real-world video platform. We compare MLLM-generated features against traditional video, audio, and metadata-based features across various recommendation architectures, including embedding-based and generative models. Our results demonstrate that MLLM-derived features significantly outperform traditional baselines. Notably, multimodal captions, generated by analyzing both video and audio

RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/2025/09
https://doi.org/10.1145/3705328.3759303

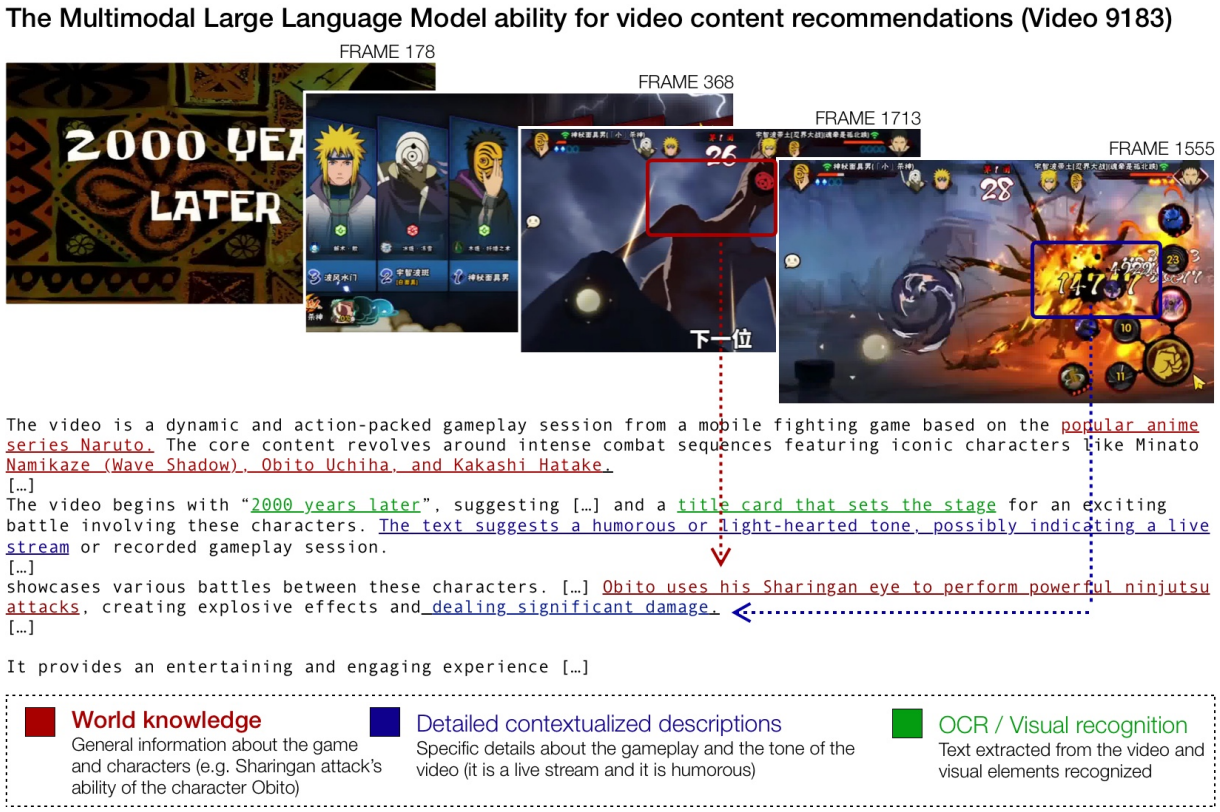

Figure 1: A qualitative breakdown of a gameplay clip from a Naruto mobile fighting game (video ID 9183), demonstrating the potential of MLLMs. A MLLM (here, Qwen-VL) (i) extracts on-screen text via OCR (green), (ii) grounds entities and actions in the franchise's world knowledge (red), and (iii) composes fine-grained, time-aligned contextual descriptions of the battle dynamics (blue). These detailed outputs could resonate with users, thereby improving the recommendation system's performance.

inputs jointly, yield the most robust representations, leading to substantial improvements in recommendation quality.

The contributions of this paper are summarized as follows:

- We present the first systematic and reproducible study of leveraging open-weight MLLMs for video recommendation. This allows practitioners to use publicly available models in both academic and industry settings.
- We release a lightweight, *zero-finetuning* pipeline that converts clips into rich textual representations. This zero-finetuning approach is particularly advantageous as end-to-end adaptation of the requisite large-scale MLLMs is often infeasible due to substantial computational and temporal costs.
- Through extensive experiments on the MicroLens-100K dataset, we show that MLLM captions consistently outperform visual, audio and metadata baselines. These gains are more pronounced in videos longer than 30 seconds.
- We make our prompts and generated data publicly available to foster reproducibility.[1]
- Our findings underscore the transformative potential of MLLMs in video recommendation tasks, offering a pathway to systems that deliver more accurate, contextually rich, and user-aligned recommendations.

2 Related Work

Video Recommendations. The lack of publicly available large-scale video recommendation datasets with raw videos has made it difficult to develop and research novel recommender system models [17]. As a result, most published research relies on collaborative signals (e.g. [8]), metadata, pre-computed features computed from images (e.g. [12, 13, 13, 32, 37, 38]) or video features extracted by proprietary systems (e.g. [11]), where raw data is kept hidden [1]. For example, MMGCN [26] use both image covers and textual metadata in a Graph Neural Network to leverage multi-modal information. Lee *et al.* [11] propose a content-based similarity learning recommendation system on 8 million videos. Singh *et al.* [21] use video embeddings for generative recommendations at Youtube.

Fortunately, the release of MicroLens [17] has allowed the direct use of raw videos to extract multiple types of information (e.g. video embeddings with custom pipelines, images) in academic research. For example, Jiang *et al.* [9] leverage video features to identify user interest groups for recommendation purposes, while Ni *et al.* [17] finetune video encoders specifically for recommendation systems. In contrast to these approaches, our method adopts a *zero-finetuning*, *model-agnostic* framework that demonstrates how MLLMs can improve video recommendation performance.

[1] https://huggingface.co/datasets/marcodena/video-recs-describe-what-you-see

MLLMs. Building on the long-standing AI goal of vision-language integration, MLLMs have surged in popularity by leveraging the power of LLMs for unprecedented cross-modal understanding and generation [2, 3, 14, 15, 25]. These models can process and synthesize information from diverse inputs, such as images, video frames, and audio signals, to produce rich, contextually aware textual descriptions. This capability allows MLLMs to move beyond superficial feature extraction, enabling them to identify complex entities, understand actions, and ground observations in extensive world knowledge, as demonstrated in tasks such as detailed scene description and event summarization.

For this reason, adaptations of MLLMs for recommendations are starting to emerge. For example, Peixuan *et al.* [19] trains an MLLM from scratch to classify movies into genres, suggesting that they can be used for recommendations. Fu *et al.* [6] finetune MLLMs for sequential recommendations. Zhou *et al.* [39] benchmarks various image MLLMs to assess whether they can be used directly for Amazon item recommendations. Ye *et al* [31] describe user preferences through text and finetune an MLLM on user interactions to predict user interests from textual metadata and the cover image of videos.

While MLLMs are beginning to be explored in recommendation contexts, a significant gap remains. To the best of our knowledge, no prior work has systematically examined how open-weight MLLMs can be effectively leveraged to capture the complex interplay of world knowledge, aesthetic features, intent, style, and cultural references embedded in video content, which are factors essential for truly understanding and aligning with user preferences in video recommendation systems. Notably, our approach does not require finetuning and does not generate recommendations directly, offering a more scalable and efficient pathway for video recommendations.

Our work directly addresses this gap by proposing a model-agnostic framework that extracts these rich, high-level semantic features from videos using MLLMs, demonstrating their superior performance over traditional methods.

3 Research Hypotheses

We investigate whether MLLMs can systematically improve the quality of video recommendations. Our core research question is: *Do MLLM-derived captions outperform classical content features in standard ranking tasks?*

MLLMs capture nuanced intent and contextual information often missed by traditional features. We hypothesize that simply replacing traditional visual, audio, and sparse textual features with semantically rich MLLM-generated descriptions will result in higher HR@K and nDCG@K scores on MicroLens-100K.

To test this, we train two representative models: one using classical video, metadata, and audio features, and another using MLLM-generated captions from videos and audio content. We evaluate both models on MicroLens-100K using a global time-based split to mimic a production setting, across multiple days.

4 The Framework

Our framework addresses the semantic sparsity inherent in traditional video recommendation systems through four guiding principles. **First**, we use widely available, state-of-the-art open-weight models. **Second**, we focus on cross-modal grounding to reduce ambiguities by fusing audio, video and textual signals. **Third**, we use frozen MLLMs, which lowers computational costs and enables frequent, low-cost training of recommendation models. **Finally**, our framework is designed to be backbone-agnostic: MLLM-generated embeddings can be seamlessly integrated into any recommendation architecture (e.g., two-towers models, generative models).

5 Experimental Results

5.1 Setting

5.1.1 Data. We use the publicly available MicroLens-100K [17], a large-scale content-driven video dataset sourced from a real-world video mobile platform with approximately 34 million users. On this platform, creators upload and share vertical format videos ranging from a few seconds to around 10 minutes in length. MicroLens-100K includes 100,000 users, 19,738 items, and 719,405 interactions, resulting in a sparsity of 99.96%. Following Ni *et al.* [17], we retain users with at least two interactions and limit each user's interaction history to their ten most recent interactions. The dataset contains raw, full-length video files along with associated metadata, making it particularly suitable for evaluating our framework. While our framework is capable of processing longer videos (up to several hours), we focus on short-term content for computational efficiency. Specifically, we restrict videos to a maximum length of 4 minutes and downscale them to a resolution of 426×224 pixels.

5.1.2 Modalities. We consider various state-of-the-art models and modalities to represent videos. To ensure a fair comparison, we select encoders with comparable parameter counts.

Metadata uses a sentence text encoder that generate embeddings from video titles (written by the video creators). We use BGE-large [29], a 335M parameters text encoder built on top of BERT.

Audio and Video use dedicated audio and video encoders, respectively, to encode the content of the video. For audio, we use CLAP [28], a 194M parameters model trained on a diverse set of (audio, text) pairs. CLAP is not speech-specific, making it well-suited for videos that include both music and speech. Video features are extracted using VideoMAE [22], following Ni *et al.* [17]. VideoMAE is a 307M parameters model that processes only 16 frames per video. To best capture the temporal dynamics, we extract these frames from the first 30 seconds. In line with recent findings, we observe that averaging multiple VideoMAE embeddings degrades performance [22]. We also tested the recently released Google VideoPrism [36], available on Hugging Face as of June 2025, but found no performance improvement on our dataset.

MLLM (audio, video) use Qwen-VL [25], an open-weights MLLM pre-trained on 1.4B image-text pairs and videos, chosen for its ability to generate knowledge-grounded captions (e.g., "a couple dancing on a Parisian rooftop at sunset"). Its hybrid encoder fuses vision tokens with text prompts, enabling effective *cross-modal grounding* while mitigating hallucinations often seen in video encoders. To bridge the audio gap, we propose a two-stage approach: 1. *Audio Transcription*: We extract speech and texture using Whisper [20], selected for its robustness to background noise. 2. *Audio Knowledge Fusion*: We feed the transcriptions and sound classifications (e.g.

"dramatic music") into Qwen-Audio [3], which generates intent-aware descriptions (e.g. "upbeat soundtrack with a lighthearted tone"). This modular design enables *decoupled audio and video analysis*, while compensating for the absence of a unified audio-visual MLLM.

5.1.3 Recommendation models. Our framework is model-agnostic. To demonstrate its versatility, we evaluate two widely used models in academia and industry: the two-towers model, originally proposed by Youtube [4], and the generative model SASRec [10]. Both models consist of an item and a user encoder. In the two-towers model, the user embedding is computed as the average embedding of the videos watched by the user, following [17]. In SASRec, the user encoder is implemented as a decoder-only Transformer. Its output is further processed through a Residual MLP before computing the loss, following the design of PinnerFormer [18]. The item encoder projects embedding features through a standard 1-layer Residual MLP with ReLU activations. We use the same hyperparameters as [17] and do not perform any hyperparameter tuning. We do not here consider solutions that finetune vision models, such as [17], due to their significantly high computational cost.

5.1.4 Evaluation. We evaluate the performance of our recommendation task using HitRate (HR) and Normalized Discounted Cumulative Gain (nDCG) @ K, with K set to 10 and 30. In line with recent evaluation standards [16], we adopt a global time-based split with a 1-day rolling window to closely mimic production settings. For each of the last seven days, we perform daily evaluations. On day k, the model is trained on data up to day k-2, with validation on day k-1 to determine the optimal number of training epochs using early stopping (patience=5). Once the best epoch count is identified, we retrain the model using data up to day k-1 and evaluate it on day k. Unless specified otherwise, all performance comparisons (e.g., between Model A and Model B) are based on a paired t-test.

5.2 Results

Prompting an off-the-shelf MLLM to summarise each clip into natural-language descriptions consistently improves performance across all tested modalities. On MicroLens-100K, replacing raw background audio features with MLLM-generated text boosts the two-towers' HR@10 from 0.0253 to 0.0405 and nDCG@10 from 0.0130 to 0.0214, a $\sim$ 60% relative gain (see Table 1). SASRec demonstrates similarly strong improvements, with a relative gain of 35%. Converting audio waveforms to text recovers aspects such as theme, mood, and world knowledge, information typically lost in traditionally spectrogram-based representations. Similarly, adding scene-level captions boosts HR@10 from 0.0393 to 0.0489 in the two-towers model (+24%) and lifts SASRec to 0.0482, outperforming raw frames and creator-written titles by 4% to 18%. In essence, pixels show *what happens on-screen*, titles reflect *what the uploader hopes* will attract clicks, but MLLM-generated text captures *why viewers might care*: conveying tone, parody, and cultural cues that standard recommenders can now begin to utilize. Because MLLM-generated descriptions condenses minute-long sequences into a single, coherent summary, they also help overcome the long-horizon limitations of conventional vision-based features [23, 27].

Table 1: Performance of various modalities features on two widely used representative models: the two-towers and SASRec. The results show that MLLM-generated descriptions improve the performance in these recommendation systems.

Representation	HR@10	HR@30	nDCG@10	nDCG@30
Two-Towers (Youtube) [4]				
Metadata	0.0414	0.0721	0.0214	0.0286
Audio	0.0253	0.0439	0.0130	0.0173
Video	0.0393	0.0729	0.0201	0.0280
Metadata + video	0.0428	0.0775	0.0222	0.0304
MLLM audio	0.0405	0.0742	0.0214	0.0294
MLLM video	**0.0489**	**0.0879**	**0.0264**	**0.355**
SASRec [10] (Generative recommendations)				
Metadata	0.0460	0.0791	0.0249	0.0326
Audio	0.0338	0.0555	0.0186	0.0237
Video	0.0456	0.0829	0.0245	0.0332
Metadata + video	0.0462	0.0813	0.0246	0.0329
MLLM audio	0.0454	0.0816	0.0245	0.0330
MLLM video	**0.0482**	**0.0877**	**0.0261**	**0.0353**

Table 2: Effect of encoder variants on SASRec performance.

Model	HR@10	HR@30	nDCG@10	nDCG@30
Baseline	0.0482	0.0877	0.0261	0.0353
Larger text encoder				
Qwen emb. [35]	0.0479	0.0878	0.0256	0.0349
Larger MLLM				
Qwen-VL 7B	0.0480	0.0862	0.0256	0.0346

Scaling encoders and backbones. Table 2 explores two scaling strategies while keeping the rest of the pipeline unchanged: (i) swapping the text encoder with a larger, higher-performing model (Qwen embeddings 0.6B [35], released in June 2025), and (ii) replacing the MLLM backbone with a larger variant (Qwen-VL 7B). While qualitative inspection reveals that Qwen-VL 7B generates richer captions with more grounded world knowledge, these upgrades do no yield measurable improvements. This suggests that, for this dataset, the baseline MLLM already captures the most relevant information, and that merely increasing model size offers diminishing returns once a coherent video-level description is achieved.

6 Conclusion

We introduce a zero-finetuning plug-and-play framework that transforms raw audio-video inputs into rich text descriptions using off-the-shelf MLLMs. These features can be seamlessly integrated into standard collaborative, content-based, and generative recommenders. Without any additional training, our framework achieves up to 60% relative gains on the MicroLens-100K dataset. As emerging omni-MLLMs [40], such as Qwen Omni [30] and Vita [5], continue to advance in jointly modeling vision, audio, and text, our approach provides a low-barrier pathway for both research prototypes and production systems to improve recommendations by recognizing not only *what* is on-screen, but also understanding *why* it may resonate with individual viewers.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[2] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. 2024. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413* (2024).

[3] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919* (2023).

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

[5] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957* (2025).

[6] Junchen Fu, Xuri Ge, Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, Kaiwen Zheng, Yongxin Ni, and Joemon M Jose. 2024. Efficient and Effective Adaptation of Multimodal Foundation Models in Sequential Recommendation. *arXiv preprint arXiv:2411.02992* (2024).

[7] Xudong Gong, Qinlin Feng, Yuan Zhang, Jiangling Qin, Weijie Ding, Biao Li, Peng Jiang, and Kun Gai. 2022. Real-time Short Video Recommendation on Mobile Devices. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3103–3112.

[8] Pan Gu, Haiyang Hu, and Guandong Xu. 2024. Modeling multi-behavior sequence via HyperGRU contrastive network for micro-video recommendation. *Knowledge-Based Systems* 295 (2024), 111841.

[9] Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. 2020. What aspect do you like: Multi-scale time-aware user interest modeling for micro-video recommendation. In *Proceedings of the 28th ACM International conference on Multimedia*. 3487–3495.

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

[11] Joonseok Lee and Sami Abu-El-Haija. 2017. Large-scale content-only video recommendation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 987–995.

[12] Chenyi Lei, Yong Liu, Lingzi Zhang, Guoxin Wang, Haihong Tang, Houqiang Li, and Chunyan Miao. 2021. Semi: A sequential multi-modal information transfer network for e-commerce micro-video recommendations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3161–3171.

[13] Youhua Li, Hanwen Du, Yongxin Ni, Pengpeng Zhao, Qi Guo, Fajie Yuan, and Xiaofang Zhou. 2024. Multi-modality is all you need for transferable recommender systems. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE, 5008–5021.

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

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *NeurIPS*.

[16] Zaiqiao Meng, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2020. Exploring data splitting strategies for the evaluation of recommendation models. In *Proceedings of the 14th acm conference on recommender systems*. 681–686.

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

[18] Nikil Pancha, Andrew Zhai, Jure Leskovec, and Charles Rosenberg. 2022. Pinnerformer: Sequence modeling for user representation at pinterest. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 3702–3712.

[19] Peixuan Qi. 2024. Movie Visual and Speech Analysis through Multi-Modal LLM for Recommendation Systems. *IEEE Access* (2024).

[20] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[21] Anima Singh, Trung Vu, Nikhil Mehta, Raghunandan Keshavan, Maheswaran Sathiamoorthy, Yilin Zheng, Lichan Hong, Lukasz Heldt, Li Wei, Devansh Tandon, et al. 2024. Better generalization with semantic ids: A case study in ranking for recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1039–1044.

[22] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems* 35 (2022), 10078–10093.

[23] Elahe Vahdani and Yingli Tian. 2022. Deep learning-based action detection in untrimmed videos: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4302–4320.

[24] Karan Vombatkere, Sepehr Mousavi, Savvas Zannettou, Franziska Roesner, and Krishna P Gummadi. 2024. Tiktok and the art of personalization: Investigating exploration and exploitation on social media feeds. In *Proceedings of the ACM Web Conference 2024*. 3789–3797.

[25] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).

[26] 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.

[27] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 2022. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13587–13597.

[28] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[29] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597 [cs.CL]

[30] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215* (2025).

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

[32] Yisong Yu, Beihong Jin, Jiageng Song, Beibei Li, Yiyuan Zheng, and Wei Zhuo. 2022. Improving micro-video recommendation by controlling position bias. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 508–523.

[33] Yisong Yu, Beihong Jin, Jiageng Song, Beibei Li, Yiyuan Zheng, and Wei Zhuo. 2023. Improving Micro-video Recommendation by Controlling Position Bias. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part I*. Springer, 508–523.

[34] Guanghu Yuan, Fajie Yuan, Yudong Li, Beibei Kong, Shujie Li, Lei Chen, Min Yang, Chenyun Yu, Bo Hu, Zang Li, et al. 2022. Tenrec: A Large-scale Multipurpose Benchmark Dataset for Recommender Systems. *arXiv preprint arXiv:2210.10629* (2022).

[35] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176* (2025).

[36] Long Zhao, Nitesh B. Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J. Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, Rachel Hornung, Florian Schroff, Ming-Hsuan Yang, David A. Ross, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Ting Liu, and Boqing Gong. 2024. VideoPrism: A Foundational Visual Encoder for Video Understanding. In *International Conference on Machine Learning (ICML)*.

[37] Yu Zheng, Chen Gao, Jingtao Ding, Lingling Yi, Depeng Jin, Yong Li, and Meng Wang. 2022. Dvr: micro-video recommendation optimizing watch-time-gain under duration bias. In *Proceedings of the 30th ACM International Conference on Multimedia*. 334–345.

[38] Ting Zhong, Jian Lang, Yifan Zhang, Zhangtao Cheng, Kunpeng Zhang, and Fan Zhou. 2024. Predicting Micro-video Popularity via Multi-modal Retrieval Augmentation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2579–2583.

[39] Peilin Zhou, Chao Liu, Jing Ren, Xinfeng Zhou, Yueqi Xie, Meng Cao, Zhongtao Rao, You-Liang Huang, Dading Chong, Junling Liu, Jae Boum Kim, Shoujin Wang, Raymond Chi-Wing Wong, and Sunghun Kim. 2025. When Large Vision Language Models Meet Multimodal Sequential Recommendation: An Empirical Study. In *Proceedings of the ACM on Web Conference 2025* (Sydney NSW, Australia) *(WWW '25)*. Association for Computing Machinery, New York, NY, USA, 275–292. https://doi.org/10.1145/3696410.3714764

[40] Tinghui Zhu, Kai Zhang, Muhao Chen, and Yu Su. 2025. Is Extending Modality The Right Path Towards Omni-Modality? *arXiv preprint arXiv:2506.01872* (2025).