A Content-Driven Micro-Video Recommendation Dataset at Scale

Yongxin Ni[1], **Yu Cheng**[1], **Xiangyan Liu**[1], **Junchen Fu**[1], **Youhua Li**[1],
Xiangnan He[2], **Yongfeng Zhang**[3], **Fajie Yuan**[1*]
[1]Westlake University
{niyongxin,chengyu,liuxiangyan,fujunchen,liyouhua,yuanfajie}@westlake.edu.cn
[2]University of Science and Technology of China
{xiangnanhe}@@gmail.com
[3]Rutgers University
{yongfeng.zhang}@rutgers.edu

Abstract

Micro-videos have recently gained immense popularity, sparking critical research in micro-video recommendation with significant implications for the entertainment, advertising, and e-commerce industries. However, the lack of large-scale public micro-video datasets poses a major challenge for developing effective recommender systems. To address this challenge, we introduce a very large micro-video recommendation dataset, named "*MicroLens*", consisting of one billion user-item interaction behaviors, 34 million users, and one million micro-videos. This dataset also contains various raw modality information about videos, including titles, cover images, audio, and full-length videos. MicroLens serves as a benchmark for content-driven micro-video recommendation, enabling researchers to utilize various modalities of video information for recommendation, rather than relying solely on item IDs or off-the-shelf video features extracted from a pre-trained network. Our benchmarking of multiple recommender models and video encoders on MicroLens has yielded valuable insights into the performance of micro-video recommendation. We believe that this dataset will not only benefit the recommender system community but also promote the development of the video understanding field. Our datasets and code are available at `https://github.com/westlake-repl/MicroLens`.

1 Introduction

Micro-videos, also known as short-form videos, have become increasingly popular in recent years. These videos typically range in length from a few seconds to several minutes and exist on various platforms, including social media, video-sharing websites, and mobile apps. Due to the brief yet captivating content, micro-videos have captured the attention of audiences worldwide, making them a powerful means of communication and entertainment. The surge in popularity of micro-videos has fueled critical research in micro-video recommender systems [66, 61, 16, 63, 13, 39, 18, 38, 30, 35].

However, the absence of large-scale public micro-video datasets containing diverse and high-quality video content, along with user behavior information, presents a significant challenge in developing reliable recommender systems (RS). Existing video recommendation datasets, such as MovieLens [19], mainly focus on longer movie-type videos and do not cover the wide range of content found in micro-videos, including but not limited to categories such as food, animals, sports, travel, education,

---

*Corresponding author. Author contributions: Fajie designed and supervised this research; Yongxin performed the research including key experiments; Chengyu, Junchen, Youhua, Xiangyan assisted a few important experiments; Xiangnan and Yongfeng provided guidance, participated in discussions, and proofread the paper; Fajie and Yongxin led the paper writing.

fashion, and music. Additionally, other datasets such as Tenrec [63] and KuaiRec [13] only contain video ID data or pre-extracted vision features from the video thumbnails, making it difficult to develop recommender models that can learn video representation directly from the raw video content data. Thus, there is an urgent need for large-scale micro-video recommendation datasets offering diverse raw content to facilitate the development of more accurate and effective recommender algorithms.

To address this challenge, we introduce a large-scale micro-video recommendation dataset, named "*MicroLens*", consisting of one billion user-item interaction behaviors, 34 million users, and one million micro-videos. Each micro-video is accompanied by original modalities, such as title, cover image, audio, and video information, providing a rich and diverse set of features for recommender models. Then, we perform benchmarking of various recommender baselines and cutting-edge video encoders on this dataset, providing valuable insights into the recommendation accuracy. We believe MicroLens can serve as a valuable resource for developing and evaluating content-driven video recommender models. To summarize, our contribution in this paper is three-fold:

- We introduce the largest and most diverse micro-video recommendation dataset, which provides access to raw video data. MicroLens encompasses all important modalities, including image, audio, text, and full-length video, making it an ideal resource for researchers working in various areas related to multimodal recommendation.

- We provide a comprehensive benchmark for over 10 recommender models and video encoders. Additionally, we introduce new types of baselines that use end-to-end (E2E) training to optimize both recommender models and video encoders. Although computationally expensive, these E2E models achieve superior performance that remains unknown in literature.

- Through empirical study, we present several crucial insights and explore the potential relationship between video understanding and recommender systems. Our findings indicate that a significant gap exists between current video understanding technologies and video recommendation, emphasizing the need for specialized research on video understanding technologies for video recommendation tasks.

2 MicroLens

2.1 Dataset Construction

Seed Video Collection. The data for MicroLens is sourced from an online micro-video platform with a focus on social entertainment. The recommendation scenario is described in Appendix Figure 7. The data collection process spanned almost a year, from June 2022 to June 2023. To begin with, we collected a large number of seed videos from the homepage. To ensure the diversity of videos, we frequently refreshed the homepage, allowing us to obtain a new set of videos every time. To ensure the quality of the collected videos, we filtered out unpopular content by only including videos with more than 10,000 likes in this stage. It is important to note that the platform does not directly provide user-video like and click interactions due to privacy protection. Instead, it provides user-video comment behaviors, which are publicly available and can serve as an implicit indicator of strong user preference towards a video. On average, there is approximately one comment for every 100 likes. That is, we mainly collected videos with positive interactions greater than 100 to ensure a reasonable level of engagement.[2] In total, we collect 400,000 micro-videos, including their video title, cover image, audio and raw video information.

Dataset Expansion. In this stage, we accessed the webpages of the videos collected in the previous stage. Each video page contains numerous links to external related videos, from which we randomly selected 10 video links. Note that (1) the related video links on each video page change with each visit; (2) the related videos have very diverse themes and are not necessarily of the same category as the main video. We collected approximately 5 million videos in this stage and retained the same metadata as in the previous stages.

Data Filtering. After collecting the videos, we conducted a data filtering process to remove a large number of duplicates and filter the data based on different modalities. For the text modality, we required that the length of the video titles, after removing meaningless characters, should not be less than 3. For the image modality, we used color uniformity checks and removed images with

---

[2]If the items are too cold in the platform, it is almost impossible to find enough overlapping users.
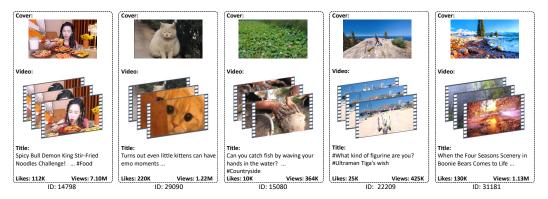
Figure 1: Dataset construction pipeline.

Figure 2: Item examples in MicroLens.

single-color areas greater than 75%. For the video modality, we set a threshold for file size and removed any videos with a file size of less than 100KB.

Overall, these filtering criteria helped to improve the quality of the collected data and ensured that only relevant and high-quality videos were included in the final dataset.

Interaction Collection. In this stage, we collected user-video interaction behaviors, primarily through the collection of comment data. We chose to collect comment data as a form of positive feedback from users for two primary reasons: (1) all user comment data on the platform is public, which eliminates potential privacy concerns that may arise with click and like data; and (2) unlike e-commerce scenarios where negative comments often indicate user dissatisfaction with the product, comments on short videos are typically about the people and events portrayed in the video, and both positive and negative comments can serve as indicators of user preferences towards the video. In fact, these preferences may be even stronger than those inferred from click behaviors.
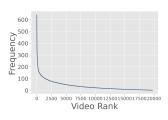
To collect comments, we accessed the webpage of each video and collected up to 5000 comments per video. This limitation was due to the fact that collecting more comments through pagination would require more time. In addition, we removed multiple comments from the same user to ensure data quality. Apart from comment data, we also recorded user IDs and comment timestamps. Although user and video IDs are public, we still anonymize them to avoid any privacy concerns.
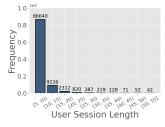
Data Integration. Due to the large volume of collected data, we employed a specialized data integration process. Our approach involved using a distributed large-scale download system, consisting of collection nodes, download nodes, and a data integration node. The technical details are in Appendix Section A. We provide the dataset construction process in Figure 1, and samples in Figure 2.

2.2 Privacy and Copyrights

MicroLens only includes public user behaviors for privacy protection. Both user and item IDs have been anonymized. To avoid copyright issues, we provide video URLs and a special tool to permanently access and download related videos. This is a common practice in prior literature [42, 65] when publishing multimedia datasets, e.g. YouTube8M[3] [1]. We will also provide the original dataset with reference to the ImageNet license, see `https://www.image-net.org/download.php`.

---

[3]Note that YouTube8M does not include user interaction data and therefore is not a recommendation dataset.

| (a) Item popularity. | (b) User session length. | (c) Video duration (in seconds) |

Figure 3: Statistics of MicroLens-100K.

Table 1: Data statistics of MicroLens. VAIT represents the video, audio, image and text data.

Dataset	#User	#Item	#Interaction	Sparsity	#Tags	Duration	VAIT
MicroLens-100K	100,000	19,738	719,405	99.96%	15,580	161s	✔
MicroLens-1M	1,000,000	91,402	9,095,620	99.99%	28,383	162s	✔
MicroLens	34,492,051	1,142,528	1,006,528,709	99.997%	258,367	138s	✔

2.3 Dataset Analysis

As the original MicroLens dataset is too large for most academic research, we have created two subsets of MicroLens by randomly selecting 100,000 users and 1 million users, named MicroLens-100K and MicroLens-1M, respectively. We consider MicroLens-100K as the default dataset to evaluate recommender models and provide some key results on MicroLens-1M in the Appendix.

Figure 3 illustrate some statistics of MicroLens-100K. (a) shows that item popularity aligns with the long-tail distribution which is commonly observed in most recommender systems. (b) indicates that users with interaction sequence lengths between 5 and 15 constitute the majority group. (c) depicts the distribution of video duration, with the majority of micro-videos less than 400 seconds in length.

We present the detailed statistical information of MicroLens-100K, MicroLens-1M, and the original MicroLens in Table 1. MicroLens-100K comprises 100 thousand users, 19,738 items, and 719,405 interactions, with the sparsity of 99.96%. MicroLens-1M includes 1 million users, 91,402 items, and 9,095,620 interactions, with the sparsity of 99.99%. The original MicroLens dataset consists of 34,492,051 users, 1,142,528 items, and 1,006,528,709 interactions. The three datasets, in ascending order, contain 15,580, 28,383, and 258,367 tags, respectively, with each tag representing a fine-grained category to which the videos belong. In addition to the raw multimodal information, we have also included additional features such as the number of views and likes per video, user gender information, and comment content.

2.4 Comparison to Existing Datasets

Over the past two decades, the field of RS has accumulated a large number of benchmark datasets. The most representative of these, MovieLens [19], has been extensively utilized for various recommendation tasks, particularly the rating prediction and top-N item recommendation tasks. Additionally, both academia and industry have released high-quality datasets, including Alibaba's various CTR prediction datasets [67, 69], Tencent's Tenrec dataset [63], and Kuaishou's KuaiRec and KuaiRand datasets [13, 14]. However, the majority of public RS datasets only offer user IDs, item IDs, and click behaviors, with relatively few public datasets providing multimodal information about the items. While KuaiRec, Flickr [57], and Behance [22] offer multimodal features, it is noteworthy that the image features are pre-extracted from vision encoders (e.g. ResNet [21]) without raw pixel features. Recently, Microsoft released the MIND dataset [56], which is the largest news recommendation dataset to date. Amazon [23] and POG [5] provided a large product purchase dataset that includes raw images of products. In addition, H&M[4], Yelp[5], and GEST [58] (Google restaurant) also released

---

[4]https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations
[5]https://www.yelp.com/dataset

several image datasets for business recommendation purposes. A list of related multimodal datasets in the recommender system community are shown in Appendix Section B.

However, to our best knowledge, there is currently no micro-video recommendation dataset that provides original video content. Reasoner [6] and MovieLens are two relevant datasets to MicroLens. However, the size of the Reasoner dataset is significantly smaller and only offers five frames of images for each video, whereas our MicroLens includes about 10,000 times more users and user-video behaviors. Although MovieLens provides the URLs of the movie trailer, it has a limited range of video categories, as it only includes videos of movie categories. Additionally, MovieLens was collected from a simulated user-movie rating website and does not accurately represent actual user watching behavior. For instance, in MovieLens, many users can rate over 10 movies in a matter of seconds, which does not reflect actual movie watching behaviors. As a result, despite its frequent use, it may not be ideal for certain types of research, e.g., sequential recommendation. Please refer to [55] for in-depth analysis of MovieLens.

3 Experiments Setup

Despite MicroLens's potential for multiple research areas, our primary focus here is on the micro-video recommendation task. As mentioned, we chose MicroLens-100K as our default dataset for basic evaluation, while additional results using MicroLens-1M are reported in Appendix.

3.1 Baselines and Evaluation

According to prior literature, video recommender models can be broadly categorized into two groups: those relying on pure item IDs (i.e., video content-agnostic) and those that incorporate pre-extracted video or visual features (from a frozen video encoder) along with item IDs. Among them, models based on pure item IDs (called IDRec) can typically be further divided into classical collaborative filtering (CF) models, such as DSSM [29], LightGCN [26], DeepFM [17] and NFM [25], and sequential recommendation (SR) models, such as GRU4Rec [27], NextItNet [62], and SASRec [31]. Regarding the latter models that utilize pre-extracted video features as side information, IDs continue to be the main features, except in very cold or new item recommendation scenarios. We simply call this approach VIDRec.[6] VIDRec can share a similar network architecture with IDRec, but with additional video features incorporated into the ID embeddings.

Beyond the above traditional baseline models, we also introduce a new family of recommender models called VideoRec. This model simply replaces the item ID in IDRec with a learnable video encoder. Unlike VIDRec, VideoRec uses end-to-end (E2E) training to optimize both the recommender model and the video encoder simultaneously. With the exception of the item representation module (ID embedding vs. video encoder), all other components of VideoRec and VIDRec are identical. Although VideoRec achieves the highest recommendation accuracy, it has not been studied in literature due to its high training costs.

In terms of training details, we exploit the in-batch softmax loss function [60] widely adopted in both academic literature and industrial systems. For evaluation, we utilized the leave-one-out strategy to split the datasets where the last item in the interaction sequence was used for evaluation, the item before the last was used for validation, and the remaining items were used for training. As nearly 95% of user behaviors involve less than 13 comments, we limited the maximum user sequence length to the most recent 13 for sequential models. We employ two popular rank metrics [62, 31], i.e., hit ratio (HR@N) and normalized discounted cumulative gain (NDCG@N). Here, N was set to 10 and 20.

3.2 Hyper-parameter Tuning

As IDRec is the most efficient baseline, we conducted a hyper-parameter search for IDRec as the first step. Specifically, we first extensively search two key hyper-parameters: the learning rate η from a set of values $\{1e-5, 5e-5, 1e-4, 5e-4, 1e-3\}$ and the embedding size from a set of values $\{64, 128, 256, 512, 1024, 2048, 4096\}$. Batch sizes b were also empirically tuned for

---

[6]In fact, research on VIDRec is relatively scarce compared to text and image-based recommendations. Even the most widely recognized model, the YouTube model, primarily relies on video ID and other categorical features, without explicitly leveraging the original video content features.

Table 2: Benchmark results on MicroLens-100K. VideoMAE and SlowFast are used as video encoder for VIDRec and VideoRec, respectively (see Footnote[8]). The fusion of video and ID embedding features can be achieved through either summation or concatenation, which shows similar results.

Class	Model	HR@10	NDCG@10	HR@20	NDCG@20
IDRec (CF)	DSSM [29]	0.0394	0.0193	0.0654	0.0258
	LightGCN [26]	0.0372	0.0177	0.0618	0.0239
	NFM [25]	0.0313	0.0159	0.0480	0.0201
	DeepFM [17]	0.0350	0.0170	0.0571	0.0225
IDRec (SR)	NexItNet [62]	0.0805	0.0442	0.1175	0.0535
	GRU4Rec [27]	0.0782	0.0423	0.1147	0.0515
	SASRec [31]	0.0909	0.0517	0.1278	0.0610
VIDRec (Frozen Encoder)	YouTube_{ID}	0.0461	0.0229	0.0747	0.0301
	YouTube_{ID+V} [7]	0.0392	0.0188	0.0648	0.0252
	MMGCN_{ID}	0.0141	0.0065	0.0247	0.0092
	MMGCN_{ID+V} [54]	0.0214	0.0103	0.0374	0.0143
	GRCN_{ID}	0.0282	0.0131	0.0497	0.0185
	GRCN_{ID+V} [53]	0.0306	0.0144	0.0547	0.0204
	DSSM_{ID+V}	0.0279	0.0137	0.0461	0.0183
	SASRec_{ID+V}	0.0799	0.0415	0.1217	0.0520
VideoRec (E2E Learning)	NexItNet_V [62]	0.0862	0.0466	0.1246	0.0562
	GRU4Rec_V [27]	0.0954	0.0517	0.1377	0.0623
	SASRec_V [31]	0.0948	0.0515	0.1364	0.0619

individual models from a set of values $\{64, 128, 256, 512, 1024, 2048\}$. With regards to VIDRec and VideoRec, we first applied the same set of hyper-parameters obtained from IDRec and then performed some basic searches around these optimal values. It is worth mentioning that extensively tuning VideoRec is not feasible in practice, as it requires at least 10-50 times more compute and training time than IDRec.[7] Due to the very high computational cost involved, we only optimized the top few layers for the video encoder network in VideoRec. Along with other hyper-parameters, such as the layer numbers of NexItNet, GRU4Rec, and SASRec, we report them in Appendix Section C. In addition, for VIDRec and VideoRec, we follow the common practice (e.g., in Video Swin Transformer [41]) by selecting a consecutive sequence of five frames from the midsection per video, with a frame interval of 1, to serve as the video input.

4 Experimental Results

The lack of high-quality video datasets has limited research on the effective utilization of raw video content in recommender systems. Here, we provide preliminary exploration with the aim of drawing the community's attention and inspiring more research on content-driven video recommendation.

4.1 Benchmark Results & Analysis

We evaluate multiple recommender baselines on MicroLens, including IDRec (which does not use video features), VIDRec (which incorporates video features as side information), & VideoRec (which uses video features exclusively).[8] The results are reported in Table 2 with the below findings.

Firstly, regarding IDRec, all sequential models, including SASRec, NexItNet, and GRU4Rec, outperform non-sequential CF models, namely DSSM, LightGCN, DeepFM, NFM and YouTube. Among all models, SASRec with Transformer backbone performs the best, improving CNN-based NexItNet and RNN-based GRU4Rec by over 10%. The findings are consistent with much prior literature [68, 31, 50, 49].

---

[7]A recent study [59] proposed a highly promising solution to improve training efficiency, which appears to be feasible for VideoRec.

[8]Regarding VIDRec, we extracted video features from VideoMAE [45], which demonstrates state-of-the-art (SOTA) accuracy for multiple video understanding tasks (e.g., action classification). For VideoRec, we utilized the SlowFast video network [11], which offers the best accuracy through E2E learning. Extensive results on more video encoders are reported Figure 4.
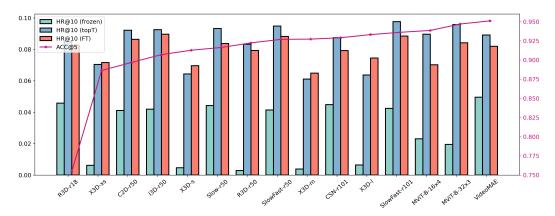
Figure 4: Video recommendation accuracy (bar charts) vs. video classification accuracy (purple line). Frozen means that the video encoder is fixed without parameter update, topT means that only the top few layers of the video encoder are fine-tuned, and FT means full parameters are fine-tuned.

Secondly, surprisingly, incorporating pre-extracted video features in VIDRec (i.e., GRCN$_{\text{ID+V}}$, MMGCN$_{\text{ID+V}}$, YouTube$_{\text{ID+V}}$, DSSM$_{\text{ID+V}}$ and SASRec$_{\text{ID+V}}$) does not necessarily result in better performance compared to their IDRec counterparts (e.g., YouTube$_{\text{ID}}$, DSSM$_{\text{ID}}$, and SASRec$_{\text{ID}}$). In fact, VIDRec, which treats video or visual features as side information, is mostly used to assist cold or new item recommendation where pure IDRec is weak due to inadequate training [24, 48, 34]. However, **for non-cold or warm item recommendation, such side information may not always improve performance, as ID embeddings may implicitly learn these features.** Similar findings were also reported in [64], which demonstrated that incorporating visual features leads to a decrease in accuracy for non-cold item recommendation. These results imply that the common practice of using pre-extracted features from a frozen video encoder may not always yield the expected improvements in performance.

Thirdly, the recent study [64] suggested that the optimal way to utilize multimodal features is through E2E training of the recommender model and the item (i.e., video in this case) modality encoder. Similarly, we observe that VideoRec (i.e., NextItNet$_{\text{V}}$, GRU4Rec$_{\text{V}}$, and SASRec$_{\text{V}}$) achieves the highest recommendation accuracy among all models. In particular, NextItNet$_{\text{V}}$ largely outperforms NextItNet (i.e., NextItNet$_{\text{ID}}$), GRU4Rec$_{\text{V}}$ largely outperforms GRU4Rec. The comparison clearly demonstrates that learning item representation from raw video data through end-to-end (E2E) training of the video encoder, as opposed to utilizing pre-extracted offline features in VIDRec or pure ID features, leads to superior results (see more analysis in Section 4.2). This is likely because E2E training can incorporate both raw video features and collaborative signals from user-item interactions.

Our above findings suggest that **utilizing raw video features instead of pre-extracted frozen features is crucial for achieving optimal recommendation results, underscoring the significance of the MicroLens video dataset**.

4.2 Video Understanding Meets Recommender Systems

In the CV field, numerous video networks have been developed. Here, we aim to explore whether these networks designed for video understanding helps the recommendation task and to what extent.

Given its top performance in Table 2, we employed SASRec as the recommender backbone and evaluated 15 well-known video encoders that were pre-trained on Kinetics [32], a well-known video (action) classification dataset. These encoders include R3D-r18 [47], X3D-xs [10], C2D-r50 [52], I3D-r50 [4], X3D-s [10], Slow-r50 [8], X3D-m [10], R3D-r50 [47], SlowFast-r50 [11], CSN-r101 [46], X3D-l [10], SlowFast-r101 [11], MViT-B-16x4 [9], MViT-B-32x3 [9], and VideoMAE [45] with details in Appendix Section D.

Q1: Can the knowledge learned from video understanding be beneficial for video recommendation?

Figure 5 (a) shows that the recommender model SASRec$_{\text{V}}$ with pre-trained SlowFast-r50, SlowFast-r101, and MVIT-B-32x3 video encoders exhibited a solid improvement in performance compared

(a) OT v.s. WT   (b) SlowFast-r50   (c) SlowFast-r101   (d) MViT-B-32x3
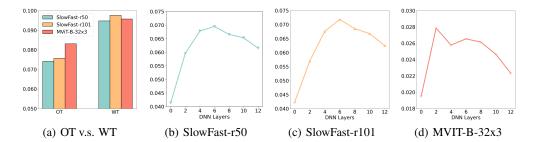
Figure 5: Ablation study of video encoders. (d) "WT" refers to the video encoders in $SASRec_V$ have pre-trained weights from the video classification task, while "OT" denotes that they are randomly initialized. (b) (c) (d) are performance change by adding DNN layers on top of three frozen encoders.

to their random initialization versions. These results clearly suggest that **the parameters learned from the video understanding task in the CV field are highly valuable for improving video recommendations.**

Q2: Does a strong video encoder always translate into a better video recommender model?

Figure 4 compares the video classification (VC) task and the video recommendation task. It is evident that **a higher performance in the VC task (purple line) does not necessarily correspond to a higher accuracy in video recommendation (bar charts).** For instance, while VideoMAE achieves optimal results in video classification, it does not necessarily guarantee the highest accuracy for item recommendation. A more pronounced example is R3D-r18, which exhibits the worst results in video classification but performs relatively well in the recommendation task. This finding differs from [64], which demonstrated that higher performance in NLP and CV models generally leads to higher recommendation accuracy. However, it should be noted that [64] only investigated image and text recommendation, and did not explore video recommendation, which could be very different.

Q3: Are the semantic representations learned from the video understanding task universal for video recommendation

In Section 4.1, we showed that incorporating pre-extracted video features may not necessarily improve recommendation accuracy when ID features are sufficiently trained. Here, we conducted a more detailed study by comparing the performance of recommender models (i.e., SASRec) with frozen (equivalent to pre-extracted video features) and end-to-end trained video encoders. Figure 4 clearly demonstrates that $SASRec_V$ with retrained video encoders, whether topT or FT, performs significantly better, with about a 2-fold improvement over the frozen approach. These results suggest that **the video semantic representations learned by the popular video classification task are not universal to the recommendation task, and retraining the video encoder on the recommendation data is necessary to achieve optimal performance.** This is because, if the pre-extracted video features were a perfect representation, a linear layer applied to these features is enough to perform equally well as the fine-tuned video encoder. Although adding more DNN layers on the pre-extracted video features significantly improves accuracy (see Figure 5(b,c,d)), it still largely falls short of the accuracy achieved by using a fine-tuned video encoder. Moreover, the results indicate that full parameter fine-tuning (FT) of the video encoder is not necessary, as fine-tuning only the top few layers (TopT) generally produces superior results. This seems reasonable since optimizing all parameters of the video encoder may result in complete catastrophic forgetting of the knowledge learned during the video pre-training task. This highlights once again the value of the knowledge gained from video understanding tasks for video recommendation.

To sum up, existing video understanding technologies, including video encoders and trained parameters, are undoubtedly valuable for video recommendation. However, there is still a significant semantic gap between video understanding tasks and recommendation systems. Therefore, not all advances made in video tasks can directly translate into improvements for recommender systems.

4.3 Additional Exploration of VideoRec

Beyond the above results, we have performed other interesting empirical experiments as below.

Table 3: Recommendation accuracy using cover images to represent videos, with three SOTA image encoders, i.e., ResNet [21], Swin Transformer [40] and MAE [20] (see Appendix Section 7 for details).

Model	MicroLens-100K			
	HR@10	NDCG@10	HR@20	NDCG@20
$SASRec_{ResNet}$	0.0858	0.0462	0.1264	0.0564
$SASRec_{MAE}$	0.0828	0.0447	0.1223	0.0546
$SASRec_{Swin}$	0.0892	0.0479	0.1299	0.0582

Q1: How would the recommendation performance be impacted if we solely rely on the cover image instead of the raw video?

To answer this question, we use three SOTA image encoders to represent video cover images. We still use the E2E learning and refer to this approach as ImageRec. Our results are given in Table 3, which suggests that VideoRec generally outperforms ImageRec when compared to the results of $SASRec_V$ in Table 2 and Figure 4. This also reflects the importance of video content for recommender systems.

Q2: Can VideoRec compete with IDRec in recommending highly popular items?

In Section 4.1, we showed that VideoRec is capable of surpassing IDRec in the regular item recommendation setting (including both popular and cold items). Here, we want to further investigate whether VideoRec still outperforms IDRec in recommending popular items. The reason why we are keen in comparing IDRec is that many recent studies [64, 37, 51, 36, 28, 12] have claimed that IDRec poses a major obstacle for *transferable* or *foundation* recommender models [15] as ID features are generally non-shareable in practice. Appendix Table 8 shows that VideoRec using the SOTA SASRec architecture can consistently outperform IDRec, even in very warm item settings.

To our best knowledge, **this study is the first to show that raw video features can potentially replace ID features in both *warm* and cold[9] item recommendation settings.** We consider this to be a significant contribution as it suggests that VideoRec may potentially challenge the dominant role of ID-based recommender systems. This is particularly noteworthy given VideoRec's natural advantage in transfer learning due to the generality of video or visual features. That is, VideoRec has taken a key step towards the grand goal of a universal "one-for-all" recommender paradigm.

At last, we have reported some key baseline results in MicroLens-1M in Appendix Table 9 and 10.

5   Conclusions and Broader Impact

This paper introduces "*MicroLens*", the most immense and diverse micro-video dataset to date. Each video in MicroLens contains rich modalities, including text descriptions, images, audio, and raw video information. We conduct an extensive empirical study and benchmark multiple classical recommender baselines. The newly proposed method, VideoRec, directly learns item representations from raw video features and achieves the highest recommendation accuracy among the compared models. We anticipate that MicroLens will become a valuable resource for the recommender system community, enabling multiple research directions in multimodal or micro-video recommendation.

Although MicroLens is primarily used for video recommendation tasks in this paper, there are other important research directions worth exploring. For instance, recent advances in foundation one-for-all models, such as ChatGPT [44] and GPT-4 [43], have achieved remarkable success in the fields of NLP and CV. However, the recommender system community has made limited progress in large foundation models, particularly in vision- or video-content driven recommender systems. This is partly due to the lack of large-scale, diverse, and high-quality multimodal recommendation datasets, which presents a significant challenge. We envision that MicroLens may serve as a valuable pre-training dataset for visually relevant recommendation, as a single micro-video in MicroLens can generate hundreds of high-quality images, resulting in a trillion-level of user-image interactions.

---

[9]More improvements can be easily observed on cold items (see Appendix Figure 6), which was also studied in much prior literature [33, 34, 64].

Moreover, the field of video understanding has recently made significant strides and is poised to become a future research hotspot [3, 2, 45, 47]. Using video understanding to drive more fine-grained recommendation, rather than simply learning user behavior similarities, is undoubtedly a more promising direction. Additionally, treating video recommendation as a downstream task for video understanding has the potential to unite the two communities and foster mutual development.

References

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

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[5] Wen Chen, Pipei Huang, Jiaming Xu, Xin Guo, Cheng Guo, Fei Sun, Chao Li, Andreas Pfadler, Huan Zhao, and Binqiang Zhao. Pog: personalized outfit generation for fashion recommendation at alibaba ifashion. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2662–2670, 2019.

[6] Xu Chen, Jingsen Zhang, Lei Wang, Quanyu Dai, Zhenhua Dong, Ruiming Tang, Rui Zhang, Li Chen, and Ji-Rong Wen. Reasoner: An explainable recommendation dataset with multi-aspect real user labeled ground truths towards more measurable explainable recommendation. *arXiv preprint arXiv:2303.00168*, 2023.

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

[8] Haoqi Fan, Tullie Murrell, Heng Wang, Kalyan Vasudev Alwala, Yanghao Li, Yilei Li, Bo Xiong, Nikhila Ravi, Meng Li, Haichuan Yang, et al. Pytorchvideo: A deep learning library for video understanding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3783–3786, 2021.

[9] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.

[10] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020.

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[12] Junchen Fu, Fajie Yuan, Yu Song, Zheng Yuan, Mingyue Cheng, Shenghui Cheng, Jiaqi Zhang, Jie Wang, and Yunzhu Pan. Exploring adapter-based transfer learning for recommender systems: Empirical studies and practical insights. *arXiv preprint arXiv:2305.15036*, 2023.

[13] Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. Kuairec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 540–550, 2022.

[14] Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. Kuairand: An unbiased sequential recommendation dataset with randomly exposed videos. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, CIKM '22, page 3953–3957, 2022.

[15] Shijie Geng, Juntao Tan, Shuchang Liu, Zuohui Fu, and Yongfeng Zhang. Vip5: Towards multimodal foundation models for recommendation. *arXiv preprint arXiv:2305.14302*, 2023.

[16] Xudong Gong, Qinlin Feng, Yuan Zhang, Jiangling Qin, Weijie Ding, Biao Li, Peng Jiang, and Kun Gai. Real-time short video recommendation on mobile devices. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3103–3112, 2022.

[17] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. Deepfm: a factorization-machine based neural network for ctr prediction. *arXiv preprint arXiv:1703.04247*, 2017.

[18] Tingting Han, Hongxun Yao, Chenliang Xu, Xiaoshuai Sun, Yanhao Zhang, and Jason J Corso. Dancelets mining for video recommendation based on dance styles. *IEEE Transactions on Multimedia*, 19(4):712–724, 2016.

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

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. Vista: A visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th ACM conference on recommender systems*, pages 309–316, 2016.

[23] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517, 2016.

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

[25] Xiangnan He and Tat-Seng Chua. Neural factorization machines for sparse predictive analytics. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 355–364, 2017.

[26] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 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*, pages 639–648, 2020.

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

[28] Yupeng Hou, Zhankui He, Julian McAuley, and Wayne Xin Zhao. Learning vector-quantized item representation for transferable sequential recommenders. In *Proceedings of the ACM Web Conference 2023*, pages 1162–1171, 2023.

[29] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338, 2013.

[30] Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. 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*, pages 3487–3495, 2020.

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

[32] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijaya-narasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[33] Yaman Kumar, Agniv Sharma, Abhigyan Khaund, Akash Kumar, Ponnurangam Kumaraguru, Rajiv Ratn Shah, and Roger Zimmermann. Icebreaker: Solving cold start problem for video recommendation engines. In *2018 IEEE international symposium on multimedia (ISM)*, pages 217–222. IEEE, 2018.

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

[35] Chenyi Lei, Yong Liu, Lingzi Zhang, Guoxin Wang, Haihong Tang, Houqiang Li, and Chunyan Miao. 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*, pages 3161–3171, 2021.

[36] Jiacheng Li, Ming Wang, Jin Li, Jinmiao Fu, Xin Shen, Jingbo Shang, and Julian McAuley. Text is all you need: Learning language representations for sequential recommendation. *arXiv preprint arXiv:2305.13731*, 2023.

[37] Ruyu Li, Wenhao Deng, Yu Cheng, Zheng Yuan, Jiaqi Zhang, and Fajie Yuan. Exploring the upper limits of text-based collaborative filtering using large language models: Discoveries and insights. *arXiv preprint arXiv:2305.11700*, 2023.

[38] Shang Liu, Zhenzhong Chen, Hongyi Liu, and Xinghai Hu. User-video co-attention network for personalized micro-video recommendation. In *The World Wide Web Conference*, pages 3020–3026, 2019.

[39] Yiyu Liu, Qian Liu, Yu Tian, Changping Wang, Yanan Niu, Yang Song, and Chenliang Li. Concept-aware denoising graph neural network for micro-video recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1099–1108, 2021.

[40] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[41] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[42] Dan S Nielsen and Ryan McConville. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3141–3153, 2022.

[43] OpenAI. Gpt-4 technical report, 2023.

[44] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[45] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

[46] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019.

[47] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[48] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. *Advances in neural information processing systems*, 26, 2013.

[49] Chenyang Wang, Zhefan Wang, Yankai Liu, Yang Ge, Weizhi Ma, Min Zhang, Yiqun Liu, Junlan Feng, Chao Deng, and Shaoping Ma. Target interest distillation for multi-interest recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2007–2016, 2022.

[50] Jiachun Wang, Fajie Yuan, Jian Chen, Qingyao Wu, Min Yang, Yang Sun, and Guoxiao Zhang. Stackrec: Efficient training of very deep sequential recommender models by iterative stacking. In *Proceedings of the 44th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 357–366, 2021.

[51] Jie Wang, Fajie Yuan, Mingyue Cheng, Joemon M Jose, Chenyun Yu, Beibei Kong, Zhijin Wang, Bo Hu, and Zang Li. Transrec: Learning transferable recommendation from mixture-of-modality feedback. *arXiv preprint arXiv:2206.06190*, 2022.

[52] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[53] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. Graph-refined convolutional network for multimedia recommendation with implicit feedback. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3541–3549, 2020.

[54] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445, 2019.

[55] Daniel Woolridge, Sean Wilner, and Madeleine Glick. Sequence or pseudo-sequence? an analysis of sequential recommendation datasets. In *Perspectives@ RecSys*, 2021.

[56] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, 2020.

[57] Le Wu, Lei Chen, Richang Hong, Yanjie Fu, Xing Xie, and Meng Wang. A hierarchical attention model for social contextual image recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 32(10):1854–1867, 2019.

[58] An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, and Julian McAuley. Personalized showcases: Generating multi-modal explanations for recommendations. *arXiv preprint arXiv:2207.00422*, 2022.

[59] Yoonseok Yang, Kyu Seok Kim, Minsam Kim, and Juneyoung Park. Gram: Fast fine-tuning of pre-trained language models for content-based collaborative filtering. *arXiv preprint arXiv:2204.04179*, 2022.

[60] Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 269–277, 2019.

[61] Yisong Yu, Beihong Jin, Jiageng Song, Beibei Li, Yiyuan Zheng, and Wei Zhuo. 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*, pages 508–523. Springer, 2023.

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

[63] Guanghu Yuan, Fajie Yuan, Yudong Li, Beibei Kong, Shujie Li, Lei Chen, Min Yang, Chenyun Yu, Bo Hu, Zang Li, et al. Tenrec: A large-scale multipurpose benchmark dataset for recommender systems. *arXiv preprint arXiv:2210.10629*, 2022.

[64] Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. Where to go next for recommender systems? id-vs. modality-based recommender models revisited. *arXiv preprint arXiv:2303.13835*, 2023.

[65] Zhaoyang Zeng, Yongsheng Luo, Zhenhua Liu, Fengyun Rao, Dian Li, Weidong Guo, and Zhen Wen. Tencent-mvse: A large-scale benchmark dataset for multi-modal video similarity evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3138–3147, 2022.

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

[67] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1059–1068, 2018.

[68] Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1893–1902, 2020.

[69] Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. Learning tree-based deep model for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1079–1088, 2018.

A   Technical Details for Data Integration

When a collection node obtained download links, it was responsible for distributing these links to the download nodes. Multiple download nodes were utilized, each equipped with large-scale storage and high-speed broadband. The download nodes were able to communicate and collaborate with each other to ensure efficient and non-redundant downloading of images, audio, and video files. Upon the completion of the downloading process, multiple high-speed transfer channels were established between the download nodes and the data integration node. This allowed for the merging of all downloaded files into a single root directory. The data integration node utilized a high-capacity hard drive to store all the downloaded data. The data integration process allowed us to effectively manage and organize a large amount of collected data, enabling us to analyze and extract valuable insights from the data.

Overall, our data integration process allowed us to effectively manage and organize a large amount of collected data, enabling us to analyze and extract valuable insights from the data.

B   Related Datasets

Table 4: Dataset comparison. "p-Image" refers to pre-extracted visual features from pre-trained visual encoders (such as ResNet), while "r-Image" refers to images with raw image pixels. "Audio and Video" means the original full-length audio and video content.

Dataset	Modality					Scale			Domain	Language
	Text	p-Image	r-Image	Audio	Video	#user	#item	#inter.		
Tenrec	✗	✗	✗	✗	✗	6.41M	4.11M	190.48M	News & Videos	✗
UserBehavior	✗	✗	✗	✗	✗	988K	4.16M	100.15M	E-commerce	✗
Alibaba CTR	✗	✗	✗	✗	✗	7.96M	66K	15M	E-commerce	✗
Amazon	✓	–	✓	✗	✗	20.98M	9.35M	82.83M	E-commerce	en
POG	✓	–	✓	✗	✗	3.57M	1.01M	0.28B	E-commerce	zh
MIND	✓	✗	✗	✗	✗	1.00M	161K	24.16M	News	en
H&M	✓	–	✓	✗	✗	1.37M	106K	31.79M	E-commerce	en
BeerAdvocate	✓	✗	✗	✗	✗	33K	66K	1.59M	E-commerce	en
RateBeer	✓	✗	✗	✗	✗	40K	110K	2.92M	E-commerce	en
Google Local	✓	✗	✗	✗	✗	113.64M	4.96M	666.32M	E-commerce	en
Flickr	✗	✓	✗	✗	✗	8K	105K	5.90M	Social Media	en
Pinterest	✗	–	✓	✗	✗	46K	880K	2.56M	Social Media	✗
WikiMedia	✗	–	✓	✗	✗	1K	10K	1.77M	Social Media	✗
Yelp	✗	–	✓	✗	✗	150K	200K	6.99M	E-commerce	✗
GEST	✓	–	✓	✗	✗	1.01M	4.43M	1.77M	E-commerce	en
Behance	✗	✓	✗	✗	✗	63K	179K	1.00M	Social Media	✗
KuaiRand	✗	✗	✗	✗	✗	27K	32.03M	322.28M	Micro-video	✗
KuaiRec	✗	✓	✗	✗	✗	7K	11K	12.53M	Micro-video	✗
ML25M	✓	–	✓	✗	✗	162K	62K	25.00M	Movie-only	en
Reasoner	✓	–	✓	✗	✗	3K	5K	58K	Micro-video	en
MicroLens	✓	–	✓	✓	✓	30M	1M	1B	Micro-video	zh/en

C   Hyper-parameter Settings for Baselines

We report some essential hyperparameters of Baselines in Table 5. The "finetuning rate" denotes the learning rate applied to the video encoder during the finetuning process.

D   Video Model Details in Video Understanding and Recommendation

E   Details of the Applied Image Encoders

We showed details of three classical image encoders in Table 7.

Table 5: Hyper-parameters settings for baselines.

Class	Model	Learning Rate	Embedding Size	Batch Size	Dropout Rate	Weight Decay	Block Number	Finetune Top Blocks	Finetuning Rate
IDRec	DSSM	1e-5	4096	64	0	0.1	-	-	-
	LightGCN	1e-3	1024	1024	0	0	-	-	-
	NFM	5e-5	1024	64	0	0.01	-	-	-
	DeepFM	1e-4	512	64	0	0.1	-	-	-
	NexItNet	1e-3	2048	64	0.1	0.1	2 (CNN Block)	-	-
	GRU4Rec	1e-4	2048	512	0.1	0.1	1 (GRU Block)	-	-
	SASRec	1e-5	2048	512	0.1	0.1	2 (Transformer Block)	-	-
VIDRec	Youtube$_{ID}$	1e-4	4096	512	0.1	0.1		-	-
	Youtube$_{ID+V}$	1e-4	4096	512	0.1	0.1		-	-
	MMGCN$_{ID}$	1e-4	4096	64	0.1	0.0	-	-	-
	MMGCN$_{ID+V}$	1e-4	4096	64	0.1	0.0	-	-	-
	GRCN$_{ID}$	1e-4	4096	64	0.1	0.0	-	-	-
	GRCN$_{ID+V}$	1e-4	4096	64	0.1	0.0	-	-	-
	DSSM$_{ID+V}$	1e-3	4096	1024	0	0.1	-	-	-
	SASRec$_{ID+V}$	1e-5	2048	64	0.1	0.1	-	-	-
VideoRec	NexItNet$_V$	1e-4	512	120	0.1	0.1	2 (CNN Block)	1	1e-4
	GRU4Rec$_V$	1e-4	512	120	0.1	0.1	1 (GRU Block)	1	1e-4
	SASRec$_V$	1e-4	512	120	0.1	0.1	2 (Transformer Block)	1	1e-4

Table 6: Performance of VideoRec with 15 video encoders. "Pretrain Settings" are the adopted frame length and sample rate from the pre-trained checkpoint. ACC@5 is the accuracy in the video classification task.

Model	Architecture	Depth	Pretrain Settings	ACC@5	HR@10 (frozen)	NDCG@10 (frozen)	HR@10 (topT)	NDCG@10 (topT)	HR@10 (FT)	NDCG@10 (FT)
R3D-r18 [47]	ResNet	R18	16x4	75.45	4.58	2.56	8.50	4.48	7.50	3.48
X3D-xs [10]	Xception	XS	4x12	88.63	0.62	0.33	7.04	3.57	6.04	2.57
C2D-r50 [52]	ResNet	R50	8x8	89.68	4.11	2.27	9.22	4.88	8.22	3.88
I3D-r50 [4]	ResNet	R50	8x8	90.70	4.19	2.36	9.25	5.01	8.25	4.01
X3D-s [10]	Xception	S	13x6	91.27	0.47	0.24	6.43	3.25	5.43	2.25
Slow-r50 [8]	ResNet	R50	8x8	91.63	4.42	2.42	9.32	4.99	8.33	3.99
X3D-m [10]	Xception	M	16x5	92.72	0.38	0.20	6.11	3.13	5.11	2.13
R3D-r50 [47]	ResNet	R50	16x4	92.23	0.28	0.14	8.33	4.34	7.33	3.34
SlowFast-r50 [11]	ResNet	R50	8x8	92.69	4.14	2.35	9.48	5.15	8.48	4.15
CSN-r101 [46]	ResNet	R101	32x2	92.90	4.48	2.52	8.74	4.71	7.74	3.71
X3D-l [10]	Xception	L	16x5	93.31	0.64	0.34	6.37	3.32	5.37	2.32
SlowFast-r101 [11]	ResNet	R101	16x8	93.61	4.25	2.36	**9.76**	**5.3**	**8.76**	**4.31**
MViT-B-16x4 [9]	VIT	B	16x4	93.85	2.30	1.33	8.96	4.79	7.96	3.79
MViT-B-32x3 [9]	VIT	B	32x3	94.69	1.95	1.11	9.57	5.11	8.57	4.11
VideoMAE [45]	Transformer	VIT-B	16x4	**95.10**	**4.96**	**2.76**	8.91	4.77	7.91	3.77

Table 7: Network architecture, parameter size, and download URL of the vision encoders for image baselines. L: number of Transformer blocks, H: number of multi-head attention, C: channel number of the hidden layers in the first stage, B: number of layers in each block.

Image encoder	Architecture	#Param.	URL
ResNet18	C = 64, B={2, 2, 2, 2}	12M	https://download.pytorch.org/models/resnet18-5c106cde.pth
Swin-T	C = 96, B={2, 2, 6, 2}	28M	https://huggingface.co/microsoft/swin-tiny-patch4-window7-224
MAE$_{base}$	L=12, H=768	86M	https://huggingface.co/facebook/vit-mae-base

Table 8: Comparison of VideoRec and IDRec in regular and warm settings using SASRec as the backbone. "Warm-20" denotes that items with less than 20 interactions were removed from the original MicroLens-100K.

Model	Regular		Warm-20		Warm-50		Warm-200	
	H@10	N@10	H@10	N@10	H@10	N@10	H@10	N@10
IDRec	0.0909	0.0517	0.1068	0.0615	0.6546	0.4103	0.7537	0.4412
SlowFast-r101	0.0976	0.0531	0.1130	0.0606	0.7458	0.4463	0.8482	0.4743
MViT-B-32x3	0.0957	0.0511	0.1178	0.0639	0.7464	0.4530	0.9194	0.4901
SlowFast-r50	0.0948	0.0515	0.1169	0.0642	0.7580	0.4614	0.8141	0.4870

F Warm-up Recommendation on MicroLens-100K

G Baseline Evaluation and Warm-up Recommendation on MicroLens-1M

We reported the results of baseline evaluation and warm-up recommendation on MicroLens-1M in Table 9 and Table 10, respectively. Please note that due to excessive GPU memory consumption, some baselines could not be trained on MicroLens-1M, and we do not report their results. In general, we observed that the trends on MicroLens-1M (in terms of both baseline evaluation and warm-up recommendation) are consistent with that observed on MicroLens-100K.

Table 9: Benchmark results on MicroLens-1M.

Class	Model	HR@10	NDCG@10	HR@20	NDCG@20
IDRec (CF)	DSSM	0.0133	0.0065	0.0225	0.0087
	LightGCN	0.0150	0.0072	0.0253	0.0098
IDRec (SR)	NexItNet	0.0389	0.0209	0.0584	0.0258
	GRU4Rec	0.0444	0.0234	0.0683	0.0294
	SASRec	0.0476	0.0255	0.0710	0.0314
VIDRec (Frozen Encoder)	YouTube$_{ID}$	0.0256	0.0129	0.0578	0.0246
	YouTube$_{ID+V}$	0.0180	0.0089	0.0303	0.0119
VideoRec (E2E Learning)	NexItNet$_V$	0.0521	0.0272	0.0792	0.0340
	GRU4Rec$_V$	0.0510	0.0264	0.0782	0.0332
	SASRec$_V$	0.0582	0.0309	0.0871	0.0382

Table 10: Comparison of VideoRec and IDRec in regular and warm-start settings using SASRec as the user backbone. Warm-20 denotes that items with less than 20 interactions were removed from the original MicroLens-1M.

Model	Regular		Warm-20		Warm-50		Warm-200	
	H@10	N@10	H@10	N@10	H@10	N@10	H@10	N@10
IDRec	0.0476	0.0255	0.0508	0.0272	0.0562	0.0306	0.5533	0.3105
SlowFast-r101	0.0554	0.0291	0.0574	0.0303	0.0603	0.0318	0.5766	0.3107
MViT-B-32x3	0.0569	0.0300	0.0562	0.0294	0.0608	0.0318	0.6046	0.3399
SlowFast-r50	0.0582	0.0309	0.0603	0.0319	0.0638	0.0339	0.6217	0.3556

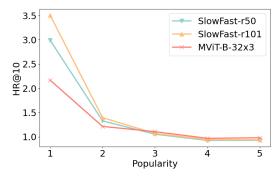# H Recommendation in Cold-start Scenarios

Figure 6: Results in different cold-start scenarios, with the y-axis representing the relative improvement of HR@10, calculated as the ratio of VideoRec to IDRec. The x-axis represents item groups divided by popularity level, the larger number indicates that items in the group are more popular.

Table 11: Recommendation results with side features on MicroLens-100K.

Model	HR@10	NDCG@10	HR@20	NDCG@20
SASRec$_{\text{ID}}$	0.0909	0.0517	0.1278	0.0610
SASRec$_{\text{ID+Pop}}$	0.0709	0.0396	0.1037	0.0479
SASRec$_{\text{ID+Tag}}$	0.0908	0.0499	0.1320	0.0603
SASRec$_{\text{ID+Pop+Tag}}$	0.0778	0.0423	0.1138	0.0513

I Recommendation with Side Features

In this section, we investigate the impact of other features on recommendation performance using MicroLens-100K dataset. We introduce two types of side features: item popularity level (Pop) and tag categories (Tag). For popularity features, we divide the item popularity into 10 uniform bins. The first bin represents the top 10% of popular items, while the last bin represents the bottom 10%. We assign a Pop ID to each item according to its popularity level. Regarding the tag features, we also handle them as categorical features with a category of $15, 580$.

We conducted experiments on SASRec$_{\text{ID}}$ (ID) with different feature combinations: SASRec$_{\text{ID}}$, SASRec$_{\text{ID+Pop}}$, SASRec$_{\text{ID+Tag}}$, and SASRec$_{\text{ID+Pop+Tag}}$. The "+" symbol denotes feature combination achieved by summing and averaging them. We report the results in Table 11.

We found that incorporating item popularity level and tag categories as side features did not clearly improve the algorithm's performance. One possible reason is that in typical recommendation scenarios, item ID embeddings have already been extensively trained, implicitly learning latent factors including similarity and popularity. For instance, we observed that many videos recommended in the top-10 recommendation list share similar categories and have relatively high popularity, indicating that ID-based methods can already capture popularity and category information. In such scenarios, incorporating many unimportant features may have a negative impact on overall performance. It is worth noting that in the very cold-start setting, the item ID feature is very weak and adding other features is necessary for better performance.

J Comparison between Textual Features and Video Content

Table 12: Comparsion results of ID, textual features and video content on MicroLens-100K.

Model	HR@10	NDCG@10	HR@20	NDCG@20
SASRec$_{\text{ID}}$	0.0909	0.0517	0.1278	0.0610
SASRec$_{\text{T}}$	0.0916	0.0490	0.1343	0.0598
SASRec$_{\text{V}}$	0.0953	0.0520	0.1374	0.0626

We used BERT[10] as the text encoder and SlowFast16x8-r101 as the video encoder and perform end-to-end training as mentioned in section 3.1. We fixed the learning rate of recommender model as $1e-4$, and searched for the optimal learning rates for the text encoder and video encoder from $\{1e-3, 1e-4\}$. The comparison results are reported in Table 12. Our results demonstrate that using only text features yields similar performance to the itemID feature. By analyzing the data, we have observed that some short videos have only a few words in their descriptions, which may contribute to the performance not being particularly competitive. On the other hand, the amount of information contained in the original videos far exceeds that of the video titles. Therefore, we believe that in the future, utilizing more powerful video understanding techniques can lead to better recommendation results.

K Recommendation Scenario of Collected Platform

Figure 7 illustrates the recommendation scenario of the micro video platform from which our MicroLens collected data. In this example, a user is recommended a video about trucks. After

---
[10]https://huggingface.co/prajjwal1/bert-small

a. The currently watched video   b. Swipe to the next video   c. New video

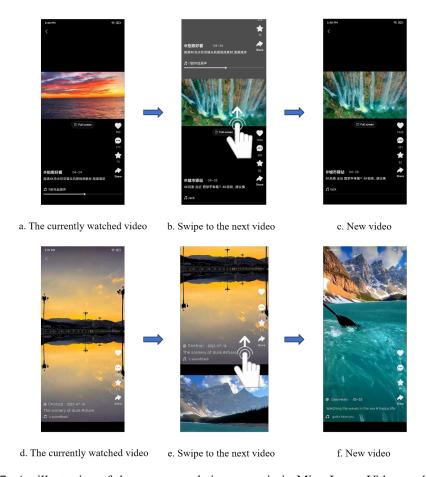d. The currently watched video   e. Swipe to the next video   f. New video

Figure 7: An illustration of the recommendation scenario in MicroLens. Videos a, b, and c are displayed in landscape format, while videos d, e, and f are displayed in portrait format. Please note that the format of the next video is random and can be either landscape or portrait. English translation is provided for all video titles.

watching a short segment, the user swipes up to the next video. All these videos allow user engagement through buttons for liking, sharing, and commenting, which are visible on the right side of the videos. On this platform, there are multiple ways to define positive and negative examples. For instance, the duration of video views, presence of likes, comments, or shares can all be considered as different levels of user feedback. However, among these behaviors, only comment behaviors are public without any access restrictions. Also, note that the videos and comments are publicly accessible both on the mobile app and the web. In the mobile app, users navigate to the next video by swiping gestures, while on the web, users use mouse scrolling to move to the next video. The web scene is displayed in the same way as the mobile app scene.

In the micro-video application, users are typically presented with a continuous stream of videos. The recommendation process continues uninterrupted through the user swiping up or mouse scrolling, ensuring a seamless flow of video recommendations.