A Content-Driven Micro-Video Recommendation Dataset at Scale

Yongxin Ni, Yu Cheng, Xiangyan Liu, Junchen Fu, Youhua Li, Xiangnan He,
Yongfeng Zhang, Fajie Yuan

https://arxiv.org/abs/2309.15379

https://github.com/westlake-repl/MicroLens

西湖大學
WESTLAKE UNIVERSITY

CONTENTS

01 Motivation

Motivation (MoRec)

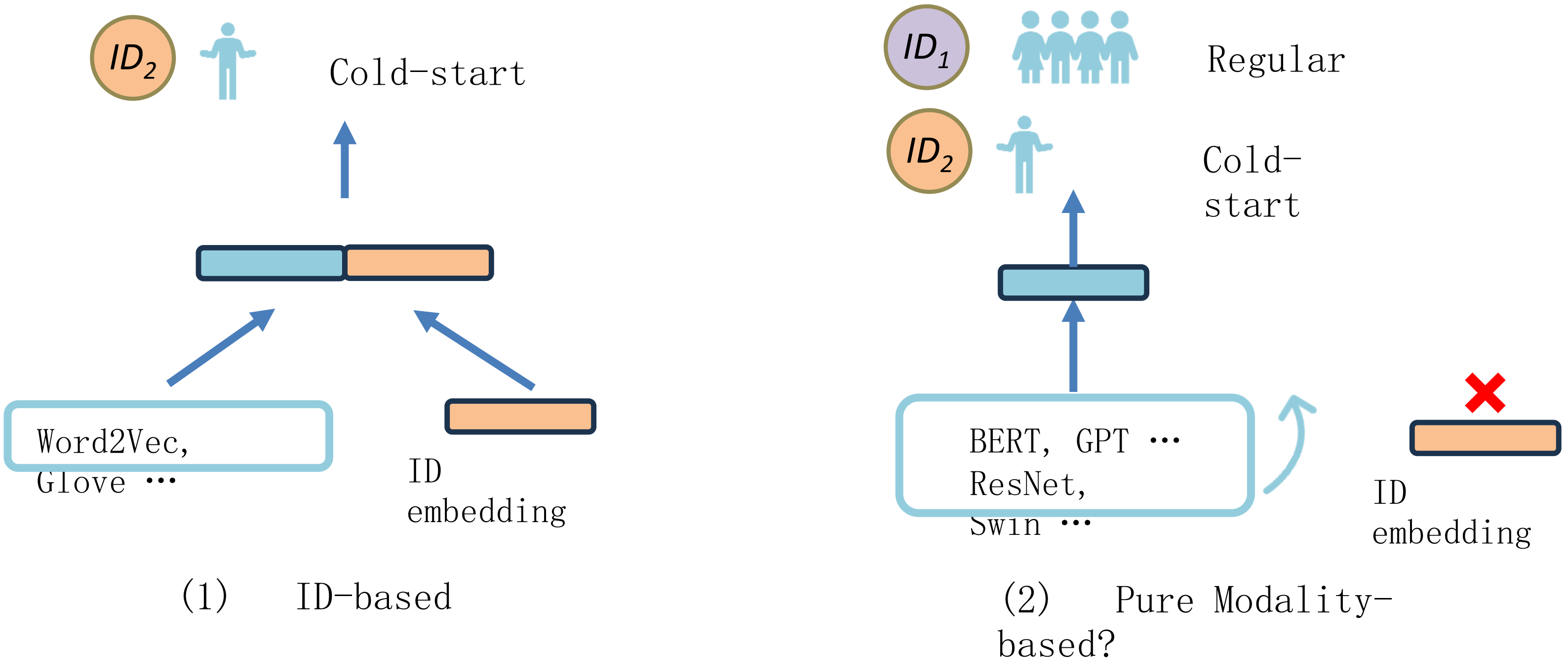
Where to Go Next for Recommender Systems? ID- vs. Modality-based Recommender Models Revisited

Zheng Yuan[1], Fajie Yuan[1], Yu Song[1], Youhua Li[1], Junchen Fu[1],
Fei Yang[2], Yunzhu Pan[1], Yongxin Ni[1]

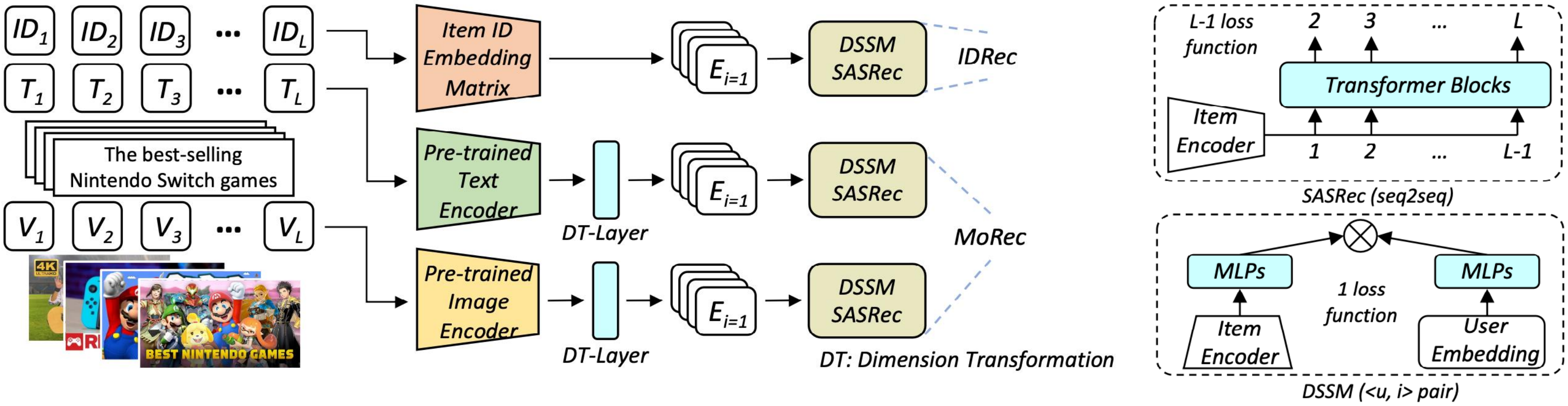
[1]Westlake University; [2]Zhejiang Lab

Code & datasets: https://github.com/westlake-repl/IDvs.MoRec

Motivation (MoRec)

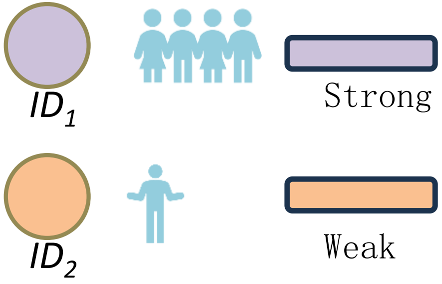


Motivation (MoRec)

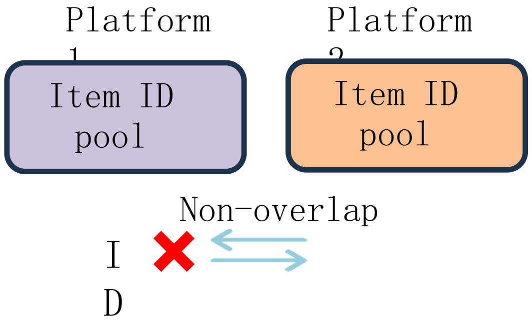


Nowadays, with the help of current textual/visual encoders, MoRec can be comparable to or even better than IDRec

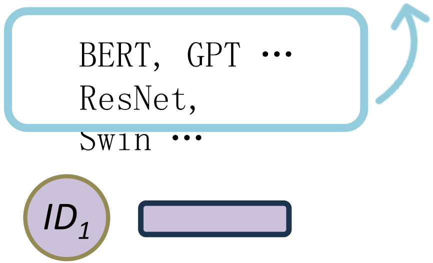
Motivation (MoRec)



(1) Cold-start setting



(2) Transfer



(3) Benefit from CV/NLP/MM

Motivation (Future Direction)

- Modality-based Recommendation
- Micro-video Recommendation
- Foundation Models for Recommender Systems
- “one4all” Paradigm

Motivation (Lack of Datasets)

- Domain
- Raw Content
- Scale
- Modality Diversity

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

02

MicroLens

MicroLens (Dataset)

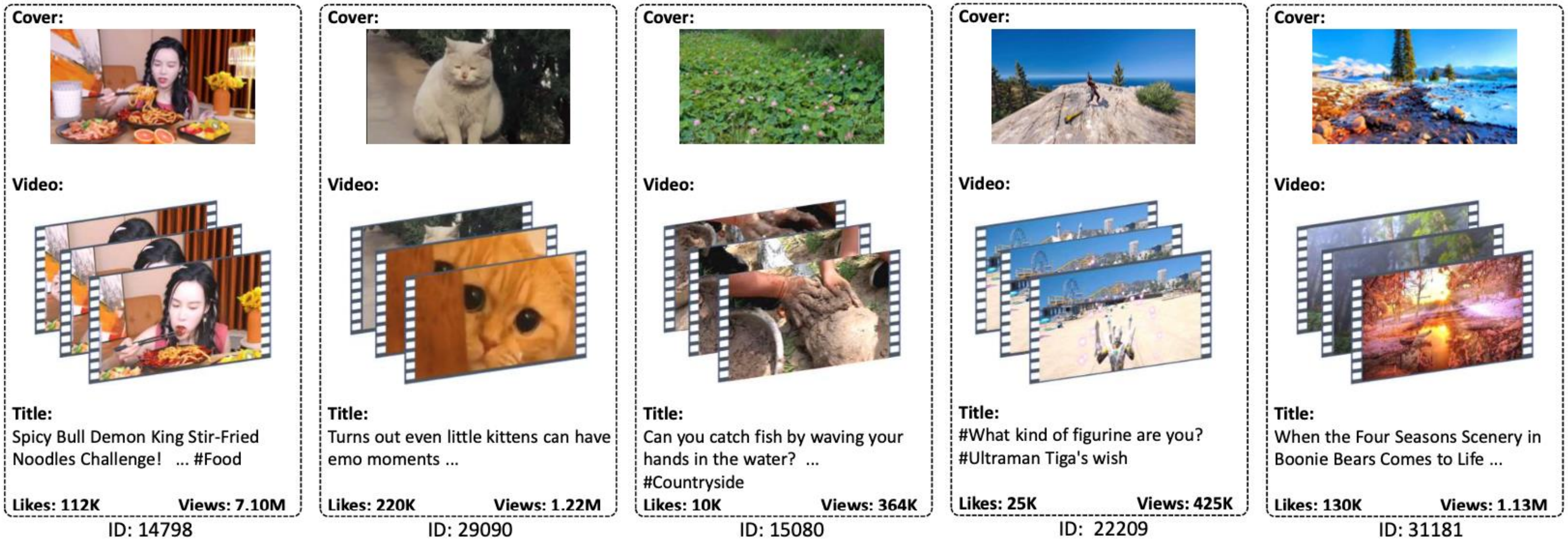
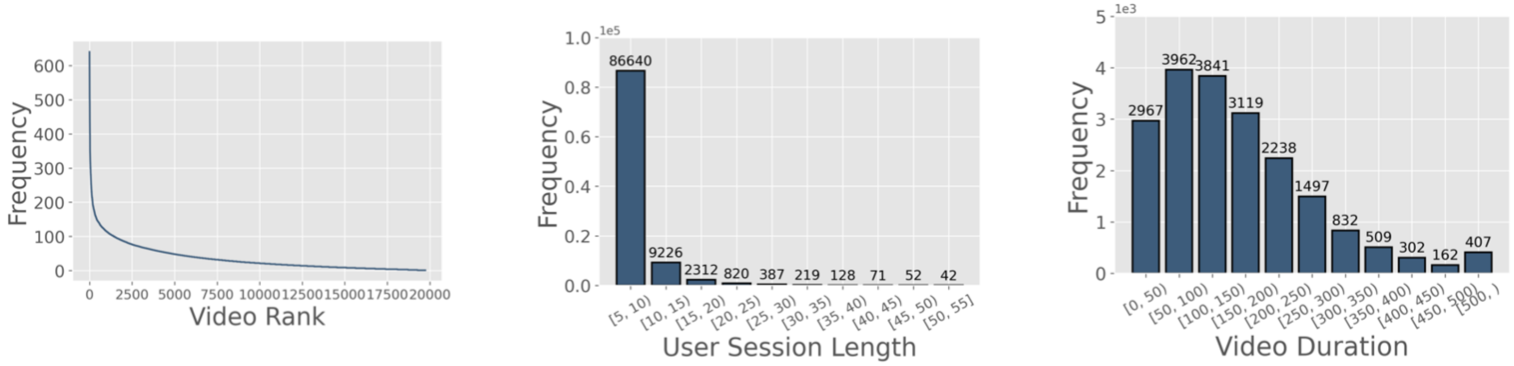


Figure 2: Item examples in MicroLens.

MicroLens (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	✔

MicroLens (Experiments)

- VideoRec
  - End-to-end manner
  - Train recommender model and video encoder simultaneously
- Investigate how *RS* benefits from *Video Understanding*
- 3 recommender models
  - CNN-based (NextItNet)
  - RNN-based (GRU4Rec)
  - Transformer-based (SASRec)
- 15 video encoders
  - R3D-r18, X3D-xs, C2D-r50, I3D- r50, X3D-s, Slow-r50, X3D-m, R3D-r50, SlowFast-r50, CSN-r101, X3D-l, SlowFast-r101, MViT-B-16x4, MViT-B-32x3, and VideoMAE

03 Findings

Findings (Benchmark 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)	$\text{YouTube}_{\text{ID}}$	0.0461	0.0229	0.0747	0.0301
	$\text{YouTube}_{\text{ID+V}}$ [7]	0.0392	0.0188	0.0648	0.0252
	MMGCN_{ID}	0.0141	0.0065	0.0247	0.0092
	$\text{MMGCN}_{\text{ID+V}}$ [54]	0.0214	0.0103	0.0374	0.0143
	GRCN_{ID}	0.0282	0.0131	0.0497	0.0185
	$\text{GRCN}_{\text{ID+V}}$ [53]	0.0306	0.0144	0.0547	0.0204
	$\text{DSSM}_{\text{ID+V}}$	0.0279	0.0137	0.0461	0.0183
	$\text{SASRec}_{\text{ID+V}}$	0.0799	0.0415	0.1217	0.0520
VideoRec (E2E Learning)	$\text{NexItNet}_{\text{V}}$ [62]	0.0862	0.0466	0.1246	0.0562
	$\text{GRU4Rec}_{\text{V}}$ [27]	0.0954	0.0517	0.1377	0.0623
	SASRec_{V} [31]	0.0948	0.0515	0.1364	0.0619

Findings (Benchmark Results)

- Methods: IDRec, VIDRec and VideoRec
  - We do not search parameters exhaustively for VideoRec
  - Only 5 frames of each video were used
- Findings: raw video content > pre-extracted frozen features

Findings (Video Understanding Meets Recommender Systems)

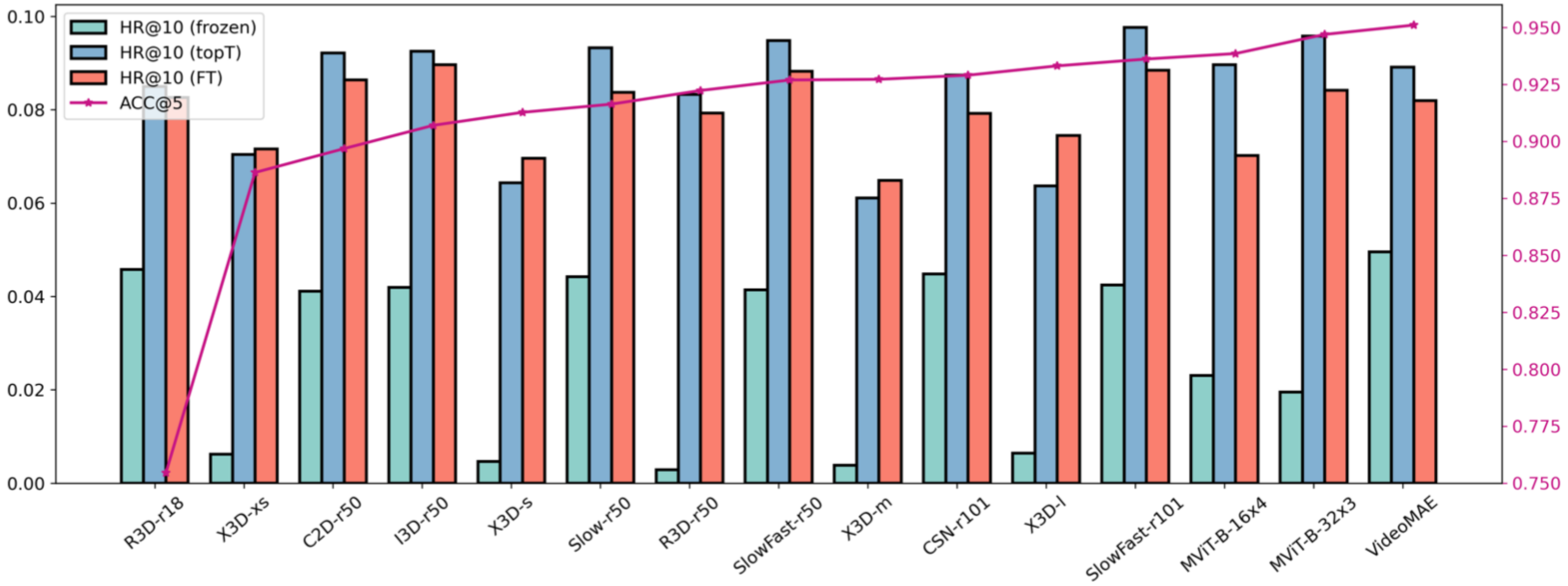


Figure 4: Video recommendation accuracy (bar charts) vs. video classification accuracy (purple line). Frozen means that the video encoder is fixed without parameter update, topT means that only the top few layers of the video encoder are fine-tuned, and FT means full parameters are fine-tuned.

Findings (Video Understanding Meets Recommender Systems)

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

Findings (Video Understanding Meets Recommender Systems)

- Better CV performance ≠ Higher recommendation accuracy
  - E.g., the worst video classification model R3D-r18
- In RS, finetuning top layers > full finetuning
  - full finetuning the video encoders is not necessary in recommender systems

Findings (Video Understanding Meets Recommender Systems)

- Knowledge learned from video understanding helps video recommendation
- Video semantic representations learned from CV task are not universal
  - a linear layer is not enough produce the same results as finetuning

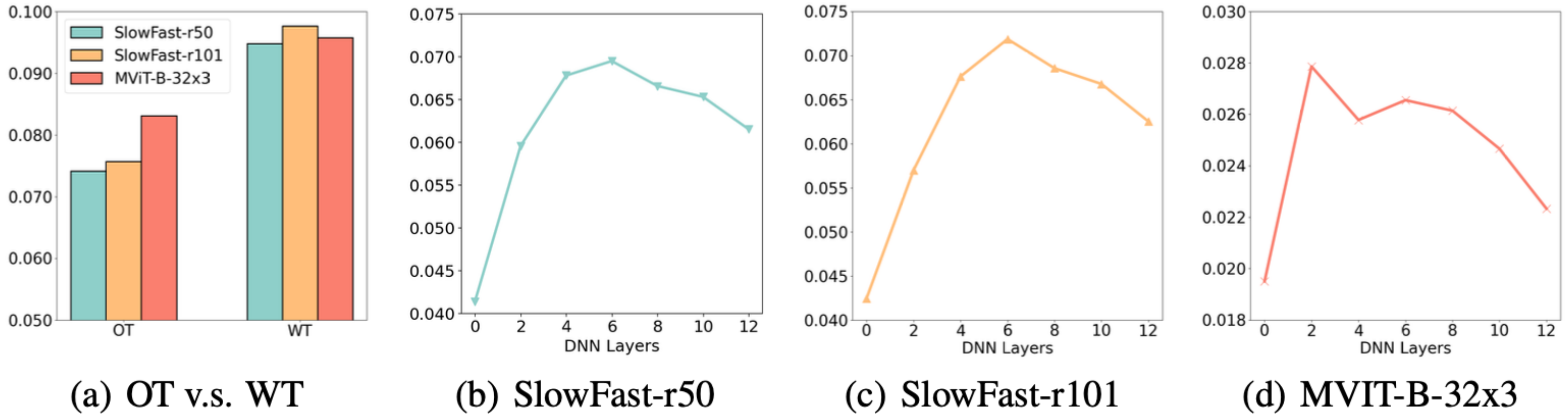


Figure 5: Ablation study of video encoders. (d) "WT" refers to the video encoders in SASRec_V have pre-trained weights from the video classification task, while "OT" denotes that they are randomly initialized. (b) (c) (d) are performance change by adding DNN layers on top of three frozen encoders.

Findings (Video Understanding Meets Recommender Systems)

- Our study is the first to show that raw video features can potentially replace ID features in both warm and cold item recommendation settings

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

Findings (Video Understanding Meets Recommender Systems)

- Our study is the first to show that raw video features can potentially replace ID features in both warm and cold item recommendation settings

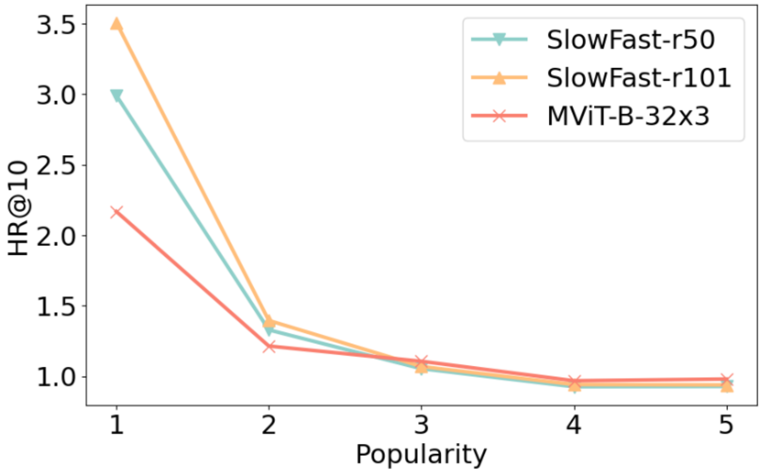


Figure 6: Results in different cold-start scenarios, with the y-axis representing the relative improvement of HR@10, calculated as the ratio of VideoRec to IDRec. The x-axis represents item groups divided by popularity level, the larger number indicates that items in the group are more popular.

Findings (Video Understanding Meets Recommender Systems)

- Summary: This work has taken a key step towards the goal of a universal "one-for-all" recommender paradigm
  - Dataset Support
  - VideoRec Paradigm Exploration

- Other Works

MoRec: Where to go next for recommender systems? id-vs. modality-based recommender models revisited

PixelRec: An Image Dataset for Benchmarking Recommender Systems with Raw Pixels

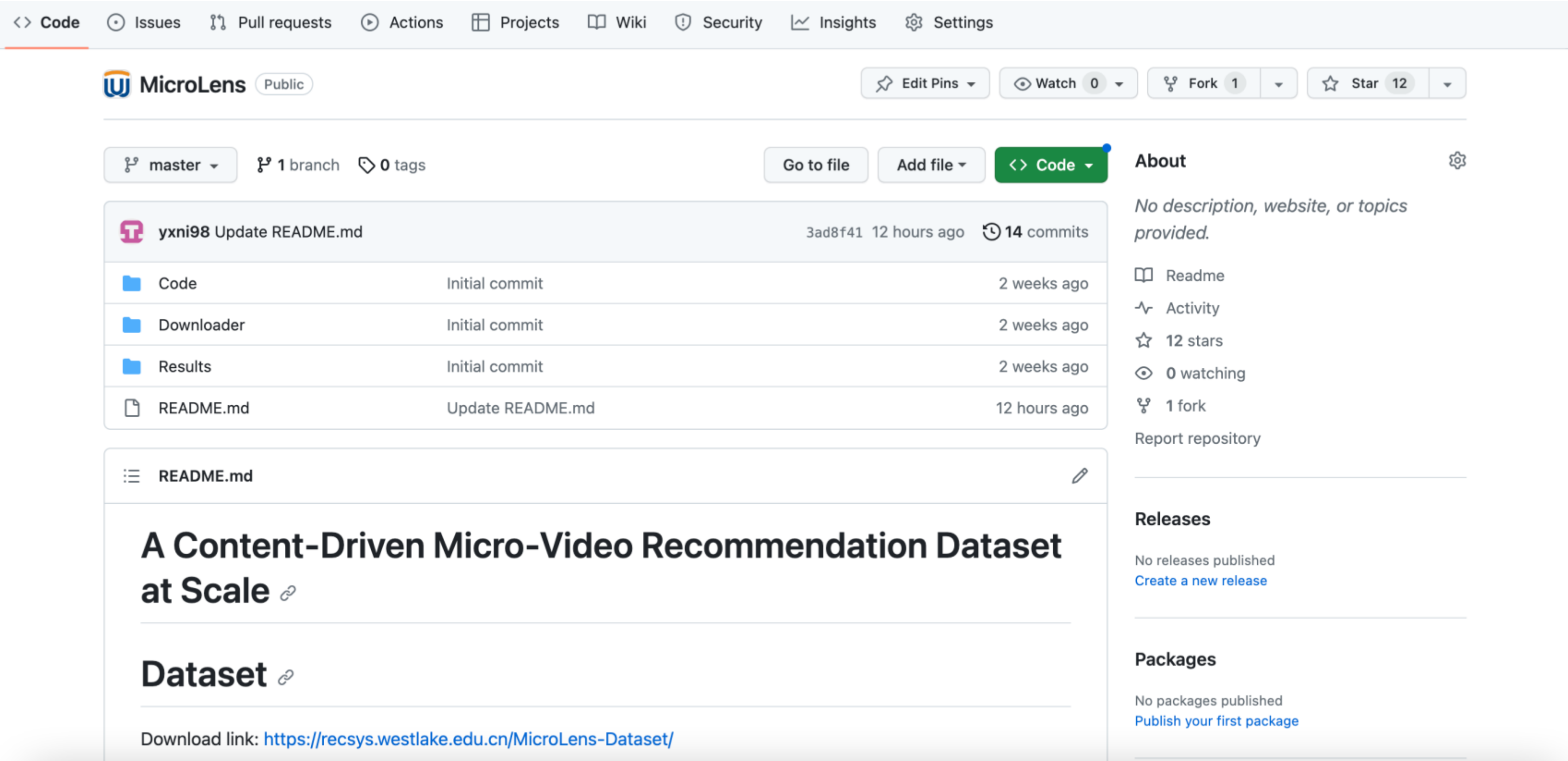
NineRec: A Benchmark Dataset Suite for Evaluating Transferable Recommendation

04

Code & Data

Code & Data

Find our GitHub:



西湖大學
WESTLAKE UNIVERSITY

THANKS

Yongxin Ni