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

Zheng Yuan*
Westlake University
yuanzheng@westlake.edu.cn

Fajie Yuan*†
Westlake University
yuanfajie@westlake.edu.cn

Yu Song
Westlake University
songyu@westlake.edu.cn

Youhua Li
Westlake University
liyouhua@westlake.edu.cn

Junchen Fu
Westlake University
fujunchen@westlake.edu.cn

Fei Yang
Zhejiang Lab
yangf@zhejianglab.com

Yunzhu Pan
Westlake University
panyunzhu@westlake.edu.cn

Yongxin Ni
Westlake University
niyongxin@westlake.edu.cn

ABSTRACT

Recommendation models that utilize unique identities (IDs for short) to represent distinct users and items have been state-of-the-art (SOTA) and dominated the recommender systems (RS) literature for over a decade. Meanwhile, the pre-trained modality encoders, such as BERT [9] and Vision Transformer [11], have become increasingly powerful in modeling the raw modality features of an item, such as text and images. Given this, a natural question arises: can a purely modality-based recommendation model (MoRec) outperforms or matches a pure ID-based model (IDRec) by replacing the itemID embedding with a SOTA modality encoder? In fact, this question was answered ten years ago when IDRec beats MoRec by a strong margin in both recommendation accuracy and efficiency.

We aim to revisit this 'old' question and systematically study MoRec from several aspects. Specifically, we study several sub-questions: (i) which recommendation paradigm, MoRec or IDRec, performs better in practical scenarios, especially in the general setting and warm item scenarios where IDRec has a strong advantage? does this hold for items with different modality features? (ii) can the latest technical advances from other communities (i.e., natural language processing and computer vision) translate into accuracy improvement for MoRec? (iii) how to effectively utilize item modality representation, can we use it directly or do we have to adjust it with new data? (iv) are there any key challenges that MoRec needs to address in practical applications? To answer them, we conduct rigorous experiments for item recommendations with two popular modalities, i.e., text and vision. We provide the first empirical evidence that MoRec is already comparable to its IDRec counterpart with an expensive end-to-end training method, even for warm item recommendation. Our results potentially imply that the dominance of IDRec in the RS field may be greatly challenged in the future. We release our code and other materials at https://github.com/westlake-repl/IDvs.MoRec.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Recommender Systems, ID-based Recommendation, Modality-based Recommendation, End-to-end Training

ACM Reference Format:
Zheng Yuan*, Fajie Yuan*†, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to Go Next for Recommender Systems? ID- vs. Modality-based Recommender Models Revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591932

1 INTRODUCTION

Recommender system (RS) models learn the historical interactions of users and items and recommend items that users may interact with in the future. RS is playing a key role in search engines, advertising systems, e-commerce websites, video and music streaming services, and various other Internet platforms. The modern recommendation models usually use unique identities (ID) to represent users and items, which are subsequently converted to embedding vectors as learnable parameters. These ID-based recommendation models (IDRec) have been well-established and dominated the RS field for over a decade until now [28, 49, 77].

Despite that, IDRec has key weaknesses that can not be ignored. First, IDRec highly relies on the ID interactions, which fails to provide recommendations when users and items have few interactions [74, 76], a.k.a. the cold-start setting. Second, pre-trained IDRec is not transferable across platforms given that userIDs and itemIDs are in general not shareable in practice. This issue seriously limits the development of big & general-purpose RS models [1, 10, 61], an emerging paradigm in the deep learning community. Third, pure

∗ Equal Contribution.
† Corresponding author. Fajie designed and supervised this research; Zheng performed this research, in charge of key technical parts; Fajie, Zheng, Yu wrote the manuscript. Yunzhu collected the Bili dataset, other authors assisted partial experiments.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591932

IDRec cannot benefit from technical advances in other communities, such as powerful foundation models (FM) [1] developed in NLP (natural language processing) and CV (computer vision) areas. Moreover, maintaining a large and frequently updated ID embedding matrix for users and items remains a key challenge in industrial applications [56]. Last but not the least, recommendation models leveraging ID features have obvious drawbacks in terms of interpretability, visualization and evaluation.

One way to address these issues is to replace the ID embedding (of IDRec) with an item modality encoder (ME), especially when item modality features such as images and text are available. We refer to such recommendation models as MoRec. In fact, such MoRec appeared in literature many years ago but it was mainly used to solve cold-start or cross-domain recommendation problems [3, 14, 58]. In other words, MoRec is rarely adopted when recommending non-cold or popular items unless combined with other effective features, such as the itemID features, e.g., in [20, 26, 60]. A key reason might be that these item ME developed in the past years (e.g., word embedding [40] and some shallow neural networks [58]) are not as expressive as typical itemID embeddings. Today, however, given the recent great success of FM, we think it is time to revisit the key comparison between modern MoRec and IDRec, especially for regular (or non cold-item) recommendation. For example, BERT [9], GPT-3 [2] and various Vision Transformers (ViT) [11, 37] have revolutionized the NLP and CV fields when representing textual and visual features. Whether item representations learned by them are better suited for the *regular* recommendation task than ID features remains unknown.

In this paper, we intend to rethink the potential of MoRec and investigate a key question: *should we still stick to the IDRec paradigm for future recommender systems?* We concentrate on item recommendation based on the text and vision modalities — the two most common modalities in literature. To be concise, we attempt to address the following sub-questions:

Q(i): Equipped with strong modality encoders (ME), can MoRec be comparable to or even surpass IDRec in regular, especially in warm-start item recommendation scenario? To answer this question, we conduct empirical studies by taking into account the **two** most representative recommendation architectures (i.e., two-tower based DSSM [24, 50] and session-based SASRec [25]) equipped with **four** powerful ME evaluated on **three** large-scale recommendation datasets with **two** modalities (text and vision).

Novelty clarification: Though much previous literature has studied MoRec and compared with many baselines [31, 41, 63, 64, 79], unfortunately none of them provided a *fair* or rigorous comparison between their proposed MoRec and the corresponding IDRec counterparts in regular or even warm item recommendation setting. Fair comparison here means that MoRec and IDRec should at least be compared with the same *backbone network* and *experimental settings*, such as samplers and loss functions. Without a fair comparison, the community can not truly assess the progress of MoRec and the expressive power of ME for recommendation.

Q(ii): If Q(i) is yes, can the recent technical advances developed in NLP and CV fields translate into accuracy improvement in MoRec when using text and visual features? We address this question by performing three experiments. First, we evaluate MoRec by comparing smaller *vs* larger ME given that pre-trained ME with larger model sizes tends to perform better than their smaller counterparts in various downstream tasks; second, we evaluate MoRec by comparing weaker *vs* stronger ME where weaker and stronger are determined by NLP and CV tasks; third, we evaluate MoRec by comparing ME with *vs* without pre-training on corresponding NLP and CV datasets.

Q(iii): Are the representations learned by these foundation models as general as claimed? How can we effectively use item modality representations derived from an NLP or CV encoder network? A desirable goal of FM research is to develop models that generate universal representations that can be directly used for various downstream tasks [34]. We examine this by first extracting frozen modality features from well-known ME and then adding them as common features for recommendation models, often referred to as the two-stage (TS) paradigm. This is a common practice for large-scale industrial recommender systems due to training efficiency consideration [7, 39]. We then compare TS with joint or end-to-end (E2E) training of both the recommendation architecture and ME.

Novelty clarification: Though several recent literature has explored E2E learning [64, 66, 69, 70] for recommendation, few of them explicitly discussed the substantial accuracy and efficiency gap (more than 100x) between TS and E2E paradigms. More importantly, most of them only discussed the DSSM architecture (or other two-tower variants) without considering more powerful and computationally more expensive sequence-to-sequence (seq2seq) training approach (e.g., used in SASRec and NextItNet [75]). Furthermore, all of them are only for text recommendation, and so far there is no *modern* (last 5 years) *peer-reviewed* literature considering the E2E learning paradigm for image recommendation.

In addition to the aforementioned key questions, we have also identified several challenges that remain unexplored for MoRec when utilizing the end-to-end learning paradigm.

2 IDREC & MOREC

One core function of a recommendation model is to represent items and users and calculate their matching score. Denote $\mathcal{I}$ (of size $|\mathcal{I}|$) and $\mathcal{U}$ (of size $|\mathcal{U}|$) as the set of items and users, respectively. For an item $i \in \mathcal{I}$, we can represent it either by its unique ID i or its modality content, such as text and visual features. Likewise, for a user $u \in \mathcal{U}$, we can represent her either by the unique ID u or the profile of u, where a profile can be the demographic information or a sequence of interacted items.

In IDRec, an ID embedding matrix $\mathbf{X}^{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}| \times d}$ is initialized, where d is the embedding size. Each vector in $\mathbf{X}^{\mathcal{I}}$ represents the latent space of an item i, and can be viewed as a simple item encoder. During training and inference, IDRec retrieves $\mathbf{X}^{\mathcal{I}}{}_i \in \mathbb{R}^d$ from $\mathbf{X}^{\mathcal{I}}$ as the embedding of i and then feeds it to the recommendation network.

In MoRec, items are assumed to contain modality information. For item i, MoRec uses ME to generate the representation for the raw modality feature of i and uses it to replace the ID embedding vector in IDRec. For instance, in the news recommendation scenario, we can use the pre-trained BERT or RoBERTa [36] as text ME and represent a piece of news by the output textual representation of its

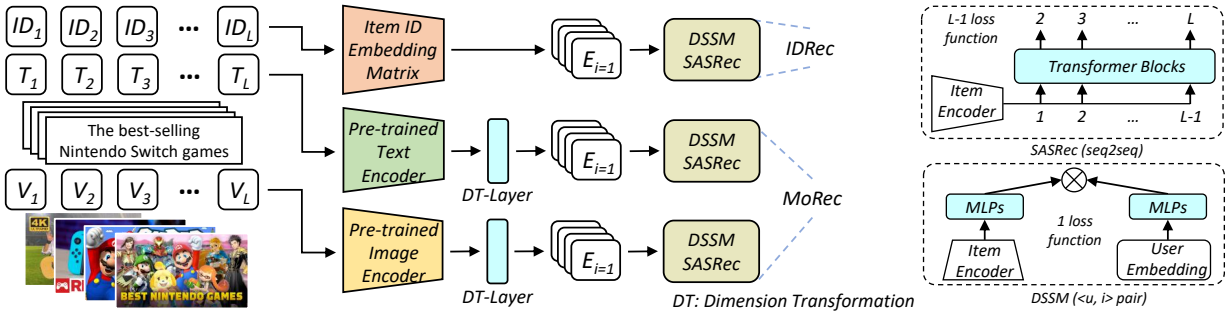


Figure 1: Illustration of IDRec *vs* MoRec. V_i and T_i denote raw features of vision and text modalities. E_i is the item representation vector fed into the recommender model. The only difference between IDRec and MoRec is the item encoder. IDRec uses an itemID embedding matrix as the item encoder, whereas MoRec uses the pre-trained ME (followed by a dense layer for the dimension transformation, denoted by DT-layer) as the item encoder.

title. Similarly, when items contain visual features, we can simply use a pre-trained ResNet or ViT as vision ME.

In this paper, we perform rigorous empirical studies on two most commonly adopted recommendation paradigms: DSSM [24] and SASRec [25].[1] The original DSSM model is a two-tower based architecture where users/items are encoded by their own encoder networks with user and item IDs as input. SASRec is a well-known sequential recommendation model based on multi-head self-attention (MHSA) [59] which describes a user by her interacted item ID sequence. As mentioned before, by replacing ID embeddings with an item ME, we obtain their MoRec versions for both DSSM and SASRec. We illustrate IDRec and MoRec in Figure 1.

2.1 Training Details

Denote $\mathcal{R}$ as the set of all observed interactions in the training set. For each positive $< u, i > \in \mathcal{R}$, we randomly draw a negative sample $< u, j > \notin \mathcal{R}$ in each training epoch, following [21, 49]. The positive and sampled negative interactions can form the training set $\mathcal{R}^{train}$. Following [21, 25], we adopt the widely used binary cross entropy loss as the objective function for both DSSM and SASRec,[2] and their MoRec versions for a fair comparison:

$$\begin{cases} min - \sum_{u \in \mathcal{U}} \sum_{i \in [2,\ldots,L]} \left\{ \log(\sigma(\hat{y}_{ui})) + \log(1 - \sigma(\hat{y}_{uj})) \right\} & \text{SASRec} \\ min - \sum_{<u,i,j> \in \mathcal{R}} \left\{ \log(\sigma(\hat{y}_{ui})) + \log(1 - \sigma(\hat{y}_{uj})) \right\} & \text{DSSM} \end{cases} \tag{1}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function, L is the interaction sequence length of user u. i and j denotes positive and negative item respectively for u, $\hat{y}_{ui}$ is the matching score between hidden vectors of user (u) encoder and item (i) encoder. Note that SASRec's user encoder (by seq2seq training) produces a different hidden vector at each position of the interaction sequence. Without special mention, all parameters of the entire recommendation model are optimized during training in the following experiments.

3 EXPERIMENTAL SETUPS

3.1 Datasets

We evaluate IDRec and MoRec on three real-world datasets, namely, the MIND news clicks dataset from the Microsoft news recommendation platform [67], the HM clothing purchase dataset from the H&M platform[3] and the Bili[4] comment dataset from an online video recommendation platform.[5] Purchases and comments can be considered implicit click signals, as it is reasonable to assume that the user has clicked on the item before making a purchase or leaving a comment. However, we cannot assume the opposite holds, which is a common property in most recommendation datasets, i.e. unobserved items can be either positive or negative for the user.

To ensure a fair comparison between IDRec and MoRec, the dataset used should guarantee that users' clicking decisions on an item are solely based on the modality content features of the item. Intuitively, the cover of an image or video and the title of a news article, play a crucial role in providing users with the very first impression of an item. This impression significantly influences their decision to click on the item. Therefore, in MIND, we represent items by their news article titles, while in HM & Bili, we represent items using their corresponding cover images. Nevertheless, it is still possible that these datasets may not *perfectly* meet the requirement. Particularly, within the e-commerce context of the HM dataset, factors such as the item's cover image, price, and sales volume may collectively influence a user's decision to click on an item (refer to Figure 2). This means relying solely on a cover image in the

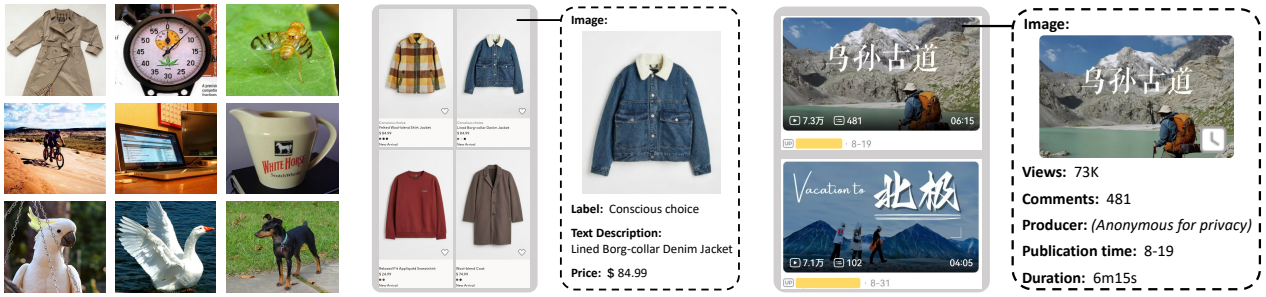
[1] We did not study other CTR (click-through rate) prediction models, as they essentially belong to the same category as DSSM, with the key difference being that many CTR models are based on single-tower backbone networks [7, 15, 21, 77]. Intuitively, such difference generally does not affect our subsequent conclusions (see section 4.1), since improvement from a two-tower backbone to a single-tower is often limited if having the same training manners [33, 81]. However, DSSM or CTR models are quite different from the seq2seq-based sequential recommendation models, such as SASRec. For example, as shown in Figure 1, SASRec has $L-1$ loss functions for each interaction sequence (input: 1, 2, ..., $L-1$, predict: 2, ..., L), while DSSM (or other CTR models) *typically* uses one loss function to predict an interaction of a $< u, i >$ pair.

[2] We also assessed other commonly used loss functions and samplers, such as the in-batch loss proposed in [71] and the softmax loss utilized in [75]. Although both IDRec and MoRec showed improved performance, their comparison results remained basically consistent.

[3] https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/overview

[4] https://www.bilibili.com/

[5] To build this dataset, we randomly crawled URLs of short videos (with duration time less than 10 minutes) from 23 different video channels of Bili from October 2021 to March 2022. Then we recorded public comments of these videos as interactions. Finally, we merged all user interactions chronologically and removed duplicate interactions.



(a) Item cases on ImageNet1K. (b) Item cases on HM. (c) Item cases on Bili.

Figure 2: Item cases on datasets. Image ME we used are all pre-trained in the ImageNet1K dataset.

Table 1: Dataset characteristics. n and m denote the numbers of users and items. $|\mathcal{R}|^{train}$, $|\mathcal{R}^{valid}|$ and $|\mathcal{R}^{test}|$ denote the number of interactions of the training set, validation set and testing set, respectively. $|\mathcal{R}|/(nm)$ represents density.

Dataset	n	m	$\|\mathcal{R}\|^{train}$	$\|\mathcal{R}^{valid}\|$	$\|\mathcal{R}^{test}\|$	$\|\mathcal{R}^{train}\|/(nm)$
MIND	630K	80K	8,407K	630K	630K	0.0167%
HM	500K	87K	5,500K	500K	500K	0.0127%
Bili	400K	128K	4,400K	400K	400K	0.0086%

HM dataset may not be adequate for MoRec to effectively capture these non-visual features, as it is the only input to the item encoder. In contrast, IDRec is known to be able to implicitly learn such features from the latent embedding space [28]. That is, MoRec's performance may still have room for improvement if a more ideal[6] dataset or more useful content features were taken into account.

To construct the datasets for experiments, we randomly select around 400K, 500K and 600K users from Bili, HM, and MIND, respectively. Then, we perform basic pre-processing by setting the size of all images to 224×224 and the title of all news articles to a maximum of 30 tokens (covering 99% of descriptions). For MIND, we select the latest 23 items for each user to construct the interaction sequence. For HM and Bili, we choose the 13 most recent interactions since encoding images requires much larger GPU memory (especially with the SASRec architecture). Following [49], we remove users with less than 5 interactions, simply because we do not consider cold user settings in this paper.

3.2 Hyper-parameters

For all methods, we employ an AdamW [38] as the default optimizer and find that the dropout rate set to 0.1 (i.e., removing 10% parameters) offers the optimal results on the validation set. Regarding other hyper-parameters, we follow the common practice and perform extensive searching. For IDRec, we tune the learning rate γ from {1e-3, 5e-4, 1e-4, 5e-5}, the embedding/hidden size d from {64, 128, 256, 512, 1024, 2048, 4096}. We set batch size b to 1024 for DSSM and 128 for SASRec. For MoRec, we set d to 512 for both DSSM and SASRec, b to 512 and 64 for DSSM and SASRec respectively due to GPU memory constraints. Given that ME (e.g., BERT and ResNet) has already well pre-trained parameters, we use relatively smaller γ than other parts in the recommender model. That is, we search γ from {1e-4, 5e-5, 1e-5} for the pre-trained ME networks, and set γ to 1e-4 for other parts with randomly initialized parameters. Finally, we tune the weight decay β from $\{0.1, 0.01, 0\}$ for both IDRec and MoRec.

For the MLPs (multilayer perceptron) used in DSSM, we initially set their middle layer size to d as well and search the layer number l from $\{0, 1, 3, 5\}$ but find that $l = 0$ (i.e., no hidden layers) always produces the best results. For the Transformer block used in SASRec, we set l to 2 and the head number of the multi-head attention to 2 for the optimal results. All other hyper-parameters are kept the same for IDRec and MoRec unless specified otherwise.

3.3 Comparison Settings

For a fair comparison, we ensure that IDRec and MoRec have exactly the same network architecture except for the item encoder. For both text and vision encoders, we pass their output item representations to a DT-layer (see Figure 1) for dimension transformation. Regarding the hyper-parameter setting, our principle is to ensure that IDRec are fully tuned in terms of learning rate γ, embedding size d, layer number l, and dropout ρ. While for MoRec, we attempt to first use the same set of hyper-parameters as IDRec and then perform some basic searching around the best choices. Therefore, without special mention, we do not guarantee that the results reported by MoRec are the best, because searching all possible hyperparameters for MoRec is very expensive and time-consuming, sometimes taking more than 100x compute and training time than IDRec, especially for vision, see Table 6. Thereby, how to efficiently find the optimal hyper-parameters of MoRec is an important but unexplored research topic.

3.4 Evaluations

We split the datasets into training, validation, and testing sets by adopting the standard leave-one-out strategy. Specifically, the latest interaction of each user was used for evaluation, while second-to-last was used as validation for hyper-parameter searching, and all others are used for training. We evaluate all models using two popular top-N ranking metrics: HR@N (Hit Ratio) and NDCG@N (Normalized Discounted Cumulative Gain), where N is set to 10.

[6]It seems that so far there is no publicly available dataset that fully satisfies the above mentioned requirement.

Table 2: Accuracy (%) comparison of IDRec and MoRec using DSSM and SASRec for regular setting. MoRec with different ME are directly denoted by their encoder names for clarity. The best results for DSSM and SASRec are bolded. 'Improv.' is the relative improvement of the best MoRec compared with the best IDRec. All results of MoRec are obtained by fine-tuning their whole parameters including both the item encoder and user encoder. Swin-T and Swin-B are Swin Transformer with different model sizes, where T is tiny and B is base. ResNet50 is a 50-layer ResNet variant.

Dataset	Metrics	DSSM			SASRec				Improv.
		IDRec	$\text{BERT}_{\text{base}}$	$\text{RoBERTa}_{\text{base}}$	IDRec	$\text{BERT}_{\text{small}}$	$\text{BERT}_{\text{base}}$	$\text{RoBERTa}_{\text{base}}$	
MIND	HR@10	**3.58**	2.68	3.07	17.71	18.50	18.23	**18.68**	+5.48%
	NDCG@10	**1.69**	1.21	1.35	9.52	9.94	9.73	**10.02**	+5.25%
		IDRec	ResNet50	Swin-T	IDRec	ResNet50	Swin-T	Swin-B	
HM	HR@10	**4.93**	1.49	1.87	6.84	6.67	6.97	**7.24**	+5.85%
	NDCG@10	**2.93**	0.75	0.94	**4.01**	3.56	3.80	3.98	-0.75%
Bili	HR@10	**1.14**	0.38	0.57	3.03	2.93	3.18	**3.28**	+8.25%
	NDCG@10	**0.56**	0.18	0.27	1.63	1.45	1.59	**1.66**	+1.84%

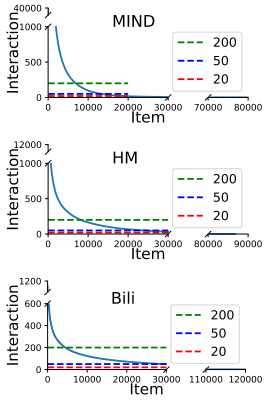


Figure 3: Item popularity distribution.

We rank the ground-truth target item by comparing it with all the left items in the item pool. Finally, we report results on the testing set, but find the best hyper-parameters via the validation set.

4 COMPARATIVE STUDIES (Q(I))

According to existing literature, MoRec can easily beat IDRec in the new item or cold-start item recommendation settings [17, 42, 57]. We report such results in the Appendix A.1. In this paper we focus on evaluating them in the more challenging setting: regular (mixture of warm and cold items) and warm-start item recommendation scenarios, where IDRec is usually very strong. To the best of our knowledge, such comparisons have not been explicitly discussed in the existing literature.

As mentioned, we evaluate IDRec and MoRec with the two most important recommendation architectures, i.e., DSSM and SASRec. We use pre-trained BERT and RoBERTa as ME when items are of text features, and use pre-trained ResNet and Swin Transformer [37] when items are of visual features.[7] Note for BERT and RoBERTa, we add the DT-layer (see Figure 1) on the final representation of the '[CLS]' token. We report results on the testing set in Table 2 for regular setting (i.e. the original distribution) and Table 3 for warm-start settings where cold items are removed.

4.1 MoRec *vs* IDRec (Regular Setting)

As shown in Table 2, we observe that DSSM always substantially underperforms SASRec, regardless of the item encoding strategy used. For instance, SASRec-based IDRec is around 4.9× better than DSSM-based IDRec in terms of HR@10 for news recommendation, although their training, validation, and testing sets are kept exactly the same. The performance gap for image recommendation is relatively small, around 1.4× and 2.7×, on HM and Bili, respectively. This is consistent with much prior literature [22, 25], where representing and modeling users with their interacted item sequence is often more powerful than dealing them as individual userIDs.

Second, we notice that with the DSSM architecture, MoRec perform much worse than IDRec in all three datasets even with the state-of-the-art (SOTA) ME, in particular for the visual recommendation scenarios. By contrast, with the SASRec architecture, MoRec consistently achieve better results than IDRec on MIND using any of the three text encoders, i.e., $\text{BERT}_{\text{small}}$, $\text{BERT}_{\text{base}}$ and $\text{RoBERTa}_{\text{base}}$. For instance, MoRec outperform IDRec by over 5% on the two evaluation metrics with the $\text{RoBERTa}_{\text{base}}$ text encoder. Meanwhile, MoRec perform comparably to IDRec when using Swin Transformer as ME but perform relatively worse when using ResNet50. The performance disparity of MoRec between DSSM and SASRec potentially implies that **a powerful recommendation backbone (SASRec *vs* DSSM) and training approach (seq2seq *vs* $< u, i >$ pair) is required to fully harness the strengths of the modality-based item encoder**. Given MoRec's poor results with DSSM, we mainly focus on the SASRec architecture in the following.

[7]We provide the URLs of all pre-trained modality encoders utilized in our study at https://github.com/westlake-repl/IDvs.MoRec.

Table 3: MoRec *vs* IDRec (HR@10) in the warm-start settings with SASRec as user backbone. Warm-20 means removing items with less than 20 interactions in the original dataset.

Dataset	MIND		HM		Bili	
	IDRec	$\text{BERT}_{\text{base}}$	IDRec	Swin-T	IDRec	Swin-T
Warm-20	20.12	**20.19**	7.89	**8.05**	3.48	**3.57**
Warm-50	20.65	**20.89**	**8.88**	8.83	**4.04**	4.02
Warm-200	**22.00**	21.73	**11.15**	11.10	**10.04**	9.98

4.2 MoRec *vs* IDRec (Warm Item Settings)

To validate the performance of MoRec and IDRec for warm item recommendation, we constructed new datasets with different item popularity. We show the item popularity distribution of the original datasets in Figure 3. For each dataset, we remove items with less than 20, 50, 200 interactions from the original datasets. We report the recommendation accuracy of all three datasets in Table 3. It can be seen that IDRec is getting stronger and stronger from warm-20, warm-50 to warm-200. In warm-20 dataset, MoRec is slightly better than IDRec, while in warm-200, MoRec is slightly worse than IDRec for text recommendation. This is reasonable since IDRec is known to to be good at modeling popular items according to the

existing literature [4, 71, 73]. But even in these warm-start setting, MoRec is still comparable to IDRec. Such property is appealing since it is well-known that MoRec can easily beat IDRec in the cold-start setting (see Appendix) and has a natural advantage for tranfer learning or cross-domain recommendation. Even further, recent work have shown that large MoRec models have the potential to be a foundation recommendation models [52, 53], capable of achieving the ambitious goal of "one model for all" [52, 61].

The above results shed the following insights: (1) the recommendation architecture (seq2seq SASRec or two-tower DSSM) of MoRec has a very large impact on its performance; (2) its item ME also influences the performance of MoRec; (3) **(Answer for Q(i)) equipped with the most powerful ME, MoRec can basically beat its IDRec counterpart for text recommendation (both cold and warm item settings) and is on par with IDRec for visual recommendation when using the sequential neural network recommendation architecture. However, it seems that there is little chance for MoRec to replace IDRec with the typical DSSM training approach in either regular or the warm-start setting;** (4) although MoRec cannot beat IDRec in terms of very popular item recommendation, they still show very competitive results. To the best of our knowledge, this is the first paper that explicitly claims that pure MoRec can be comparable to pure IDRec (when they are compared under the same sequential[8] recommendation architecture), even for the very challenging warm item recommendation.

5 INHERIT ADVANCES IN NLP & CV? (Q(II))

Intuitively, MoRec have the potential to bring powerful representation learning techniques from other communities, such as NLP and CV, to recommendation tasks. However, this has not been formally studied. Here, we ask: can recent advances in NLP and CV translate into improved accuracy for recommendation tasks? We aim to answer it from the following perspectives.

First, we investigate whether a larger pre-trained ME enables better recommendation accuracy since in NLP and CV larger pre-trained models tend to offer higher performance in corresponding downstream tasks. As shown in Figure 4, a larger vision item encoder always achieves better image recommendation accuracy, i.e., ResNet18-based MoRec < ResNet34-based MoRec < ResNet50-based MoRec, and Swin-T based MoRec < Swin-B based MoRec. Similarly, we find that $\text{BERT}_{\text{tiny}}$-based MoRec < $\text{BERT}_{\text{base}}$-based MoRec < $\text{BERT}_{\text{small}}$-based MoRec. One difference is that $\text{BERT}_{\text{base}}$-based MoRec do not outperform $\text{BERT}_{\text{small}}$-based MoRec although the latter has a smaller-size BERT variant. We conclude that, in general, a larger and more powerful ME from NLP and CV tends to improve the recommendation accuracy, but this may not strictly apply in all cases.

Second, we investigate whether a stronger encoder network enables better recommendations. For example, it is recognized that RoBERTa outperforms BERT [36], and BERT outperforms the unidirectional GPT [45], such as OPT [80], for most NLP understanding (but not generative) tasks with similar model sizes, and that Swin Transformer often outperforms ResNet in many CV tasks [37]. In addition, these modern pre-trained NLP foundation models easily outperform TexTCNN [27] and GloVe [43], two well-known shallow models developed about ten years ago. As shown in Figure 4, MoRec's performance keeps consistent with the findings in NLP and CV, i.e., $\text{RoBERTa}_{\text{base}}$-based MoRec > $\text{BERT}_{\text{base}}$-based MoRec > OPT_{125M}-based MoRec > TextCNN-based MoRec > GloVe-based MoRec, and Swin-T based MoRec > ResNet50-based MoRec (Swin-T has a similar model size to ResNet50, the same for $\text{RoBERTa}_{\text{base}}$, $\text{BERT}_{\text{base}}$ and OPT_{125M}).

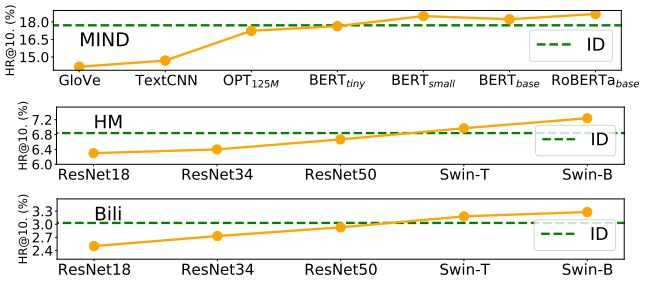


Figure 4: Accuracy with different pre-trained ME in MoRec. Parameters of the pre-trained encoder network are all fine-tuned on the recommendation task.

Table 4: Pre-trained (PE) ME *vs* TFS on the testing set regarding HR@10 (%). $\text{BERT}_{\text{base}}$ are used as text ME, and ResNet50 and Swin-T are used as vision ME. 'Improv.' indicates the relative improvement of PE over TFS.

Dataset	ME	Base			50K		
		TFS	PE	Improv.	TFS	PE	Improv.
MIND	$\text{BERT}_{\text{base}}$	17.78	**18.23**	+2.53%	**15.04**	14.35	-4.59%
HM	ResNet50	5.82	**6.67**	+14.60%	2.74	**3.26**	+18.98%
	Swin-T	6.27	**6.97**	+11.16%	2.84	**4.47**	+57.39%
Bili	ResNet50	2.67	**2.93**	+9.74%	1.07	**1.20**	+12.05%
	Swin-T	2.83	**3.18**	+12.37%	1.08	**1.46**	+35.19%

Third, we investigate whether the pre-trained ME produces higher recommendation accuracy than its training-from-scratch (TFS) version (i.e., with random initialization). There is no doubt that the pre-trained BERT, ResNet, and Swin largely improve corresponding NLP and CV tasks against their TFS versions. We report the recommendation results on the testing set in Table 4. It can be clearly seen that pre-trained MoRec obtain better final results. In particular, MoRec achieve around 10% improvements with the pre-trained ME (ResNet and Swin) on HM and Bili, which also aligns with findings in NLP and CV domains. We also construct the smaller version datasets by randomly drawing 50K users from MIND, HM, and Bili. It can be seen that the advantages of pre-trained ME over TFS are more obvious on small datasets. However, we found that the pre-trained $\text{BERT}_{\text{base}}$ is even worse than its TFS version on MIND-50K.

According to the above experiments, we conclude that **(Answer for Q(ii)) MoRec build connections for RS and other**

[8]Due to space reasons, we do not report results of other sequential models, but we have indeed evaluated them. The conclusions hold when using GRU4Rec, NextItNet, and BERT4Rec as backbones.

multimedia communities, and can in general inherit the latest advances from the NLP and CV fields. This is a very good property, which means that once there are new breakthroughs in the corresponding research fields in the future, MoRec have more opportunities and greater room to be improved.

6 ARE MODALITY REPRESENTATIONS UNIVERSAL FOR RS? (Q(III))

Foundation models in NLP and CV are expected to generate generic representation, which can then be directly used for downstream tasks in the zero-shot setting. However, most of them are only evaluated in some traditional tasks [32, 44], such as image and text classification. We argue that predicting user preference is more challenging than these objective tasks.

To see this problem clearly, we evaluate two training approaches. The first approach is to pre-extract modality features by ME and then add them into a recommendation model [19, 20], referred to as a two-stage (TS) pipeline. Due to the high training efficiency, TS is especially popular in real-world industrial applications, where there are usually hundreds of millions of training examples. The second approach is the one used in all above experiments, by optimizing user and item encoders simultaneously in an E2E manner.

As shown in Table 5, we find that TS-based MoRec show surprisingly poor results, compared to IDRec and E2E-based MoRec. In particular, with ResNet, it achieves only around 60% and 25% performance of E2E-based MoRec on HM and Bili, respectively. For better adaption, we also add many dense layers on top of these fixed modality features. As shown, this can indeed improve the performance of TS; however, it is still much worse than IDRec and E2E-based MoRec, especially for visual recommendation.

The results indicate that the modality features learned by these NLP and CV tasks are not universal enough for the recommendation problem, and thus the recommendation results are worse compared to retraining on new data (i.e., the E2E paradigm). The good thing is that by proper adaption (i.e., TS-DNN), TS-based MoRec have some potential to compete with E2E MoRec for text recommendation in the future (16.66 *vs* 18.23).

Thereby, we want to explicitly remind RS researchers and practitioners that **(Answer for Q(iii)) the popular two-stage recommendation mechanism leads to significant performance degradation (especially for image recommendation), which should not be ignored in practice.[9] Second, for NLP and CV researchers, we want to show them that, despite the revolutionary success of FM, until now their representation features are not universal enough, at least for item recommendation.**

7 KEY CHALLENGES (Q(IV))

E2E-based MoRec has been less studied before, especially for visual recommendation. Here, we present several key challenges and some unexpected findings that the community may not be aware of.

Training cost. As shown in Figure 4, MoRec with larger ME tend to perform better than smaller ME, however, the training compute, time and GPU memory consumption also increase, especially for the seq2seq-based architecture with very long interaction sequence.

[9]Unfortunately, so far, there is not even any literature showing that an E2E-based MoRec has been successfully deployed in real-world recommender systems.

Table 5: HR@10 (%) of E2E *vs* TS with additional MLP layers . 'TS-DNN 6' denotes that TS-based MoRec with 6 learnable MLPs layers on top of these fixed modality representation.

Dataset	IDRec	ME	TS	TS-DNN					E2E
				2	6	8	10	12	
MIND	17.71	$BERT_{base}$	13.93	15.20	16.26	<u>16.66</u>	16.32	16.14	**18.23**
HM	6.84	ResNet50	4.03	4.64	<u>5.40</u>	5.39	<u>5.40</u>	5.02	6.67
		Swin-T	3.45	4.46	5.28	<u>5.55</u>	5.40	5.38	**6.97**
Bili	3.03	ResNet50	0.72	1.23	<u>1.62</u>	1.47	1.28	1.24	2.93
		Swin-T	0.79	1.40	1.81	<u>2.10</u>	1.95	1.64	**3.18**

Table 6: The training cost. #Param: number of tunable parameters, FLOPs: computational complexity (we measure FLOPs with batchsize=1), Time/E: averaged training time for one epoch, 'm' means minutes, MU: GPU memory usage, e.g., 'V100-32G(2)' means that we used 2 V100s with 32G memory.

Dataset	Method	#Param.	FLOPs	Time/E	MU	GPU
MIND	IDRec	47M	0.12G	2.7m	3G	V100-32G(1)
	$BERT_{tiny}$	11M	0.63G	10m	4G	V100-32G(1)
	$BERT_{small}$	35M	16G	42m	13G	V100-32G(1)
	$BERT_{base}$	116M	107G	102m	52G	V100-32G**(2)**
HM	IDRec	114M	1G	4.3m	5G	V100-32G(1)
	ResNet18	18M	40G	95m	23G	V100-32G(1)
	ResNet34	29M	81G	136m	30G	V100-32G(1)
	ResNet50	31M	91G	83m	80G	V100-32G**(4)**
	Swin-T	34M	96G	107m	157G	A100-40G**(4)**
	Swin-B	94M	333G	102m	308G	A100-40G**(8)**

We report the training cost details on HM (close to Bili) and MIND in Table 6. In fact, it is not difficult to imagine that MoRec will consume more computing resources and time than IDRec. However, it is hard to imagine that the best MoRec (with SASRec as user encoder and Swin-B as ME) takes an astonishing more than 100x compute and training time than IDRec.[10] This has not been explicitly revealed in literature. This may also be the reason why there are no formal publications combining seq2seq user encoder and E2E-learned item ME for MoRec, especially for image recommendation. Note that in practice, it may not always be necessary to optimize all parameters of ME, and for some datasets, fine-tuning a few top layers of ME can achieve comparable results. On the other hand, although E2E-based MoRec is highly expensive[11] during training (akin to FM in NLP and CV), it has been shown to enable foundation recommendation models, which can free up more labor in training specific models [52, 53].

Extra pre-training. Performing a second round of pre-training for ME using the downstream dataset often works well in much machine learning literature [16, 54]. Here, we explore whether it offers improved results for MoRec. Following the pre-training of BERT, we adopt the "masked language model" (MLM) objective to train the text encoder of MoRec (denoted by $BERT_{base}$-MLM) on MIND and

[10]Note that the inference time of MoRec for online service is as fast as IDRec.

[11]This entire work has costed us over $140,000. For example, with 8 A100 GPUs, MoRec with Swin-B requires nearly 1 week to converge on HM, and the cost of purchasing the GPU service is about $2,000 (for one set of hyper-parameters).

Table 7: HR@10 (%) of co-training ID and modality. 'ADD' and 'CON' are two fusion methods. w/ and w/o denote whether to add extra MLP layers after the fusion layer. We search the layer number from $\{2, 4, 6, 8\}$. Adding extra DNN layers for 'ID+E2E' does not improve the accuracy, so we do not report them below for clarity. 'Improv.' means the relative improvement with ID+modality feaurs compared to the best result of pure IDRec and pure MoRec.

Dataset	ME	IDRec	TS	ID+TS				Improv.	TS-DNN	ID+TS-DNN				Improv.	E2E	ID+E2E		Improv.
				w/o		w/				w/o		w/				w/o		
				ADD	CON	ADD	CON			ADD	CON	ADD	CON			ADD	CON	
MIND	$BERT_{base}$	17.71	13.93	16.10	17.20	17.66	17.57	-0.28%	16.66	14.93	16.58	17.29	17.55	-0.90%	**18.23**	16.25	17.12	-6.09%
HM	Swin-T	6.84	3.45	5.75	4.89	5.37	5.40	-15.94%	5.55	5.27	4.00	4.77	5.11	-22.95%	**6.97**	5.40	4.95	-22.53%
Bili	Swin-T	3.03	0.79	3.01	2.61	3.02	2.86	-0.33%	2.10	2.86	2.35	2.50	2.72	-5.61%	**3.18**	2.94	2.55	-7.55%

report results in Table 8. As shown, $BERT_{base}$-MLM gains higher accuracy than $BERT_{base}$ for both the TS and E2E models. Similarly, we explore whether it holds for the vision encoder. Note that ResNet and Swin Transformer used in previous experiments are pre-trained in a supervised manner, but neither HM nor Bili contains supervised image labels. To this end, we turn to use MAE [18], a SOTA image encoder pre-trained in an unsupervised manner, similar to MLM. We find MAE_{base}-MLM clearly improves the standard MAE_{base} on HM with the TS model, but obtains marginal gains with the E2E model. By contrast, no accuracy improvements are observed on Bili. By examining image cases in Figure 2, we find that pictures in Bili have very diverse topics and are more challenging than HM (with only very simple fashion elements). Our conclusion is that the effectiveness of the second round of pre-training depends on individual datasets; more importantly, it seems difficult to achieve larger accuracy gains for the E2E MoRec.

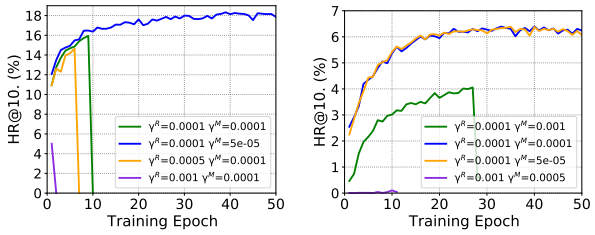
Combing ID & modality features. Given that IDRec and E2E-based MoRec both work well, a natural idea is to combine the two features (i.e., ID and modality) in one model. We have evaluated this, as shown in Table 7. We consider two types of feature combinations: additive and concatenated. Surprisingly, we find that neither TS- nor E2E-based MoRec is improved compared to the best results between IDRec and MoRec. By adding ID features, E2E-based MoRec perform even worse than pure IDRec and pure MoRec. Our results here are somewhat inconsistent with previous publications, which often claimed to achieve better results by adding modality or multimedia features for IDRec [19, 20, 29]. One reason might be that in the regular (*vs* cold-start) setting, both E2E-based MoRec and IDRec learn user preference from user-item interaction data, so they cannot complement each other, while for TS-based MoRec, since ID embeddings are too much better than frozen modality features, their combination also does not improve the results. The second reason may be that more advanced techniques are required when combining ID and modality features. In fact, from another point of view, MoRec with ID features will lose many advantages of MoRec (see Introduction). For example, with ID features MoRec are not suitable for building foundation recommendation models, because IDs are not easily transferable due to privacy and overlapping issues.[12]

Model collapse. Unlike IDRec, we find a very surprising phenomenon that training MoRec without proper hyper-parameters

[12]In this paper, we did not intend to study the effect of transfer learning, because a reliable pre-trained model requires a huge amount of training data and compute, see [52, 53].

Table 8: Comparison of HR@10 (%) w/ and w/o extra pre-training. 'Improv.' means the relative improvement of w/ extra pre-training compared to w/o extra pre-training.

Dataset	ME	TS			E2E		
		w/o	w/	Improv.	w/o	w/	Improv.
MIND	$BERT_{base}$	13.93	**14.68**	+5.38%	18.23	**18.63**	+2.19%
HM	MAE_{base}	2.50	**2.79**	+11.60%	7.03	**7.07**	+0.57%
Bili	MAE_{base}	0.57	0.57	0.00%	**3.18**	3.17	-0.31%



(a) MoRec with $BERT_{base}$ on MIND. (b) MoRec with Swin-T on HM.

Figure 5: Training collapse (on the validation set) with different learning rates γ. M and R denote learning rate for ME and the remaining modules, respectively.

(mainly the learning rate γ) can easily lead to model collapse. As shown in Figure 5, the performance of MoRec on MIND drops drastically from 16% to 0 when γ^M and γ^R are equal to 0.0001. Even worse, MoRec becomes collapsed from the beginning when $\gamma^M = 0.0001$ and $\gamma^R = 0.001$. Similarly, MoRec also have this problem when making image recommendation on HM. However, by carefully searching hyper-parameters, we find that MoRec can usually be trained well with a proper γ. It is worth noting that it is sometimes necessary to set different γ for item ME and other modules. This may be because item ME has been pre-trained on NLP and CV datasets before, and its learning stride may be different from other modules trained from scratch. By contrast, IDRec did not collapse even with many different γ. To the best of our knowledge, our findings here have not been reported in the literature.

8 RELATED WORK

ID-based recommender systems (IDRec). In the existing recommendation literature, there are countless models built entirely on

user/item ID, from early item-to-item collaborative filtering [35], shallow factorization models [28, 48], to deep neural models [21, 22]. They can be roughly divided into two categories: non-sequential models (NSM) and sequential neural models (SRM). NSM further includes various recall (e.g., DSSM, and YouTube DNN [7]) and CTR models (e.g., DeepFM [15], wide & Deep [6], and Deep Crossing [51]). These models typically take a user-item pair as input along with some additional features and predict matching scores between users and items. In contrast, a typical SRM takes a sequence of user-item interactions as input and generates the probability of the next interaction. The most representative SRM includes GRU4Rec [22], NextItNet [75, 76], SR-GNN [68], SASRec [25] & BERT4Rec [55] with RNN, CNN, GNN, Transformer & BERT as the backbone, respectively, among which SASRec often performs the best in literature [13, 77, 78].

Modality-based recommender systems (MoRec). MoRec focus on modeling the modality content features of items, such as text [67], images [39], videos [8], audio [58] and text-image multimodal pairs [65]. Previous work tended to adopt the two-stage (TS) mechanism by first pre-extracting item modality features from ME and then incorporating these fixed features into the recommendation model [19, 20, 30, 39, 51, 57, 62]. What's more, most of these work mainly use modality as side features and IDs as the main features. E2E-based MoRec is not popular until recently for several reasons: (1) the TS mechanism is architecturally very flexible for industrial applications and requires much lower compute and training cost; (2) there were few high-quality public datasets with original item modalities; (3) ME developed in past literature (e.g., word embedding) is not expressive enough even with E2E training. In the past two years, some works have begun to explore E2E-based MoRec, however, most of them focus on text recommendation [23, 52, 64, 69, 70, 72]. A recent preprint [12] introduced ResNet as ME for fashion-based recommendation but had to rely on ID features for competitive accuracy. To the best of our knowledge, none of these existing peer-reviewed literature provides an explicit and comprehensive comparative study of MoRec and its corresponding IDRec counterpart in a fair experimental setting (e.g., making sure they use the same backbone for comparison), especially in the non cold-start or even warm-start settings.

9 CONCLUSION AND FUTURE WORKS

In this paper, we investigated an ambitious but under-explored question, whether MoRec has the opportunity to end the dominance of IDRec. Obviously, this problem cannot be completely answered in one paper, and requires more study and efforts from the RS and even the NLP and CV communities. Yet, one major finding here is that with the SOTA and E2E-trained ME, *modern* MoRec could already perform on par or better than IDRec with the typical recommendation architecture (i.e., Transformer backbone) even in the non cold-start item recommendation setting. Moreover, MoRec can largely benefit from the technical advances in the NLP and CV fields, which implies that it has larger room for accuracy improvements in the future. Given this, we believe our research is meaningful and would potentially inspire more studies on E2E-based MoRec, for example, developing more powerful recommendation architectures (particular for CTR[13] prediction tasks), more expressive & generalized item encoders, better item & user fusion strategies and more effective optimizations to reduce the compute & memory costs and the longer training time. We also envision that in the long run the prevailing paradigm of RS may have a chance to shift from IDRec to MoRec when item raw modality features are available.

As mentioned above, this study is only a preliminary of MoRec and has several limitations: (1) we considered RS scenarios with only text and vision, whereas MoRec's behaviors with other modalities, e.g., voice and video, remain unknown; (2) we consider only single-modal item encoders, while the behaviors of multimodal MoRec are unknown; (3) we considered only a very basic approach to fusing ME into recommendation models, thereby MoRec may achieve sub-optimal performance; (4) our observations were made on three medium-sized dataset, and it remains unknown whether the key findings hold if we scale up training data to 100× or 1000× as in real industrial systems.

ACKNOWLEDGMENTS

This work is supported by the Research Center for Industries of the Future (No.WU2022C030) and the Key Research Project of Zhejiang Lab (No.2022PG0AC02).

A APPENDIX

A.1 MoRec *vs* IDRec on cold-start settings

Table 9: HR@10 (%) of IDRec and MoRec for cold and new item recommendation. m^{cold} and m^{new} denote the number of cold items and new items, respectively. All results are evaluated based on the SASRec architecture.

Dataset	ME	m^{cold}	IDRec	MoRec	m^{new}	IDRec	MoRec
MIND	$\text{BERT}_{\text{base}}$	32K	0.0036	**3.0637**	13K	0.0125	**0.5899**
HM	Swin-B	37K	0.3744	**1.0965**	14K	0.0115	**0.6846**
Bili	Swin-B	39K	0.3551	**0.6400**	5K	0.0078	**0.0832**

MoRec is a natural fit for cold item recommendation as their ME is specifically developed to model the raw modality features of an item, whether it is cold or not. To validate this, we evaluate IDRec and MoRec in two scenarios, i.e., COLD item setting and NEW item setting. Specifically, we counted the interactions of all items in the training set and regarded those that appeared less than 10 times as cold items. We found that the number of cold items were very small in our original testing test. So we performed dataset crawling again for one month and then selected user sequences (from this new dataset) that contained these cold items (as cold item setting) and items that did not appear in the training set (as new item setting). We report the results in Table 9. As expected, MoRec consistently and substantially improve IDRec on all three datasets for both text and vision modalities in both cold and new settings. The superiority of MoRec comes from the powerful representations of ME which were pre-trained on large-scale text and image datasets beforehand.

[13] In fact, we notice that the NLP/CV communities are formulating most tasks into sequence learning problem with Transformer as the backbone [46, 47], e.g., GPT-3 and pixelGPT [5]. It will be interesting to see whether complex CTR models with various user/item features can be formulated in a similar fashion (the way MoRec is powerful).

REFERENCES

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Hung-Wei Chen, Yi-Leh Wu, Maw-Kae Hor, and Cheng-Yuan Tang. 2017. Fully content-based movie recommender system with feature extraction using neural network. In *2017 International conference on machine learning and cybernetics (ICMLC)*, Vol. 2. IEEE, 504–509.

[4] Jiawei Chen, Hande Dong, Yang Qiu, Xiangnan He, Xin Xin, Liang Chen, Guli Lin, and Keping Yang. 2021. Autodebias: Learning to debias for recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 21–30.

[5] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *International conference on machine learning*. PMLR, 1691–1703.

[6] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.

[7] 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.

[8] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, Pietro Piazzolla, and Massimo Quadrana. 2016. Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics* 5, 2 (2016), 99–113.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Hao Ding, Yifei Ma, Anoop Deoras, Yuyang Wang, and Hao Wang. 2021. Zero-shot recommender systems. *arXiv preprint arXiv:2105.08318* (2021).

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[12] Shereen Elsayed, Lukas Brinkmeyer, and Lars Schmidt-Thieme. 2022. End-to-End Image-Based Fashion Recommendation. *arXiv preprint arXiv:2205.02923* (2022).

[13] Elisabeth Fischer, Daniel Zoller, and Andreas Hotho. 2021. Comparison of Transformer-Based Sequential Product Recommendation Models for the Coveo Data Challenge. (2021).

[14] Wenjing Fu, Zhaohui Peng, Senzhang Wang, Yang Xu, and Jin Li. 2019. Deeply fusing reviews and contents for cold start users in cross-domain recommendation systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 94–101.

[15] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).

[16] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964* (2020).

[17] Casper Hansen, Christian Hansen, Jakob Grue Simonsen, Stephen Alstrup, and Christina Lioma. 2020. Content-aware neural hashing for cold-start recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 971–980.

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

[19] Ruining He and Julian McAuley. 2016. 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*. 507–517.

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

[21] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

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

[23] Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2022. Towards Universal Sequence Representation Learning for Recommender Systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 585–593.

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

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

[26] Donghyun Kim, Chanyoung Park, Jinoh Oh, Sungyoung Lee, and Hwanjo Yu. 2016. Convolutional matrix factorization for document context-aware recommendation. In *Proceedings of the 10th ACM conference on recommender systems*. 233–240.

[27] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[28] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

[29] Maciej Kula. 2015. Metadata embeddings for user and item cold-start recommendations. *arXiv preprint arXiv:1507.08439* (2015).

[30] 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.

[31] Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. 2022. MINER: Multi-Interest Matching Network for News Recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*. 343–352.

[32] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10965–10975.

[33] Xiangyang Li, Bo Chen, HuiFeng Guo, Jingjie Li, Chenxu Zhu, Xiang Long, Sujian Li, Yichao Wang, Wei Guo, Longxia Mao, et al. 2022. IntTower: the Next Generation of Two-Tower Model for Pre-Ranking System. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3292–3301.

[34] Yutong Lin, Ze Liu, Zheng Zhang, Han Hu, Nanning Zheng, Stephen Lin, and Yue Cao. 2022. Could Giant Pretrained Image Models Extract Universal Representations? *arXiv preprint arXiv:2211.02043* (2022).

[35] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 7, 1 (2003), 76–80.

[36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

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

[38] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[39] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*. 43–52.

[40] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 26 (2013).

[41] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 188–197.

[42] Charles Packer, Julian McAuley, and Arnau Ramisa. 2018. Visually-aware personalized recommendation using interpretable image representations. *arXiv preprint arXiv:1806.09820* (2018).

[43] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[45] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[47] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175* (2022).

[48] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*. IEEE, 995–1000.

[49] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[50] Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. 2020. Neural collaborative filtering vs. matrix factorization revisited. In *Fourteenth ACM conference on recommender systems*. 240–248.

[51] Ying Shan, T Ryan Hoens, Jian Jiao, Haijing Wang, Dong Yu, and JC Mao. 2016. Deep crossing: Web-scale modeling without manually crafted combinatorial features. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 255–262.

[52] Kyuyong Shin, Hanock Kwak, Kyung-Min Kim, Minkyu Kim, Young-Jin Park, Jisu Jeong, and Seungjae Jung. 2021. One4all user representation for recommender systems in e-commerce. *arXiv preprint arXiv:2106.00573* (2021).

[53] Kyuyong Shin, Hanock Kwak, Kyung-Min Kim, Su Young Kim, and Max Nihlen Ramstrom. 2021. Scaling law for recommendation models: Towards general-purpose user representations. *arXiv preprint arXiv:2111.11294* (2021).

[54] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification?. In *China national conference on Chinese computational linguistics*. Springer, 194–206.

[55] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.

[56] Yang Sun, Fajie Yuan, Min Yang, Guoao Wei, Zhou Zhao, and Duo Liu. 2020. A generic network compression framework for sequential recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1299–1308.

[57] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. 2019. Adversarial training towards robust multimedia recommender system. *IEEE Transactions on Knowledge and Data Engineering* 32, 5 (2019), 855–867.

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

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[60] Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 448–456.

[61] Jie Wang, Fajie Yuan, Mingyue Cheng, Joemon M Jose, Chenyun Yu, Beibei Kong, Zhijin Wang, Bo Hu, and Zang Li. 2022. TransRec: Learning Transferable Recommendation from Mixture-of-Modality Feedback. *arXiv preprint arXiv:2206.06190* (2022).

[62] 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.

[63] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 6389–6394.

[64] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering news recommendation with pre-trained language models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1652–1656.

[65] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Mm-rec: multimodal news recommendation. *arXiv preprint arXiv:2104.07407* (2021).

[66] Chuhan Wu, Fangzhao Wu, Yang Yu, Tao Qi, Yongfeng Huang, and Qi Liu. 2021. NewsBERT: Distilling pre-trained language model for intelligent news application. *arXiv preprint arXiv:2102.04887* (2021).

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

[68] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 346–353.

[69] Shitao Xiao, Zheng Liu, Yingxia Shao, Tao Di, Bhuvan Middha, Fangzhao Wu, and Xing Xie. 2022. Training large-scale news recommenders with pretrained language models in the loop. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4215–4225.

[70] Yoonseok Yang, Kyu Seok Kim, Minsam Kim, and Juneyoung Park. 2022. GRAM: Fast Fine-tuning of Pre-trained Language Models for Content-based Collaborative Filtering. *arXiv preprint arXiv:2204.04179* (2022).

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

[72] Yang Yu, Fangzhao Wu, Chuhan Wu, Jingwei Yi, Tao Qi, and Qi Liu. 2021. Tiny-NewsRec: Efficient and Effective PLM-based News Recommendation. *arXiv preprint arXiv:2112.00944* (2021).

[73] Fajie Yuan, Guibing Guo, Joemon M Jose, Long Chen, Haitao Yu, and Weinan Zhang. 2016. Lambdafm: learning optimal ranking with factorization machines using lambda surrogates. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. 227–236.

[74] Fajie Yuan, Xiangnan He, Alexandros Karatzoglou, and Liguang Zhang. 2020. Parameter-efficient transfer from sequential behaviors for user modeling and recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 1469–1478.

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

[76] Fajie Yuan, Guoxiao Zhang, Alexandros Karatzoglou, Joemon Jose, Beibei Kong, and Yudong Li. 2021. One person, one model, one world: Learning continual user representation without forgetting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 696–705.

[77] 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).

[78] Mengqi Zhang, Shu Wu, Xueli Yu, Qiang Liu, and Liang Wang. 2022. Dynamic graph neural networks for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[79] Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. 2021. UNBERT: User-News Matching BERT for News Recommendation.. In *IJCAI*. 3356–3362.

[80] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).

[81] Jieming Zhu, Quanyu Dai, Liangcai Su, Rong Ma, Jinyang Liu, Guohao Cai, Xi Xiao, and Rui Zhang. 2022. Bars: Towards open benchmarking for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2912–2923.