




MMNet: Multi-Granularity Multi-Mode Network for Item-Level Share Rate Prediction

Haomin Yu , Mingfei Liang , Ruobing Xie , Zhenlong Sun, Bo Zhang, and Leyu Lin

WeChat Search Application Department, Tencent, China
haominyu@bjtu.edu.cn, aesopliang@tencent.com, xrbsnowing@163.com,
{richardsun, nevinzhang, goshawklin}@tencent.com

Abstract. Item-level share rate prediction (ISRP) aims to predict the future share rates for each item according to the meta information and historical share rate sequences. It can help us to quickly select high-quality items that users are willing to share from millions of item candidates, which is widely used in real-world large-scale recommendation systems for efficiency. However, there are several technical challenges to be addressed for improving ISRP’s performance: (1) There is data uncertainty in items’ share rate sequences caused by insufficient item clicks, especially in the early stages of item release. These noisy or even incomplete share rate sequences strongly restrict the historical information modeling. (2) There are multiple modes in the share rate data, including normal mode, cold-start mode and noisy mode. It is challenging for models to jointly deal with all three modes especially with the cold-start and noisy scenarios. In this work, we propose a multi-granularity multi-mode network (MMNet) for item-level share rate prediction, which mainly consists of a fine-granularity module, a coarse-granularity module and a meta-info modeling module. Specifically, in the fine-granularity module, a multi-mode modeling strategy with dual disturbance blocks is designed to balance multi-mode data. In the coarse-granularity module, we generalize the historical information via item taxonomies to alleviate noises and uncertainty at the item level. In the meta-info modeling module, we utilize multiple attributes such as meta info, contexts and images to learn effective item representations as supplements. In experiments, we conduct both offline and online evaluations on a real-world recommendation system in WeChat Top Stories. The significant improvements confirm the effectiveness and robustness of MMNet. Currently, MMNet has been deployed on WeChat Top Stories.

Keywords: Share rate prediction · multi-granularity · multi-mode.

1 Introduction

With the development of social media, people are becoming more enthusiastic about publishing their created contents and sharing their opinions on the inter-

Haomin Yu, Mingfei Liang and Ruobing Xie contribute equally to this work.

net, generating massive amounts of items. Users want to quickly obtain valuable information from massive items on social media platforms. Therefore, personalized recommendations are adopted to provide appropriate items effectively and efficiently for users to read and share.

Real-world large-scale recommendation systems should deal with millions of new items per day. It is essential to pre-select high-quality items for the following personalized matching and ranking modules in recommendation systems [16, 6] for efficiency. In this work, we propose the **Item-level Share Rate Prediction (ISRP)** task, which aims to predict the future share rates for each item according to their meta information and historical share rate sequences. It can help us to quickly find appropriate items that users are interested in from millions of item candidates, which could be viewed as an item quality inspector that is essential in real-world large-scale recommendation systems.

In recent years, to better grasp the development trend of items, many scholars have studied popularity prediction by inferring the total counts of interactions between users and items (e.g., view, click and share). The popularity prediction approaches can be roughly divided into two categories, including social-based prediction methods [3] and item-based prediction methods [4]. Item-based prediction methods generally utilize item-related meta information such as images and contexts to predict popularity [15]. Inspired by this, ISRP can be regarded as a special item-based popularity prediction task in recommendation systems, which focuses on predicting item share rates only with the item-related information. Moreover, we creatively bring in the historical share rate sequence for each item containing the average item share rate at each time period. However, there are some challenges in combining different meta information and historical share rate information for ISRP in practice:

- **Item-related data uncertainty.** In ISRP, there are mainly two types of item-related data uncertainty, including the share rate uncertainty and the attribute uncertainty. Share rate uncertainty mainly occurs in the early stage of item release, which is caused by the insufficient item clicks. In addition, the share rates of an item may fluctuate greatly during the whole period, which makes it difficult for the model to obtain high-confidence information from the historical share rate trends. This uncertainty locates in every item’s lifetime, since every item has a cold-start period and most items are long-tail. In contrast, attribute uncertainty derives from the noises or missing in item-related meta information. Therefore, it is essential to introduce an uncertainty eliminator to enable a robust ISRP framework.
- **Multi-mode share rate data.** In practice, there are multiple modes of the share rate data, including the normal mode, cold-start mode and noisy mode. *Normal mode* indicates that the share rate sequences are reliable with sufficient clicks, so the model can fully rely on historical share rates. In contrast, *cold-start mode* refers to the mode influenced by the share rate uncertainty in the early stages of item release. *Noisy mode* represents that the share rate sequences are disturbed by noises and even data missing. Due to the unbalance in three modes, ISRP models can be easily dominated by the

normal mode and is prone to over-rely on historical share rate data, losing the ability to model cold-start and noisy modes. Therefore, an intelligent multi-mode learner is needed to jointly handle all scenarios.

To address the above challenges, we propose a **Multi-granularity Multi-mode Network (MMNet)** for item-level share rate prediction. MMNet is composed of a coarse-granularity module, a fine-granularity module and a meta-info modeling module, where the first two modules aim to model the historical share rate sequences. Specifically, in the fine-granularity module, we design two disturbance blocks with different masking strategies to highlight all modes during training process. The coarse-granularity module is presented to alleviate share rate uncertainty by considering global preference features anchored by item taxonomies. The meta-info modeling module aims to introduce sufficient meta features to represent item information as a supplement to the historical share rate sequences. All three features are then combined for the ISRP task.

In experiments, we conduct extensive evaluations on three datasets with normal, cold-start and noisy modes. We also deploy MMNet on a widely-used recommendation system to evaluate its online effectiveness. In summary, the contributions of this work can be summarized as follows:

- We systematically highlight the challenges in real-world item-level share rate prediction, and propose a novel MMNet framework to address them.
- We design the multi-granularity share rate modeling to alleviate the uncertainty issues in cold-start and noisy scenarios, which helps to capture user preferences from both the global and local perspectives.
- We present a multi-mode modeling strategy in the fine-granularity module with dual disturbance blocks, which can jointly learn informative messages from all three modes to build a robust model in practice.
- MMNet achieves significant improvements in both offline and online evaluations. Currently, MMNet has been deployed on WeChat Top Stories, affecting millions of users.

2 Related Works

Time Series Modeling Techniques. Time series modeling techniques have been widely used in forecasting tasks. Autoregressive integrated moving average (ARIMA) [13] model, which is a classic statistical model in the time series field. However, this model requires the time series data stationary, or stationary after differencing steps. In recent years, various sequence modeling methods based on deep learning have emerged, such as recurrent neural network (RNN). Nevertheless, RNN suffers from gradient disappearance and explosion problems. To alleviate the problem, long short-term memory (LSTM) [9] and gated recurrent unit (GRU) [5] methods appeared. However, the inherently sequential nature of recurrent models limits the ability of parallelization ability. Temporal convolutional network (TCN) [1], which utilizes convolution algorithm to solving prediction problem. It can achieve parallelization computation. Since historical

information can provide a certain degree of guidance for ISRP tasks, we introduce the historical share rate sequences to improve ISRP’s performance.

Popularity Prediction. The popularity prediction task is generally to estimate how many attentions a given content will receive after it is published on social media. The task is mainly divided into two types: social-based prediction methods and item-based prediction methods.

The social-based methods aim to predict the popularity of item spread through social relationships in social networks. DeepCas [10] and DeepHawkers [2] are popularity prediction methods by modeling information cascade. DeepCas constructs a cascade graph as a collection of cascade paths that are sampled by multiple random walk processes, which can effectively predict the size of cascades. DeepHawkers learns the interpretable factors of Hawkers process to model information cascade. Cao et al.[3] proposed CoupledGNN, which uses two coupled graph neural networks to capture the interplay between nodes and the spread of influence. However, the social-based methods concentrate on the propagation on social networks, which have a great dependence on social relationships. This limits the application scenarios of these models.

In contrast, the item-based methods extract a large number of features related to contents for popularity prediction. UHAN [18] and NPP [4] design hierarchical attention mechanisms to extract representations of multi-modalities. Different from them, Wu et al. [14] and Liao et al. [12] paid more attention to the influence of temporal information on popularity prediction. The former utilizes neighboring temporal and periodic temporal information to learn sequential popularity in short-term and long-term popularity fluctuations. The latter leverages RNN and CNN to capture item-related long-term growth and short-term fluctuation. Xie [15] proposed a multimodal variational encoder-decoder (MMVED) framework, which is the most related model of our task. It introduces the uncertain factors as the randomness for the mapping from the multimodal features to the popularity. However, in ISRP, the item-related data uncertainty and multi-mode share rate data will strongly affect the performance of existing popularity prediction models. Consequently, we propose disturbance blocks based multi-mode modeling strategy with multi-granularity share rate modeling for ISRP.

3 Preliminary

In this section, we first introduce some important notions used in this work.

Share rate Given an item I , the share rate y_t at time period t is defined as its overall shared number r_t divided by its overall click (i.e., items being clicked by users) number p_t , as shown in the following formula:

$$y_t = \frac{r_t}{p_t} \times 100\%, \quad p_t > 0. \quad (1)$$

A smaller click number p_t will result in data uncertainty of the share rate.

Multi-mode data For the historical share rate sequences, we define three modes according to different scenarios as follows:

- **Normal mode.** It represents that historical share rate sequences are reliable with sufficient clicks. The sequences can provide strong guidance.
- **Cold-start mode.** It indicates that whole historical share rate sequences are unreliable or even missing. This mode is usually caused by insufficient clicks, especially in the early stages of item release.
- **Noisy mode.** It means that there is partial uncertainty in the historical share rate sequences, which is usually caused by partial missing or noises.

Input features We deploy MMNet on a video recommendation system. The input features we use in MMNet can be mainly grouped into three categories, namely the fine-grained sequential features, the coarse-grained sequential features and the multi-modal meta-information features.

- **Fine-grained sequential features.** We calculate the share rates at all time period for each item, and arrange them into historical share rate sequences.
- **Coarse-grained sequential features.** To improve the generalization ability and reduce potential uncertainty in item-level share rates, we further bring in the coarse-grained sequential features. Precisely, we build the share rate sequences for each taxonomy (e.g., tag, category) in this item, modeling the share rate trends at the taxonomy level.
- **Multi-modal meta-information features.** These features consist of three heterogeneous parts: context features, visual features and meta features. The first part contains textual features such as item title. The second part regards the cover images as the visual features. The last part is composed of many meta information including video taxonomies and duration.

Item-level share rate prediction Formally, given the multi-modal feature set \mathbf{C} , and the historical share rate sequence $\{y\}_{t-\omega_1}^t$ with a time window of length ω_1 , our goal at time t is to predict the share rate \hat{y}_{t+h} at the next h time as:

$$F(\mathbf{C}, \{y\}_{t-\omega_1}^t) \rightarrow \hat{y}_{t+h}, \quad (2)$$

where \hat{y}_{t+h} is the predicted share rate at $t+h$, and h is the desirable prediction horizon time stamp. In most situations, the horizon h of share rate prediction task is chosen according to the practical demands of the real-world scenario. $F(\cdot)$ is the mapping function we aim to learn via MMNet.

4 Methodology

4.1 Overall Framework

Figure 1 shows the overall framework of MMNet. It mainly consists of three parallel modules, including a fine-granularity module, a coarse-granularity module,

and a meta-info modeling module. The fine-granularity module conducts a multi-mode strategy with two disturbance blocks to enable a robust share rate sequence modeling. The coarse-granularity module models the coarse-grained share rate sequential information brought by the corresponding item’s taxonomies, which can alleviate potential noises and missing in item-level share rate sequences. The meta-info modeling module further combines heterogeneous item meta features together. All three modules are then combined and fed into a gated fusion layer and a MLP (multi-layer perceptron) layer for the following prediction.

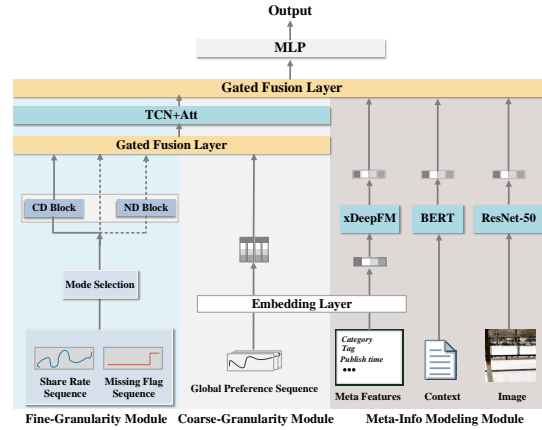


Fig. 1. Overall framework of the proposed MMNet.

4.2 Fine-Granularity Module

The fine-granularity module is responsible for encoding historical share rate sequences. However, there are two differences between the numerical share rate sequence in ISRP and other sequences (e.g., item sequences in session-based recommendation), which leads to the following challenges: (1) the share rate sequences are numerical sequences that suffer from data uncertainty and high variance caused by insufficient item clicks. (2) The fine-granularity module should jointly deal with three modes including the normal, cold-start and noisy modes. We conduct the multi-mode modeling to address these issues.

Multi-mode modeling In real-world scenarios, the multi-mode data are often unbalanced, and the normal mode data are far more than other two modes. Thus, if we directly use the original share rate sequence instances to train our MMNet, the model will be overfitting on the historical share rate information, regardless of other meta-information. Although the model can well predict the share rates of normal mode data with sufficient clicks and historical information,

it cannot deal with items in cold-start and noisy scenarios, which heavily rely on MMNet for item pre-selection in real-world recommendation systems. Therefore, for all historical sequences during training process, we randomly feed them into three modes followed by different disturbance blocks with equal probability. To simulate different mode sequences, the proposed multi-mode modeling strategy introduces two disturbance blocks, including a cold-start disturbance block (CD block) and a noisy disturbance block (ND block), as shown in 2. Note that this strategy can be regarded as a form of data augmentation. To better represent the state of the historical sequence, we introduce a missing flag sequence $\mathbf{m} = \{m\}_{t-\omega_1}^t$, where $m \in \{0, 1\}$. If the missing flag is 1, it means that the data is missing or uncertain, and otherwise, it means that the data is normal.

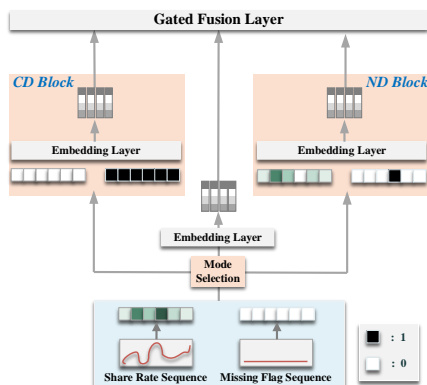


Fig. 2. The illustration of CD and ND disturbance blocks based multi-mode modelling.

Cold-start disturbance block. CD block is in charge of simulating cold-start mode data. This block can make the model learn more comprehensive features and alleviate the model’s excessive dependence on historical sequences. Formally, we exploit an all-one vector mask $\mathbf{m}^p = \{m^p\}_{t-\omega_1}^t$ ($m^p \in \{1\}$) as a missing flag vector, which means whole share rate sequence is missing or uncertain. Thus, the corresponding share rate sequence is erased by zero vector $\{y\}_{t-\omega_1}^t = \vec{\mathbf{0}}$. It aims to lead the model to focus more on other information rather than the share rates to improve the performance of all modes.

Noisy disturbance block. Similar to the CD block, we also design a ND block, which is responsible for simulating partial data uncertainty. For each missing flag in the sequence, we randomly sample a value T from the uniform distribution $U[0, 1]$, and then set a threshold τ . When T is greater than τ , the missing flag is set as 0, otherwise, it is set to 1. Note that τ can be regarded as a missing rate. When τ is large, there are more missing. Considering the input sequence also contains missing data, we should keep the missing data of origin input sequence unchanged. Thus, the final missing flag sequence is $\mathbf{m}^c = \mathbf{m}^c \vee \mathbf{m}$,

where \vee represents logical or. Consequently, the corresponding input share rate sequence \mathbf{y} is reset through $\mathbf{y} = \mathbf{m}^c \odot \mathbf{y}$, where \odot denotes Hadamard product. After the input sequence is processed by a mode, it is sent into the embedding layer $Emb(\cdot)$ [12] to obtain the item-level share rate representation sequence $\{\mathbf{h}^m\}_{t-\omega_1}^t$ as:

$$\{\mathbf{h}^m\}_{t-\omega_1}^t = Emb(CD(\{y\}_{t-\omega_1}^t) \text{ or } ND(\{y\}_{t-\omega_1}^t) \text{ or } \{y\}_{t-\omega_1}^t). \quad (3)$$

4.3 Coarse-Granularity Module

The fine-granularity module focuses on the historical share rates at the item level, which is precise but noisy due to the possible insufficient clicks and even data missing. Hence, we build the coarse-granularity module as a supplement, which is in charge of encoding the coarse-grained sequential information at the taxonomy level. A temporal mining layer is designed to encapsulate the trend information from both coarse and fine sequential information in two modules.

Global Preference Features Users have different priori preferences on different taxonomies. For example, considering the difference attractions of the item categories, we analyze the share rates of different categories in our system. As shown in Fig. 3, there are significant differences in the share rates of different categories (e.g., health-related videos have the highest share rate). Moreover, items with the same taxonomies (e.g., tags, categories) may have similar share rate trends. For instance, during the World Cup, the share rates of football-related videos generally grow higher than others. Therefore, it is essential to consider the share rate trends at the taxonomy level as a supplement to the item level.

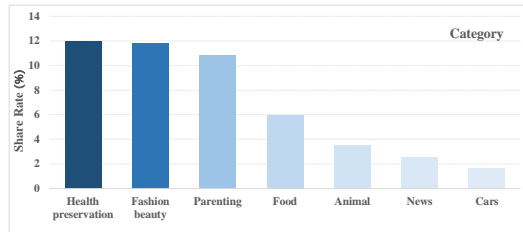


Fig. 3. The illustration of the global average share rates of different item categories.

Specifically, we introduce global preference features as a priori and generalized anchor to give coarse-granularity guidance for ISRP. Taking category for example, given an item I belonging to Category $Cat(I)$, the category global preference $g_t^{Cat(I)}$ at time t is obtained by calculating the global average share rate of whole item collection belonging to the $Cat(I)$ category:

$$g_t^{Cat(I)} = Mean_{\hat{I} \in C(Cat(I))} \{y_t(\hat{I})\}, \quad (4)$$

where $C(Cat(I))$ indicates the collection of all items which belongs to Category $Cat(I)$, \hat{I} represents an item in collection $C(Cat(I))$, $y_t(\hat{I})$ stands for the share rate of item \hat{I} , and $Mean(\cdot)$ represents the average operation. Note that we calculate the global preferences for all time periods to build the global category-level share rate sequence. Other taxonomies' modeling is the same as the category's. Next, at each time step t , we acquire global preference features \mathbf{g}_t , where $\mathbf{g}_t = Concat\{g_t^{Cat(I)}, g_t^{Tag(I)}, \dots\}$. Similarly, we use the same embedding layer on the global feature sequence $\{\mathbf{g}\}_{t-\omega_1}^t$ to obtain the taxonomy-level share rate representation sequence as follows:

$$\{\mathbf{h}^g\}_{t-\omega_1}^t = Emb(\{\mathbf{g}\}_{t-\omega_1}^t). \quad (5)$$

Temporal Mining Layer Temporal mining layer is responsible for encoding both fine and coarse sequential information from item-level share rate representation $\{\mathbf{h}^m\}_{t-\omega_1}^t$ and taxonomy-level share rate representation $\{\mathbf{h}^g\}_{t-\omega_1}^t$. To better balance fine and coarse representations, we utilize an attention mechanism [17] to obtain the aggregated representation as:

$$\{\mathbf{h}^{mg}\}_{t-\omega_1}^t = Att(\{\mathbf{h}^m\}_{t-\omega_1}^t, \{\mathbf{h}^g\}_{t-\omega_1}^t). \quad (6)$$

To reveal the inherent regularity and encapsulate the share rate trend information of historical information, we adopt a temporal convolutional network (TCN) [1] to learn the final sequential representation $\{\mathbf{h}^q\}_{t-\omega_1}^t$ as:

$$\{\mathbf{h}^q\}_{t-\omega_1}^t = TCN(\{\mathbf{h}^{mg}\}_{t-\omega_1}^t). \quad (7)$$

Then, we utilize a temporal attention mechanism on the sequence $\{\mathbf{h}^q\}_{t-\omega_1}^t$ to automatically learn the impacts of share rate representations at different times:

$$\mathbf{q} = Temporal_Att(\{\mathbf{h}^q\}_{t-\omega_1}^t), \quad (8)$$

where \mathbf{q} is the final historical share rate representation we use for prediction.

4.4 Meta-info Modeling Module

Meta-info modeling module is exploited to capture multi-modal features from heterogeneous item profiles. Multi-modal information can introduce complementary information for ISRP, thereby alleviating attribute uncertainty.

Specifically, this module is responsible for extracting interactive meta feature representations, context representations and visual representations:

- **Meta features:** there are potential relationships between different meta features. To capture effective feature interactions, we feed them into xDeepFM [11] for extracting the interactive representation \mathbf{u} .
- **Context features:** we directly use a pre-trained BERT [7], and acquire a context representation \mathbf{c} from the input contexts (e.g., video titles).
- **Visual features:** we use a pre-trained ResNet-50 [8], which utilizes skip connections, or shortcuts to jump over some layers. Precisely, we feed the cover image into a ResNet-50 model, and obtain the visual representation \mathbf{v} .

These features are essential especially in the cold-start and noisy scenarios.

4.5 Optimization Objectives

We jointly consider the share rate sequential representation \mathbf{p} , meta representation \mathbf{u} , context representation \mathbf{c} , and visual representation \mathbf{v} in ISRP. To automatically determine the influence of these representations, We concatenate these features, and send them into a gated fusion layer as follows:

$$\mathbf{h}^{\text{fuse}} = \text{Gating}([\mathbf{p}; \mathbf{u}; \mathbf{c}; \mathbf{v}]), \quad (9)$$

where $\text{Gating}(\cdot)$ is similar as the attention mechanism in [17]. \mathbf{h}^{fuse} represents the aggregated feature which encloses multi-modal information. Finally, we feed the aggregated feature \mathbf{h}^{fuse} into a MLP layer to generate the predicted share rate \hat{y}_{t+h} at the $t+h$ time as follows:

$$\hat{y}_{t+h} = \text{MLP}(\mathbf{h}^{\text{fuse}}). \quad (10)$$

In this work, we optimize the proposed MMNet by minimizing a mean square error (MSE) between the predicted and real share rates \hat{y}_{t+h} and y_{t+h} as follows:

$$\text{MSE} = \frac{1}{N} \sum_{t=1}^N (\hat{y}_{t+h} - y_{t+h})^2, \quad (11)$$

where N is the number of training samples. Note that it is also convenient to transfer MMNet to other rate prediction tasks (e.g., CTR, complete rate).

5 Online Deployment

We have deployed our MMNet model on a well-known real-world recommendation system, which is widely used by millions of users per day. This online system should deal with massive numbers of new items generated everyday. Therefore, based on the classical two-stage recommendation framework containing matching (i.e., candidate generation) and ranking modules introduced in [6], we further deploy MMNet on the pre-matching module to judge item quality according to item’s meta-information and historical behaviours for efficiency. The predicted share rates of each item candidate are used in two manners: (1) we directly filter low-quality items according to a low-standard rule-based threshold, and (2) the predicted share rates are fed into ranking modules as features. For efficiency, we use 200 workers equipped with 1 core and 8GB memory for online inference. The source code is in <https://github.com/MingFL/MMNET>.

6 Experiments

6.1 Datasets

To thoroughly evaluate the performance of our methods, we build an online video dataset from a widely-used recommendation system named WeChat Top Stories.

The dataset contains nearly 40 million share instances on 35 thousand items. All data are pre-processed via data masking for user privacy. We divide the dataset by a ratio of 8:1:1 for training, validation and testing. Due to the uncertainty in share rates, we discard all items that have low clicks in the test set, for we want all instances in the test set to have high confidence. This test setting is named as the *Normal Dataset*, since it mainly contains items with sufficient clicks and reliable historical share rates. To further investigate the model abilities for multi-mode data, the normal dataset is further processed into two other datasets, namely the *Cold-start dataset* and the *Noisy dataset*. To simulate the cold-start mode where all historical share rate data is unreliable or empty, we mask out all historical share rates on the normal dataset for generating the cold-start dataset. Similarly, in order to verify that the model deals with the data of noisy mode, we mask out the historical data with a certain probability.

6.2 Baselines and Experimental Settings

Baselines. The main contributions of MMNet locate in the share rate sequence modeling. Therefore, we compare MMNet with five competitive baselines in the share rate modeling. For fair comparisons, all baselines also contain the same meta-info modeling module, where the encoding of multi-modal features is consistent with MMNet (i.e., BERT processing context features, ResNet-50 processing visual features, and xDeepFM processing meta features). All models including MMNet and baselines share the same input features. We have:

- **HA.** Historical average (HA) is a straightforward method. Here, we use the average share rate value of the most recent 6 time periods (the same as MMNet) to predict the share rates in the next horizon time.
- **GRU.** Gated recurrent unit [5] is a classical model that can alleviate the problem of vanishing gradient in RNN. It performs well in solving time series forecasting problems.
- **Encoder-Decoder.** The encoder-decoder model [5] is a classical sequence modeling method, which is widely utilized in real-world tasks.
- **MMVED.** Multimodal variational encoder-decoder framework [15] is designed for sequential popularity prediction task, which considers the uncertain factors as randomness for the mapping from the multimodal features to the popularity.
- **DFTC.** The approach of deep fusion of temporal process and content features [12] is utilized in online article popularity prediction. It utilizes RNN and CNN to capture long-term and short-term fluctuations, respectively.

Ablation settings. Furthermore, to verify the advantages of each component of MMNet, we conduct four ablation versions of MMNet implemented as follows:

- **MMNet-M.** It is an incomplete MMNet, in which the multi-mode modeling strategy is removed, in order to verify the multi-mode modeling influence.

- **MMNet-C**. It is an incomplete MMNet, in which the coarse-granularity module is removed on the basis of MMNet-M, in order to verify the role of global preference features on three modes.
- **MMNet_{norm/noisy}**. It is a variant of MMNet, which lets the historical sequence select the normal mode and the noisy mode with equal probability without considering the cold-start mode.
- **MMNet_{norm/cold}**. It is a variant of MMNet, which lets the historical sequence select the normal mode and the cold-start mode with equal probability without considering the noisy mode.

Experimental settings. The proposed method is implemented with Tensorflow. The learning rate is set as 0.001, the batch size is set as 64, and the model is trained by minimizing the mean squared error function. The historical sequence window lengths (i.e. ω_1) of the share rates and global preference features are set to 6. In MMNet, the representations after embedding layer are all set as 64, including sequence representation, visual representation, context representation and meta representation. For TCN, we set three channels, and the hidden layers of these channels are 256, 128 and 64 respectively. Meanwhile, the size of the convolution kernel in TCN is 2. The missing rate τ is set to be 0.5, and the parameter sensitivity experiment of τ can be seen in Sec. 6.6.

6.3 Offline Item-level Share Rate Prediction

Evaluation Protocol We adopt two representative evaluation metrics for ISRP, including mean squared error (MSE) and precision@N% (P@N%). (1) MSE is a classical metric that is calculated by the average squared error between predicted and real share rates. It aims to measure the ability of MMNet in predicting share rates. (2) As for P@N%, we first rank all items in the test set via their predicted share rates, and then calculate the precision of top N% items as P@N%. It reflects the real-world performance of ISRP in recommendation systems. To simulate the practical settings, we report P@5% and P@10% in evaluation.

Experimental Results Table 1 presents the offline ISRP results of all models. We analyze the experimental results in details:

(1) MMNet achieves the best overall performance on all three datasets. The improvements of three metrics on the cold-start/noisy datasets, and the improvement of MSE on the normal dataset are significant with the significance level $\alpha = 0.01$. Since the proposed multi-granularity multi-mode strategy mainly aims to solve the cold-start and noisy issues, it is natural that the improvements on the cold-start and noisy datasets are much more significant. It indicates that MMNet can well deal with all three scenarios in ISRP, especially in the cold-start and noisy scenarios.

(2) Comparing with baselines, we find that the results of baselines are not ideal in cold-start and noisy datasets. It is because that the multi-mode data is not balanced, where the normal mode is the dominating mode. Therefore, most baselines are strongly influenced by the normal mode data during training. In

contrast, our MMNet is armed with the multi-granularity sequence modeling that can alleviate the cold-start and low click issues. Moreover, the multi-mode modeling also brings in robustness for these two scenarios. It can be regarded as a certain data argumentation, which can improve both the generalization ability of the share rate sequence modeling as well as the feature interactions between sequential and meta information in different scenarios.

Table 1. Calibration results for three datasets.

Method	Normal Dataset			Cold-start Dataset			Noisy Dataset		
	MSE	P@5%	P@10%	MSE	P@5%	P@10%	MSE	P@5%	P@10%
HA	1.650	0.919	0.920	42.143	0.054	0.101	7.174	0.747	0.791
GRU	0.260	0.975	0.973	16.410	0.219	0.293	3.299	0.688	0.761
Encoder-Decoder	0.256	0.975	0.973	15.663	0.233	0.291	3.990	0.656	0.725
MMVED	1.431	0.882	0.880	22.072	0.052	0.108	2.109	0.840	0.847
DFTC	0.257	0.976	0.974	14.980	0.284	0.348	3.677	0.760	0.786
MMNet	0.149	0.977	0.976	3.442	0.755	0.786	0.175	0.968	0.969

6.4 Online A/B Tests

Evaluation Protocol To further evaluate MMNet in practice, we deploy our model on a real-world recommendation system as introduced in Sec. 5. Specifically, MMNet is deployed in the pre-matching module and predicts item-level share rates for all items, which is used as (1) a coarse filter, and (2) features for the next matching and ranking modules. We conduct an online A/B test with other modules unchanged. The online base model is an ensemble of some rule-based filterers. In this online A/B test, we focus on two metrics: (1) average item-level share rate (AISR), (2) average dwell time per user (ADT/u).

Similarly, we further transfer the idea of MMNet on ISRP to the item-level complete rate prediction task. The complete rate is calculated by *user-finished duration divided by video’s full duration*, which reflects the qualities of items from another aspect. Precisely, we build a similar MMNet model with different parameters, and train it under the supervision of item-level complete rates of videos. We deploy this MMNet as in Sec. 5, and focus on (1) average dwell time per user (ADT/u), and (2) average dwell time per item (ADT/i). We conduct this online A/B test for 14 days, affecting nearly 6 million users.

Table 2. Online A/B tests on a real-world recommendation system.

Settings	supervised by share rates		supervised by complete rates	
	AISR	ADT/u	ADT/u	ADT/i
MMNet	+0.91%	+0.93%	+1.02%	+1.31%

Experimental Results Table 2 shows the relative improvements of MMNet over the online base model, from which we can observe that:

(1) MMNet achieves significant improvements in both item-level share rate and average dwell time. It indicates that our MMNet can well capture multi-granularity features, distinguish multi-mode share rate sequences, and combine multi-modal features for all normal, cold-start and noisy scenarios in ISRP.

(2) The successes in MMNet supervised by complete rates verify that our proposed framework is robust and easy to transfer to other scenarios.

6.5 Ablation Studies

Table 3 lists the results of the above-mentioned ablation settings with MSE, P@5% and P@10%. Note that since the multi-granularity and multi-mode modeling are mainly designed for the cold-start and noisy scenarios, we focus on these two datasets in ablation studies. We can observe that:

(1) MMNet achieves the best performance on all metrics in the noisy dataset and normal dataset, and the second best performance in the cold-start scenario. It verifies that all components in MMNet are essential in ISRP.

(2) Comparing with MMNet-C and MMNet-M, we can find that the global preference features are more suitable for cold-start and noisy scenarios. Meanwhile, the results also show that multi-granularity and multi-mode modeling are effective in capturing informative messages for all three modes in ISRP.

(3) Comparing with $\text{MMNet}_{norm/noisy}$ and $\text{MMNet}_{norm/cold}$, we find that both CD and ND disturbance blocks are effective for the cold-start and noisy scenarios respectively. It is worth noting that $\text{MMNet}_{norm/cold}$ focuses on the cold-start mode, so it is natural that it has better cold-start performance. In practice, we can flexibly set the weights of different disturbance blocks for specific motivations.

Table 3. Ablation study results for three datasets.

Method	Normal Dataset			Cold-start Dataset			Noisy Dataset		
	MSE	P@5%	P@10%	MSE	P@5%	P@10%	MSE	P@5%	P@10%
MMNet-M	0.248	0.974	0.974	14.594	0.365	0.406	0.633	0.931	0.938
MMNet-C	0.671	<u>0.975</u>	0.975	15.369	0.316	0.373	1.537	0.907	0.922
$\text{MMNet}_{norm/noisy}$	0.210	0.974	0.973	11.944	0.466	0.503	<u>0.315</u>	<u>0.964</u>	<u>0.966</u>
$\text{MMNet}_{norm/cold}$	<u>0.165</u>	0.973	0.976	2.440	0.802	0.831	0.836	0.888	0.925
MMNet	0.149	0.977	0.976	<u>3.442</u>	<u>0.755</u>	<u>0.786</u>	0.175	0.968	0.969

6.6 Parameter Analyses

We further study the parameter sensitivity of MMNet. We vary the missing rate τ from 0.01 to 0.9, which is essential in model training. The results are reported in

Table 4, from which we can find that: (1) The results of the parameter changes are relatively stable on the normal dataset. (2) In the cold-start dataset, the performance gradually improves as the missing rate increases. The main reason is that the missing rate is higher, and the data in the noisy dataset and the cold-start dataset will be more similar. (3) In the noisy dataset, as the missing rate increases, the performance has a gradual improvement followed by a slight decrease. The size of the missing rate can reflect the model’s dependence on historical data to a certain extent, so it can be concluded that the appropriate dependence on historical data is helpful to the model performance improvement. We select $\tau = 0.5$ according to the overall performance on three modes.

Table 4. Parameter analysis with different missing rates τ .

Method	Normal Dataset			Cold-start Dataset			Noisy Dataset		
	MSE	P@5%	P@10%	MSE	P@5%	P@10%	MSE	P@5%	P@10%
0.01	0.594	0.974	0.975	4.073	0.742	0.783	0.883	0.945	0.956
0.05	0.156	0.977	0.976	3.453	0.757	0.789	0.287	0.959	0.961
0.1	0.161	0.975	0.975	3.585	0.749	0.785	0.252	0.963	0.963
0.2	0.160	0.977	0.975	3.535	0.751	0.784	0.238	0.966	0.966
0.3	0.154	0.977	0.976	3.441	0.759	0.789	0.226	0.968	0.969
0.4	0.159	0.975	0.975	3.607	0.747	0.783	0.176	0.966	0.968
0.5	0.149	0.977	0.976	3.442	0.755	0.786	0.175	0.968	0.969
0.6	0.154	0.975	0.975	3.184	0.769	0.796	0.223	0.966	0.969
0.7	0.148	0.975	0.975	2.919	0.779	0.813	0.217	0.967	0.968
0.8	0.152	0.974	0.974	2.744	0.790	0.815	0.223	0.964	0.967
0.9	0.150	0.975	0.975	2.346	0.813	0.837	0.224	0.964	0.965

7 Conclusion and Future Work

In this paper, we present MMNet for ISRP. We propose a multi-granularity sequence modeling to improve the generalization ability from item taxonomies. Moreover, we design two multi-mode disturbance blocks to enhance the robustness of MMNet against potential data noises and uncertainty. Both offline and online evaluations confirm the effectiveness and robustness of MMNet in WeChat Top Stories. In the future, we will design an adaptive mode selection strategy based on the characteristics of the instance itself, so as to fully learn feature representations from existing instances. We will also explore more sophisticated feature interaction modeling between all types of features.

References

1. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271(2018)

2. Cao, Q., Shen, H., Cen, K., Ouyang, W., Cheng, X.: Deephawkes: Bridging the gap between prediction and understanding of information cascades. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp.1149–1158 (2017)
3. Cao, Q., Shen, H., Gao, J., Wei, B., Cheng, X.: Popularity prediction on social platforms with coupled graph neural networks. In: Proceedings of the 13th International Conference on Web Search and Data Mining. pp. 70–78 (2020)
4. Chen, G., Kong, Q., Xu, N., Mao, W.: Npp: A neural popularity prediction model for social media content. *Neurocomputing*333, 221–230 (2019)
5. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
6. Covington, P., Adams, J., Sargin, E.: Deep neural networks for youtube recommendations. In: Proceedings of RecSys (2016)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805(2018)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation*9(8), 1735–1780 (1997)
10. Li, C., Ma, J., Guo, X., Mei, Q.: Deepcas: An end-to-end predictor of information cascades. In: Proceedings of the 26th international conference on World Wide Web. pp. 577–586 (2017)
11. Lian, J., Zhou, X., Zhang, F., Chen, Z., Xie, X., Sun, G.: xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1754–1763 (2018)
12. Liao, D., Xu, J., Li, G., Huang, W., Liu, W., Li, J.: Popularity prediction on online articles with deep fusion of temporal process and content features. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 200–207 (2019)
13. Makridakis, S., Hibon, M.: Arma models and the box-jenkins methodology. *Journal of Forecasting*16(3), 147–163 (1997)
14. Wu, B., Cheng, W.H., Zhang, Y., Huang, Q., Li, J., Mei, T.: Sequential prediction of social media popularity with deep temporal context networks. arXiv preprint arXiv:1712.04443 (2017)
15. Xie, J., Zhu, Y., Zhang, Z., Peng, J., Yi, J., Hu, Y., Liu, H., Chen, Z.: A multi-modal variational encoder-decoder framework for micro-video popularity prediction. In: Proceedings of The Web Conference 2020. pp. 2542–2548 (2020)
16. Xie, R., Qiu, Z., Rao, J., Liu, Y., Zhang, B., Lin, L.: Internal and contextual attention network for cold-start multi-channel matching in recommendation. In: Proceedings of IJCAI (2020)
17. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. pp. 1480–1489 (2016)
18. Zhang, W., Wang, W., Wang, J., Zha, H.: User-guided hierarchical attention network for multi-modal social image popularity prediction. In: Proceedings of the 2018 world wide web conference. pp. 1277–1286 (2018)