DVR: Micro-Video Recommendation Optimizing Watch-Time-Gain under Duration Bias

Yu Zheng
Department of Electronic Engineering, Tsinghua University
Beijing, China

Chen Gao*
Department of Electronic Engineering, Tsinghua University
Beijing, China

Jingtao Ding
Department of Electronic Engineering, Tsinghua University
Beijing, China

Lingling Yi
WeChat Technical Architecture Department, Tencent Inc.
Shenzhen, China

Depeng Jin
Department of Electronic Engineering, Tsinghua University
Beijing, China

Yong Li
Department of Electronic Engineering, Tsinghua University
Beijing, China

Meng Wang
School of Computer Science and Information Engineering, Hefei University of Technology
Hefei, China

ABSTRACT

Recommender systems are prone to be misled by biases in the data. Models trained with biased data fail to capture the real interests of users, thus it is critical to alleviate the impact of bias to achieve unbiased recommendation. In this work, we focus on an essential bias in micro-video recommendation, duration bias. Specifically, existing micro-video recommender systems usually consider *watch time* as the most critical metric, which measures how long a user watches a video. Since videos with longer duration tend to have longer watch time, there exists a kind of *duration bias*, making longer videos tend to be recommended more against short videos. In this paper, we empirically show that commonly-used metrics are vulnerable to duration bias, making them NOT suitable for evaluating micro-video recommendation. To address it, we further propose an unbiased evaluation metric, called **WTG** (short for ***W**atch **T**ime **G**ain*). Empirical results reveal that WTG can alleviate duration bias and better measure recommendation performance. Moreover, we design a simple yet effective model named **DVR** (short for ***D**ebiased **V**ideo **R**ecommendation*) that can provide unbiased recommendation of micro-videos with varying duration, and learn unbiased user preferences via adversarial learning. Extensive experiments based on two real-world datasets demonstrate that DVR successfully eliminates duration bias and significantly improves recommendation performance with over 30% relative progress. Codes and datasets are released at https://github.com/tsinghua-fib-lab/WTG-DVR.

CCS CONCEPTS

• **Information systems → Personalization**.

*Corresponding author (chgao96@gmail.com).

This work is licensed under a Creative Commons Attribution International 4.0 License.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9203-7/22/10.
https://doi.org/10.1145/3503161.3548428

KEYWORDS

Recommendation, micro-video, duration bias, fairness

ACM Reference Format:
Yu Zheng, Chen Gao, Jingtao Ding, Lingling Yi, Depeng Jin, Yong Li, and Meng Wang. 2022. DVR: Micro-Video Recommendation Optimizing Watch-Time-Gain under Duration Bias. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Oct. 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3503161.3548428

1 INTRODUCTION

Today's micro-video platforms, such as TikTok[1], have been taking the majority of Internet traffic. With millions of micro-videos uploaded per day, recommender systems have become the fundamental channel that users access micro-videos [6, 9, 10, 14, 15, 19, 20, 39, 49, 50, 56]. Existing approaches usually consider watch time as a critical index of user satisfaction and activeness, thus recommend micro-videos with higher estimated watch time [14]. Specifically, recommender systems take rich features like user profiles and video attributes as input, and predict watch time with a parametric model such as deep neural networks [14]. Micro-videos with longer predicted watch time are ranked higher and recommended to the users. However, longer watch time does not necessarily indicate that the user is more interested in the micro-video, since watch time is highly correlated with the duration of the video. Such duration bias makes it challenging to evaluate the performance and learn user preferences for micro-video recommendation.

The duration bias hidden in user-video interaction data means that micro-videos with longer duration tend to have longer watch time, since users usually decide whether to continue watching or switch to the next one until watching a certain fraction of the micro-video. Here the *duration* is defined as *the total length of a micro-video*. Figure 1 (a) shows the duration and average watch time of a real-world micro-video dataset, Wechat Channels (details of the dataset will be introduced later in Section 4.1.1). The watch

[1] https://www.tiktok.com/

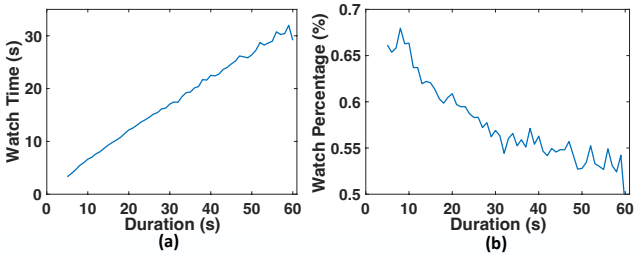

Figure 1: Duration bias of micro-videos with different duration. (a) Mean watch time (b) Mean watch percentage .

time grows as duration increases, which demonstrates the existence of duration bias. To address it, an intuitive solution is to use "watch percentage" instead of watch time. Unfortunately, micro-videos of short duration tends to have larger watch percentage, which means that duration bias still exists but in the opposite direction, illustrated in Figure 1 (b). It is worthwhile to notice that Figure 1 (a) and Figure 1 (b) actually describe the same phenomenon and Figure 1 (b) can be obtained by normalizing Figure 1 (a) with the video duration (x-axis). In addition, our finding of duration bias is in line with related literature [51] on measuring user engagement on online videos. In the following, we will elaborate on how the duration bias leads to two main undesired consequences: *inaccurate* recommendation and *unfair* recommendation.

Inaccurate recommendation caused by duration bias. Unlike traditional scenarios that deal with discrete user feedback, such as rating prediction [30–32], implicit collaborative filtering (CF) [25, 26, 45] and click-through rate (CTR) prediction [40, 47, 60], user engagement towards videos is mainly reflected by the watch time, which is continuous [33, 46]. Specifically, a user tends to continue watching if he/she is interested in the current video, and otherwise, he/she may switch to the next one. In other words, the continuous value, watch time, serves as a substantial indicator of user preference. However, caused by duration bias, longer watch time does not necessarily mean that users are more interested, which we have shown in Figure 1. As a consequence, a recommendation model can be easily misled by the duration bias, and recommend too many micro-videos that do not match user preference but with long duration. It is worthwhile to notice that micro-video platforms like TikTok insert advertisements between different micro-videos, thus simply recommending long micro-videos will NOT bring higher advertising revenue, which is different from platforms like YouTube that insert advertisements inside the videos.

Unfair recommendation caused by duration bias. On the other hand, different users upload micro-videos of different duration, ranging from a few seconds like short funny videos to longer ones of a few minutes like VLogs. As we have mentioned, such duration bias makes longer videos more likely to be recommended than shorter videos, which favors long video publishers and is unfair for short video publishers. To show this point, we compute the average duration of the uploaded micro-videos for each user on the above Wechat Channels dataset, and separate all the micro-video producers from the middle into two groups, which are long micro-video producers and short micro-video producers. Figure 2 (a) illustrates the quite different distributions of the published micro-videos with respect to the duration of the two user groups. Then we implement the famous Factorization Machine (FM) model [44], and show the

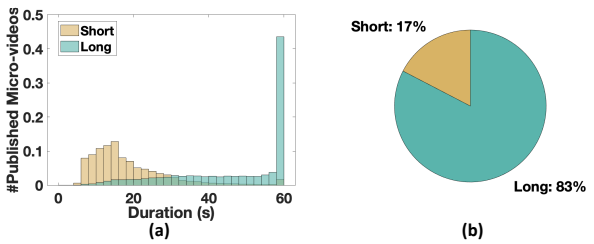

Figure 2: (a) The distribution of uploaded micro-videos for two user groups; (b) The recommendation traffic received by the two user groups.

recommendation chances received by the two groups in Figure 2 (b). We can observe that "long" group receives much more recommendation chances with over 80% than "short" group with less than 20%, although the two groups have the same number of users. Such results show that recommendation based on watch time lead to unfairness for shorter-video publishers.

In this paper, we take the first step to eliminate duration bias for micro-video recommendation. Specifically, in order to reduce the impact of duration bias on both evaluation and learning of micro-video recommendation, we investigate two research questions.

- **RQ1: How to measure users' watch time towards micro-videos in an unbiased way?** Since traditional metrics, such as total watch time and total watch percentage of the top k recommended micro-videos, all suffer from duration bias, which favor either long or short micro-videos, it is crucial to define an unbiased metric that does not favor either side.
- **RQ2: How to learn unbiased user preferences on micro-videos of different duration and provide accurate recommendation?** Existing recommendation approaches are vulnerable to the duration bias since the duration of micro-videos is a strong feature when predicting watch time. Therefore, designing recommender systems that are free from duration bias is useful to capture users' real interests in micro-videos.

Alleviating duration bias for micro-video recommendation is largely unexplored, and we face two main challenges. **First**, micro-videos of different duration can not be compared directly. The final watch time of a micro-video is determined by both user preference and video duration. Therefore, watch time and video duration need to be compared jointly to evaluate the performance with respect to user preference. **Second**, since the structural differences between recommendation models vary widely, the bias alleviation design is supposed to be general and model-agnostic. In other words, it needs to be compatible with any recommendation model that ranks micro-videos according to rich input features.

For the first research question, we propose an unbiased evaluation metric **W**atch **T**ime **G**ain (WTG), which measures a user's relative engagement on a video against the average engagement of all users on videos with the same duration-level. The proposed metric overcomes the influence of video duration, and videos of different duration are forced to be *flattened* equally, and they are comparable with each other, which addresses the first challenge. Meanwhile, to emphasize the order of recommended micro-videos, *i.e.* micro-videos of larger WTG are best to rank higher in the recommendation list, we further propose a **D**iscounted **C**umulative version of WTG (DCWTG) inspired by the widely adopted NDCG

metric in recommendation literature [7, 25, 59]. The proposed WTG and DCWTG provide unbiased evaluation protocols for micro-video recommendation. For the second research question, we further propose a framework named **D**ebiased **V**ideo **R**ecommendation (DVR), which can learn user preference with simple and effective strategies to remove the duration bias and facilitate accurate recommendation. The proposed DVR framework adds an adversarial layer on the predicted value of existing recommendation models, and it does not have any preset requirements for the structure of the backbone models. Therefore, it can be combined with any off-the-shelf recommender systems, which addresses the second challenge.

We conduct experiments on two real-world datasets collected from the largest micro-video platforms in China. Specifically, we perform a large-scale analysis to investigate the impact of duration bias and the shortcomings of existing metrics of micro-video recommendation. In addition, we demonstrate that the proposed metric WTG can measure users' watch time on micro-videos in an unbiased way which does not favor long or short videos. Furthermore, we show that WTG can help construct unbiased recommender systems. We combine DVR with various backbone models, and experimental results show that DVR can improve state-of-the-art recommendation approaches with over 30% relative progress.

The main contributions of the paper are summarized as follows:

- We take the pioneering step to alleviate duration bias for micro-video recommendation. We conduct a large-scale analysis to show how duration bias leads to inaccurate recommendation.
- We propose a new metric, WTG, to achieve unbiased measurement of users' watch time on micro-videos, which eliminates duration bias. We further propose a novel model DVR to learn unbiased user preference on micro-videos of different duration.
- Extensive experiments on two real-world datasets show that our proposed metrics and DVR model successfully achieve unbiased recommendation of micro-videos.

2 DATA ANALYSIS

As introduced previously in Figure 1, micro-videos with longer duration tend to have longer watch time. In this section, we further investigate the impact of such duration bias on recommendation models. We conduct analysis on the same dataset as Figure 1 and 2, and the details of the adopted dataset will be introduced in Section 3. Specifically, we select representative and state-of-the-art recommendation models that aim to predict watch time, following the common paradigm in existing recommender systems. Then we demonstrate that these models are influenced by the duration bias, which makes them recommend too many micro-videos that do not match user preference but with long duration. Meanwhile, we also compare these methods with several intuitive and trivial approaches, such as always recommending long videos, and reveal the shortcomings of existing metrics.

Distribution shift of recommendation results. In order to study the influence of duration bias on micro-video recommendation, we use recommendation models to predict the watch time of micro-videos and analyze the results. We adopt classical algorithms including LibFM [44], Wide&Deep [11], DeepFM [23], NFM [24] and AFM [53], as well as state-of-the-art approaches including AutoInt [47] and AFN [12]. We first calculate the average predicted watch

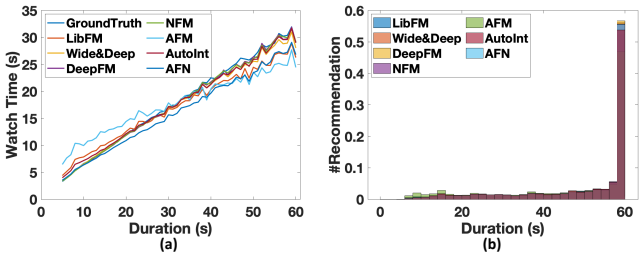

Figure 3: (a) Average predicted watch time of different models. (b) Histogram of recommended micro-videos with different duration of different models.

time of micro-videos with different duration. We discretize duration into equally sized bins with a width of 1 second. Figure 3 (a) shows the predicted and groundtruth watch time of micro-videos in different bins. We can observe that models tend to *amplify* duration bias in the data. Specifically, we can find in 3 (a) that the slope of curves of all models are much higher than the groundtruth curve (blue). In other words, the predicted watch time of long (short) videos is much longer (shorter) than it supposed to be. Such bias amplification of recommendation models is damaging to user experience, since they recommend too many micro-videos with long duration. To illustrate this point, we select the top k recommended micro-videos of all models and plot the histogram of their duration in Figure 3 (b). We can discover that recommended micro-videos concentrate on the long duration side, while all the models almost do not recommend any short micro-videos. We now show that such biased recommendation is not only inaccurate but also unfair.

- *Inaccurate recommendation due to the distribution shift.* Simply recommending micro-videos with a long duration can not meet users' needs which leads to inaccurate recommendation, since there are a large amount of *bad* cases of long micro-videos, *i.e.* users may quickly find that they have no interest in the long micro-video and switch to the next one. Therefore, we calculate the number of bad cases for these recommendation models. Specifically, we define bad cases as the recommended micro-videos with groundtruth watch time lower than 2 seconds. In order to reveal the shortcomings of using watch time for micro-video recommendation, we add two trivial models **LongRec** and **RandomRec**, where LongRec model ranks the micro-videos directly according to the duration thus long micro-videos rank higher, and RandomRec model just randomly shuffles the micro-videos to provide a recommendation list. Table 1 shows the results of all the above models with respect to Mean Absolute Error (**MAE**), Root Mean Squared Error (**RMSE**), total Watch Time of top k videos (**WatchTime@k**) and the number of bad cases (**#BC**). We have two important observations. First, both classical models and state-of-the-art models achieve comparable WatchTime@k with the trivial Long model. In other words, although these models have a strong capacity with thousands of learnable parameters, they fail to learn much more than the duration bias. Second, bad cases generated by these models are only slightly less than the *Random* model, which means that it is not a reasonable choice to recommend blindly according to the predicted watch time. Such many bad cases indicate that the duration bias results in low recommendation accuracy. Therefore,

Table 1: The impact of duration bias (larger WatchTime@k and smaller MAE, RMSE, #BC means better performance). We can observe that personalized models are even as poor as non-personalized RandomRec or LongRec.

Model	MAE	RMSE	WatchTime@k	#BC
RandomRec	12.18	18.21	117.41	3850
LongRec	6.30	12.50	202.92	3679
LibFM	5.48	8.13	204.72	3560
Wide&Deep	5.26	7.85	205.75	3558
DeepFM	5.29	7.85	205.83	3553
NFM	5.23	7.82	206.08	3550
AFM	6.67	9.86	158.48	3515
AutoInt	5.23	7.86	205.77	3568
AFN	5.72	8.44	201.56	3536

it is crucial to define an unbiased metric to measure user engagement towards micro-videos. In addition, an unbiased metric can also facilitate user preference learning to make it free from the influence of video duration.

- *Unfair recommendation due to the distribution shift.* We compare the recommendation traffic received by long and short micro-video publishers. Specifically, we use the above well-trained models, and rank the micro-videos according to the estimated watch time. Then for each user, we recommend k micro-videos with the highest estimated watch time. We vary the value of k, and Figure 4 illustrates the recommendation traffic received by users who mainly produce long or short micro-videos. We can observe that short micro-video publishers hardly receive any recommendation when k is small, and they only receive less than 20% of recommendation traffic even with a large enough value of k. Meanwhile, long micro-video publishers obtain much more recommendation for their videos than short micro-video publishers. Comparison of the recommendation traffic verifies that recommending micro-videos based on predicted watch time leads to serious unfairness for different micro-video publishers.

In short, we have the following observations from data analysis.

- Duration bias is amplified by recommendation models, leading to unbalanced recommendation results: the model recommends much more long micro-videos than short ones.
- Such unbalance leads to inaccurate recommendation and a large number of bad recommendation cases.
- The unbalance leads to unfairness, favoring long micro-video producers, which is unfair for short micro-video producers.

3 METHOD

To alleviate the duration bias that leads to inaccurate and unfair recommendation, we propose a new unbiased metric of watch time, WTG, and an unbiased recommendation model, DVR.

3.1 WTG: An Unbiased Metric of Watch Time

Watch Time Conditioned on Duration. Based on the above analysis, we can conclude that watch time can not be directly used as an indicator for user engagement/preference since watch time is to a great extent dominated by the duration bias. However, if we condition on the value of micro-video duration, watch time can be regarded as a reasonable metric on whether the micro-video matches

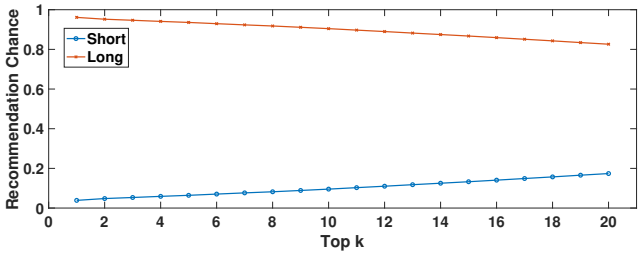

Figure 4: Recommendation chances (frequency of being recommended) of two user (producer) groups *w.r.t* top-k.

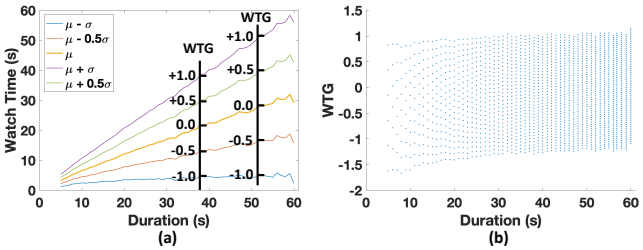

Figure 5: (a) Illustration of how WTG is calculated from Watch Time and Video Duration. (b) Distribution of WTG on Wechat Channels dataset. We divide duration into bins with 1 second per bin.

the user's preference [51]. For example, a user u may watch 50 seconds of a 60s-duration micro-video v_i, and only watch 5 seconds of another micro-video v_j with the same duration of 60 seconds. Then we can confidently infer that the user u prefers micro-video v_i, while he/she might have little interest in micro-video v_j. In other words, the watch time of a micro-video can be used as a metric only when it is compared with other data points of micro-videos with similar duration. Therefore, we define a new metric called Watch Time Gain **(WTG)**, which measures the relative user engagement on a micro-video compared with the average engagement of all users on micro-videos with a similar duration. Specifically, we first divide all the micro-videos into equally wide bins according to their duration, and each micro-video can be mapped to its corresponding duration bin as follows,

$$\mathbf{B} = [b_1, \cdots, b_m], \tag{1}$$

$$B(v) = f_b(d_v), \tag{2}$$

where m is the number of bins, d_v is the duration of micro-video v, and f_b is a function mapping a duration to the corresponding bin.

Since videos in the same bin share similar duration, we compare the watch time using data points within each bin, rather than all bins. Formally, we calculate the mean and standard deviation of watch time in each bin, and then **WTG** is computed as follows,

$$\mathbf{WTG} = \frac{\mathbf{WT} - \mu_{B(v)}}{\sigma_{B(v)}}, \tag{3}$$

where **WT** represents watch time of the data point, $\mu_{B(v)}$ is the mean of watch time in the micro-video v's corresponding bin, and $\sigma_{B(v)}$ is the standard deviation of watch time in that bin. Both $\mu_{B(v)}$ and $\sigma_{B(v)}$ are calculated from the records of all the users on the whole dataset. In other words, we standardize the watch

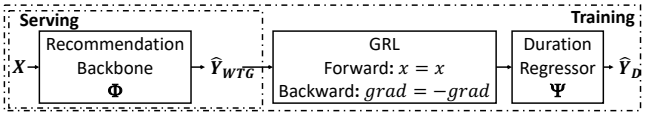

Figure 6: Illustration of the proposed DVR framework.

time within each bin, which makes WTG independent with micro-video duration. Intuitively, WTG eliminates the gap between watch time in different bins by normalization, and provides an unbiased measurement of user engagement in micro-videos with different duration. Figure 5 (a) illustrates the calculation of WTG, as well as how it is related to watch time and micro-video duration. By normalizing inside each bin, micro-videos of different duration can be compared by the WTG. For example, micro-videos of 0 WTG means they are just normal videos regardless of their duration. And a short video of 1 WTG is definitely better than a long video of 0.5 WTG, even though the original watch time of the long video might be longer. Figure 5 (b) shows the distribution of WTG on different micro-video duration, based on the same dataset above. We can observe that WTG successfully alleviates duration bias since it is distributed more uniformly and does not favor long or short micro-videos. Moreover, the proposed WTG metric can be efficiently implemented, and the computational details are introduced in Section A.1.

3.2 DVR: Unbiased Recommendation Model

With the unbiased WTG measurement of user engagement towards micro-videos, we now show how to achieve unbiased recommendation under the guidance of WTG. Since the structures of backbone models can be quite different, the debiasing design need to be general and compatible for different models. Therefore, we propose a simple yet effective framework called Debiased Video Recommendation (**DVR**) which is model-agnostic and it has no preset requirements for backbone models. Figure 6 illustrates the overall design of DVR, where Φ is the backbone which can be any off-the-shelf recommendation models. Specifically, we add an adversarial model Ψ on the predicted value of Φ, to make it independent with micro-video duration, thus reduce the impact of duration bias. We now elaborate on the proposed DVR framework.

3.2.1 Input Features. User profiles and micro-video attributes constitute the input features of recommender systems. In existing approaches, micro-video duration is included as input features and fed into a machine learning model to predict watch time. Particularly, it serves as an important feature, to some extent even the most important one, due to the duration bias. In other words, video duration in the input feature becomes a shortcut for recommendation models to predict watch time directly from it and ignore other features that are related to user preference modeling. From our above analysis, such duration bias is the key reason of inaccuracy and unfairness. Therefore, we propose to remove micro-video duration from input features, which eliminates duration bias fundamentally.

3.2.2 Prediction Target. As shown previously in Section 2, recommending directly according to watch time can not well capture user preference and is unfair for short micro-video publishers. One trivial solution is to transform the predicted watch time to WTG, and then recommend micro-videos according to WTG. However, it is difficult to predict watch time accurately due to the unbalanced distribution of watch time and duration. Specifically, the predicted watch time of long (short) videos tend to be longer (shorter), *i.e.* bias amplification shown in Figure 3 (a). Thus, we utilize the proposed unbiased measurement WTG as the prediction target. During the model training, we optimize recommendation models to predict WTG as accurately as possible. As for the final recommendation, we rank candidate micro-videos according to the predicted WTG, and then top-k micro-videos with higher ranks are recommended to the user. Since the distribution of WTG is more uniform and independent with duration as shown in Figure 5 (b), it is much easier to predict accurately than the biased watch time target.

To evaluate the performance of top-k micro-video recommendation, we further propose **WTG@**k which is the average of the groundtruth WTG of the top recommended micro-videos as follows,

$$\textbf{WTG@}k = \frac{1}{k}\sum_{i=1}^{k} \mathrm{WTG}(l_i), \tag{4}$$

where l_i is the video with the i-th highest predicted WTG. Thus higher WTG@k means better micro-video recommendation performance, since the user is willing to spend more time watching the recommended micro-videos, compared with the watch time of a random list of micro-videos with a similar duration.

The above WTG@k metric is insensitive to the order of the k recommended micro-videos, which means the same k micro-videos of different orders for a given user will share the same WTG@k. Moreover, in the research field of recommender systems, it is widely acknowledged that higher positions are more important compared with lower ones [25]. Inspired by the commonly adopted Normalized Discounted Cumulative Gain (NDCG) metric in recommendation [7, 25, 59] which emphasize the order of recommended items by assigning larger weights to higher positions, we further propose DCWTG@k, which is the discounted cumulative version of WTG. Specifically, DCWTG@k adds a decaying factor which imposes larger weights on the head of the list, and it is calculated as follows,

$$\mathrm{DCWTG@}k = \sum_{i=1}^{k} \frac{\mathrm{WTG}(l_i)}{\log_2(1+i)}. \tag{5}$$

3.2.3 Model Training with Adversarial Learning. We now investigate how to capture unbiased user preference that is free from the influence of micro-video duration. Although we remove duration from input features and use the unbiased WTG as the prediction target, the influence of duration bias can not be fully eliminated and it still hides implicitly in the data, *e.g.* duration can be correlated with other input features like micro-video category. Therefore, in order to make the predicted WTG independent with micro-video duration, we add an extra regression layer, denoted as Ψ, to predict duration from the estimated WTG and train the recommendation model, denoted as Φ, in an adversarial way. Specifically, we encourage the extra regression layer to predict micro-video duration as accurately as possible, and force the recommendation model to best fool the regression layer. In short, it follows a manner of adversarial learning, which can be formally denoted as follows,

$$\hat{Y}_{WTG} = \mathbf{\Phi}(X), \tag{6}$$

$$\hat{Y}_D = \mathbf{\Psi}(\hat{Y}_{WTG}), \tag{7}$$

where X is the input features, $\hat{Y}_{WTG}$ and $\hat{Y}_D$ are the predicted WTG and duration, respectively.

As for duration regression model $\mathbf{\Psi}$, we force it to discover possible correlations between the predicted WTG from $\mathbf{\Phi}$ and micro-video duration as much as possible. As for the recommendation model $\mathbf{\Phi}$, the adversarial learning encourages it to squeeze out all the information about micro-video duration. In other words, by adding an extra regression model $\mathbf{\Psi}$, the recommendation model $\mathbf{\Phi}$ learns to predict WTG without being disturbed by micro-video duration. Inspired by the recent advances[17, 58], we implement the adversarial learning by inserting a Gradient Reversal Layer (GRL) between $\mathbf{\Phi}$ and $\mathbf{\Psi}$, as illustrated in Figure 6. In this way, the recommendation model $\mathbf{\Phi}$ captures unbiased user preference, which is free from the notorious duration bias. Then we have two loss functions for regression as follows,

$$L_{WTG} = \mathbf{MSE}(\hat{Y}_{WTG}, Y_{WTG}), \tag{8}$$

$$L_D = \mathbf{MSE}(\hat{Y}_D, Y_D), \tag{9}$$

where Y_{WTG} and Y_D are the groundtruth value of WTG and duration. Here **MSE** represents the *Mean Squared Loss* function. To balance the two loss functions, we add a hyper-parameter α, which controls the intensity of adversarial learning. The two components, $\mathbf{\Phi}$ and $\mathbf{\Psi}$, are optimized with L_{WTG} and L_D in an end-to-end manner. We show the whole process of DVR in Algorithm 1.

3.2.4 Discussion of Backbone Recommendation Model. It is worthwhile to notice that the proposed DVR approach is highly general and can be integrated with any off-the-shelf recommendation models, since we impose no restrictions on the structure of $\mathbf{\Phi}$. Specifically, the extra duration regression component $\mathbf{\Psi}$ can be appended on any appropriate $\mathbf{\Phi}$ that can perform real-value regression from high-dimensional input features. For example, existing deep learning based recommendation models [12, 23, 24, 47] are perfect candidates for $\mathbf{\Phi}$. We will show in experiments (Section 4) that DVR can achieve consistent improvements in both fairness and accuracy when combined with different backbone recommendation models.

Remark. In real-world micro-video applications, multi-task learning framework [18, 28] is usually adopted, which includes targets other than watch time, such as like, comment, follow, and so on. Although these signals may not be affected by the duration bias, they are hard to collect (very sparse in the real world), while watch time is the most fundamental user feedback in micro-video platforms [14]. Therefore, our solution is essential and practical.

Algorithm 1 Debiased Video Recommendation (DVR)

Input: Training data O with features X, watch time labels Y_{WT}, and duration labels Y_D
Models: WTG regression model $\mathbf{\Phi}$, duration regression model $\mathbf{\Psi}$

```
remove duration from features X
compute WTG labels Y_WTG by Algorithm 2/3
while not converge do
    for batch in O do
        compute Ŷ_WTG, Ŷ_D according to (6)-(7)
        compute L_WTG, L_D according to (8)-(9)
        optimize Ψ with αL_D
        optimize Φ with L_WTG − αL_D
    end for
end while
```

4 EXPERIMENTS

4.1 Experimental Settings

4.1.1 ***Datasets.*** We utilize two public real-world datasets, which are collected from two large micro-video platforms, Wechat Channels[2] and Kuaishou[3]. Each dataset is composed of abundant micro-videos of different duration, and both exhibits strong duration bias, *e.g.* duration bias of Wechat dataset has been shown in Figure 1. The details of the adopted datasets are introduced in Section A.2.

4.1.2 ***Backbone Models.*** To investigate the recommendation performance, we experiment with both classical and state-of-the-art recommendation backbone models, including **LibFM** [44], **Wide&Deep** [11], **DeepFM** [23], **NFM** [24], **AFM** [53], **AutoInt** [47] and **AFN** [12]. Details of all the models are in Section A.3.

4.1.3 ***Metrics.*** To evaluate the performance of learning user preference, we calculate the two proposed metrics, WTG@k and DCWTG@k. Both metrics are calculated for each user, and we report the average value of all users. Higher WTG@k and DCWTG@k mean better recommendation performance. We also evaluate the number of bad cases (#BC@k) for each model, which is defined previously in Table 1 as the number of recommended videos with watch time less than 2 seconds. It is worthwhile to note that lower #BC@k means better recommendation performance. k is set as 10 in our experiments, a widely selected value [25], which measures the quality of the top 10 recommended videos.

Implementation details are introduced in Section A.4.

4.2 Effectiveness of WTG (RQ1)

For each recommendation model, we train two versions of it, using watch time or WTG as the target, respectively. We stop training when the regression accuracy on the validation set converges. During the evaluation, for each user, the two versions predict watch time or WTG of the micro-videos in the test set, then micro-videos are ranked according to the estimated watch time or WTG, respectively. Finally, top k micro-videos with the highest estimated watch time or WTG are recommended to each user. We compare the two versions with respect to both accuracy and fairness.

Accuracy Comparison of Watch Time v.s. WTG. Recommendation according to watch time may lead to low accuracy since there are many bad cases where the user only watches a few seconds of a long micro-video. To illustrate this point, we investigate the accuracy of recommended micro-videos of different duration. Specifically, we separate the recommended micro-videos into bins according to their duration, and then calculate the average groundtruth watch time of the recommended micro-videos. Figure 7 shows the results of recommending according to watch time or WTG. We

[2]https://www.wechat.com/en
[3]https://www.kuaishou.com

Table 2: Recommendation performance comparison of different backbone models with/without DVR on two datasets.

Method		Wechat			Kuaishou		
Backbone	**Debias**	WTG@10	DCWTG@10	#BC@10	WTG@10	DCWTG@10	#BC@10
FM	None	0.0209	0.2985	6381	0.0571	0.4178	5854
	DVR-	0.1249	1.3813	5994	0.1662	1.1318	5728
	DVR	**0.1332**	**1.5100**	**5947**	**0.2094**	**1.6137**	**5240**
WDL	None	0.0265	0.3880	6326	0.0532	0.4031	5851
	DVR-	0.1342	1.4683	5926	0.2002	1.4810	5511
	DVR	**0.1468**	**1.6539**	**5881**	**0.2087**	**1.5833**	**5226**
DeepFM	None	0.0236	0.3648	6345	0.0550	0.4161	5843
	DVR-	0.1372	1.5086	5894	**0.2132**	1.5664	5426
	DVR	**0.1469**	**1.6551**	**5866**	0.2066	**1.5902**	**5261**
NFM	None	0.0234	0.3334	6345	0.0561	0.4478	5826
	DVR-	0.1302	1.4338	5952	**0.2089**	1.5632	5368
	DVR	**0.1444**	**1.6226**	**5899**	0.2081	**1.6050**	**5230**
AFM	None	0.0335	0.4028	6349	0.1052	0.7237	6337
	DVR-	0.1203	1.3318	5986	0.1260	0.8890	5726
	DVR	**0.1391**	**1.5656**	**5930**	**0.2082**	**1.6068**	**5209**
AutoInt	None	0.0272	0.3862	6330	0.0504	0.3823	5868
	DVR-	0.1351	1.4841	5924	**0.2124**	1.5561	5343
	DVR	**0.1458**	**1.6420**	**5874**	0.2086	**1.5905**	**5237**
AFN	None	0.0157	0.2599	6358	0.0536	0.4037	5832
	DVR-	0.1254	1.3714	6064	0.1691	1.2442	5552
	DVR	**0.1408**	**1.5858**	**5917**	**0.2015**	**1.5551**	**5229**

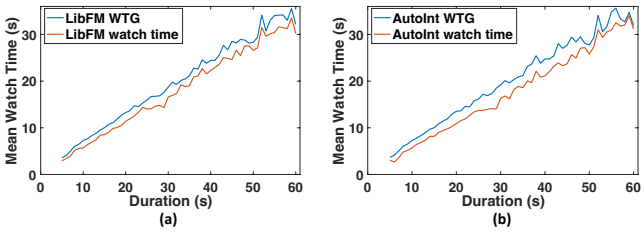

Figure 7: Accuracy comparison between Watch Time and WTG. We plot the mean watch time of recommended videos. Two selected models: (a) LibFM (b) AutoInt.

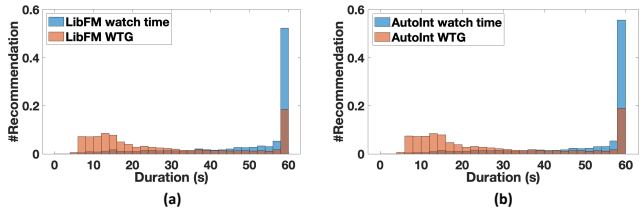

Figure 8: Fairness comparison between Watch Time and WTG. We plot the histogram on recommended videos of different duration. Two selected models: (a) LibFM (b) AutoInt.

can observe that models trained with watch time as target result in inferior recommendation quality, since the average ground-truth watch time of recommended micro-videos is consistently lower than trained with WTG. With the unbiased WTG as the prediction target, models can generate much more high-quality recommendations of both long and short micro-videos.

Fairness Comparison of Watch Time v.s. WTG. Figure 8 illustrates the histogram on the duration of recommended micro-videos for LibFM and AutoInt. Results of the other five backbones are similar and we omitted due to the space limitation. We can observe that the recommended micro-videos from models trained with watch time as target mainly concentrate on the long duration side, which is because watch time is to a great extent dominated by the duration bias. For example, videos of 60s (the maximum length in Wechat dataset) take over 52.3% and 55.6% of recommendation traffic for LibFM and AutoInt, respectively, while short videos of less than 20s almost receive no recommendation with recommendation traffic lower than 1.46%. Using WTG as the target can largely solve this problem, and we can discover that the duration distribution of recommended micro-videos from models trained with WTG as the target is much more balanced compared with using watch time as the target. Specifically, videos of 60s only receive about 18.47% and 18.95% of recommendation for LibFM and AutoInt, which is far less than using watch time as the target, and short videos of less than 20s also receive fair enough recommendation chance with about 8.59% of total recommendation. In other words, the proposed WTG serves as an unbiased target to train recommendation models, and achieves fair recommendation of micro-videos with different duration, which does not favor long or short micro-video publishers.

In summary, the proposed unbiased target WTG successfully improves both the accuracy and fairness of recommendation models against the biased watch time target.

4.3 Effectiveness of DVR (RQ2)

We combine DVR with all the recommendation backbones, and Table 2 shows the results. We also include a simplified version of the proposed model called **DVR-**, which means that the backbone

model is trained with watch time as target while we transform the predicted watch time to the proposed metric WTG for ranking. We have the following observations:

- **Worse performance of existing recommendation models.** Without special designs to eliminate duration bias, existing recommendation approaches can not well capture user preference, and they are easily misled to blindly recommend micro-videos with long duration. However, these long micro-videos fail to meet users' interest, and users only watch a few seconds of them, leading to low WTG and DCWTG. Note that in equation (3) WTG is a normalized metric by subtracting mean watch time then divided by the standard deviation, thus WTG close to zero means that the recommended micro-videos are almost as the same quality as random recommendation. We can observe that the WTG of top-10 micro-videos is very low for all recommendation models without debiasing, which verifies that duration bias leads to bad recommendation accuracy. Meanwhile, #BC@10 of backbone models without any debiasing design is much higher than DVR- and DVR, which means that recommending according to watch time provides a large amount of unsatisfactory micro-videos, which may directly lead to user churn.
- **Steady improvement of our DVR model.** The proposed DVR can improve recommendation accuracy significantly. Specifically, the progress of WTG@10 is over 300% in most cases of seven backbones on two datasets. For the state-of-the-art method AFN, DVR can improve WTG by over 500%. In addition, the number of bad cases for DVR is significantly less than simply using the backbone models. For example, #BC@10 of DVR is about 21.65% less than AFM on Kuaishou dataset. Meanwhile, consistent improvements across different backbones demonstrate that DVR is a highly general framework that can be integrated smoothly with existing recommendation approaches. Another interesting finding is that although DVR- is worse than DVR in most cases, it outperforms backbone models with significant improvements. In fact, DVR- utilizes well-trained biased models, and corrects the duration bias directly from the predicted watch time, by transforming it to the unbiased WTG value. The huge improvements of DVR- over backbone models indicate that it is easy to apply our proposed WTG to existing recommendation systems.

Due to space limit, more experimental results including ablation study and hyper-parameter study of DVR can be found in Section A.5.

5 RELATED WORK

Video Recommendation. Users are spending more and more time in video apps, especially micro-video apps such as TikTok and Kuaishou. As the number of uploaded videos is quite large, it is critical to utilize a recommender system to provide personalized videos to users [6, 9, 10, 14, 15, 20, 27, 39, 41, 49, 50, 54]. For example, the YouTube recommendation has evolved from rule-based systems [15], to Deep Neural Networks (DNN) based models [14], then Recurrent Neural Networks (RNN) based models [6], and now Reinforcement Learning (RL) based models [9]. Li *et al.* [39] proposed to capture user interest by leveraging multiple user behaviors towards micro-videos such as click, like and follow, with a graph-based Long Short-Term Memory (LSTM) model. In addition, Wei *et al.* [50] proposed a Graph Convolutional Networks (GCN) based model which leverages multi-modal information to enhance the performance of short video recommendation. However, these approaches either focus on traditional discrete user feedback like clicks or predict the continuous watch time feedback, which exhibits strong bias. To the best of our knowledge, we are the first to reduce duration bias for micro-video recommendation, which is crucial for learning users' real interests that are independent of video duration.

Fairness-aware Recommendation. As recommender systems grow increasing impact on users, fairness becomes a critical issue [3–5, 16, 21, 37, 38, 48, 52, 55, 61], especially in user-generated content (UGC) platforms where multi-stakeholders are involved, such as micro-video applications. Fairness-aware recommendation is generally studied from two perspectives [38, 42], including user fairness which focuses on algorithmic bias towards specific individuals or user groups [34, 36], and item fairness which means fair recommendation traffic received by different items [2, 38, 42]. Unlike existing fairness-aware recommendation literature, in this paper, we address a specific fairness issue in micro-video platforms, where micro-videos of different duration tend to receive unfair recommendation traffic.

Duration Bias. Bias in recommender systems [8] has been studied from several directions, such as popularity bias [22, 57, 59] and position bias [13, 43]. However, duration bias in video recommendation has been unexplored until a recent study [51], in which Wu *et al.* investigated the bias of watch time and watch percentage from an aggregated level, *i.e.* the average of the watch time of all users towards each video. In other words, it merges all samples of the same video into one single data point, and compares with other videos to measure the video quality. Unlike [51], our study focuses on the personalized duration bias, where different users have distinct WTG values towards the same micro-video. Our setting is closer to the real-world recommendation scenarios, and the proposed WTG metric can be directly integrated into online recommender systems.

6 CONCLUSION AND FUTURE WORK

In this paper, we investigate a largely unexplored duration-bias problem in micro-video recommendation. We conduct large-scale data analysis to show that the duration bias leads to inaccurate and unfair recommendation. A new measurement of watch time on micro-videos, WTG, is proposed which eliminates duration bias and can evaluate recommendation performance without favoring either long or short videos. A general model DVR is further designed to help recommendation models learn unbiased user preferences. Experiments demonstrate that the proposed metric and model successfully eliminate duration bias, which can achieve accurate and fair recommendation. As for the future work, we plan to apply WTG in online systems to evaluate the performance of micro-video recommendation. We also plan to evaluate DVR with online A/B tests.

ACKNOWLEDGMENTS

This work is supported in part by The National Key Research and Development Program of China under grant 2020AAA0106000. This work is also supported in part by the National Natural Science Foundation of China under U1936217, 61971267, 61972223, U20B2060.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI)*. 265–283.

[2] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys)*. 42–46.

[3] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. 2019. One-network adversarial fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2412–2420.

[4] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2212–2220.

[5] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).

[6] Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H Chi. 2018. Latent cross: Making use of context in recurrent recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM)*. 46–54.

[7] Jianxin Chang, Chen Gao, Yu Zheng, Yiqun Hui, Yanan Niu, Yang Song, Depeng Jin, and Yong Li. 2021. Sequential Recommendation with Graph Neural Networks. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 378–387.

[8] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240* (2020).

[9] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k off-policy correction for a REINFORCE recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM)*. 456–464.

[10] Xusong Chen, Dong Liu, Zheng-Jun Zha, Wengang Zhou, Zhiwei Xiong, and Yan Li. 2018. Temporal hierarchical attention at category-and item-level for micro-video click-through prediction. In *Proceedings of the 26th ACM international conference on Multimedia*. 1146–1153.

[11] 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.

[12] Weiyu Cheng, Yanyan Shen, and Linpeng Huang. 2020. Adaptive factorization network: Learning adaptive-order feature interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 3609–3616.

[13] Andrew Collins, Dominika Tkaczyk, Akiko Aizawa, and Joeran Beel. 2018. A study of position bias in digital library recommender systems. *arXiv preprint arXiv:1802.06565* (2018).

[14] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the tenth ACM conference on recommender systems (RecSys)*. 191–198.

[15] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems (RecSys)*. 293–296.

[16] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, et al. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 69–78.

[17] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning (ICML)*. PMLR, 1180–1189.

[18] Chen Gao, Xiangnan He, Dahua Gan, Xiangning Chen, Fuli Feng, Yong Li, Tat-Seng Chua, and Depeng Jin. 2019. Neural multi-task recommendation from multi-behavior data. In *2019 IEEE 35th international conference on data engineering (ICDE)*. IEEE, 1554–1557.

[19] Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, et al. 2021. Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions. *arXiv preprint arXiv:2109.12843* (2021).

[20] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. 2017. A unified personalized video recommendation via dynamic recurrent neural networks. In *Proceedings of the 25th ACM international conference on Multimedia*. 127–135.

[21] Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. 2022. Toward Pareto Efficient Fairness-Utility Trade-off inRecommendation through Reinforcement Learning. *arXiv preprint arXiv:2201.00140* (2022).

[22] Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. 2018. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM)*. 198–206.

[23] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 1725–1731.

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

[25] 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 (TheWebConf)*. 173–182.

[26] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Eighth IEEE International Conference on Data Mining (ICDM)*. Ieee, 263–272.

[27] Hao Jiang, Wenjie Wang, Yinwei Wei, Zan Gao, Yinglong Wang, and Liqiang Nie. 2020. What Aspect Do You Like: Multi-scale Time-aware User Interest Modeling for Micro-video Recommendation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3487–3495.

[28] Bowen Jin, Chen Gao, Xiangnan He, Depeng Jin, and Yong Li. 2020. Multi-behavior recommendation with graph convolutional networks. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*. 659–668.

[29] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[30] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 426–434.

[31] Yehuda Koren. 2009. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (KDD)*. 447–456.

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

[33] Coco Krumme, Manuel Cebrian, Galen Pickard, and Sandy Pentland. 2012. Quantifying social influence in an online cultural market. *PloS one* 7, 5 (2012), e33785.

[34] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2018. User fairness in recommender systems. In *Companion Proceedings of the The Web Conference 2018*. 101–102.

[35] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. *Mining of massive data sets*. Cambridge University Press.

[36] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021 (TheWebConf)*. 624–632.

[37] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards Personalized Fairness based on Causal Notion. *arXiv preprint arXiv:2105.09829* (2021).

[38] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. 2021. Tutorial on Fairness of Machine Learning in Recommender Systems. SIGIR Tutorial.

[39] Yongqi Li, Meng Liu, Jianhua Yin, Chaoran Cui, Xin-Shun Xu, and Liqiang Nie. 2019. Routing micro-videos via a temporal graph-guided recommendation system. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. 1464–1472.

[40] Bin Liu, Chenxu Zhu, Guilin Li, Weinan Zhang, Jincai Lai, Ruiming Tang, Xiuqiang He, Zhenguo Li, and Yong Yu. 2020. Autofis: Automatic feature interaction selection in factorization models for click-through rate prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD)*. 2636–2645.

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

[42] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*. 2243–2251.

[43] Maeve O'Brien and Mark T Keane. 2006. Modeling result-list searching in the World Wide Web: The role of relevance topologies and trust bias. In *Proceedings of the 28th annual conference of the cognitive science society*, Vol. 28. Citeseer, 1881–1886.

[44] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining (ICDM)*. 995–1000.

[45] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 452–461.

[46] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market.

Science 311, 5762 (2006), 854–856.

[47] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. 2019. Autoint: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*. 1161–1170.

[48] Christina Wadsworth, Francesca Vera, and Chris Piech. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199* (2018).

[49] Yinwei Wei, Zhiyong Cheng, Xuzheng Yu, Zhou Zhao, Lei Zhu, and Liqiang Nie. 2019. Personalized hashtag recommendation for micro-videos. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1446–1454.

[50] 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 (MM)*. 1437–1445.

[51] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. 2018. Beyond views: Measuring and predicting engagement in online videos. In *Twelfth international AAAI conference on web and social media (ICWSM)*.

[52] Yao Wu, Jian Cao, Guandong Xu, and Yudong Tan. 2021. TFROM: A Two-sided Fairness-Aware Recommendation Model for Both Customers and Providers. *arXiv preprint arXiv:2104.09024* (2021).

[53] Jun Xiao, Hao Ye, Xiangnan He, Hanwang Zhang, Fei Wu, and Tat-Seng Chua. 2017. Attentional factorization machines: learning the weight of feature interactions via attention networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 3119–3125.

[54] Feng Yu, Zhaocheng Liu, Qiang Liu, Haoli Zhang, Shu Wu, and Liang Wang. 2020. Deep interaction machine: A simple but effective model for high-order feature interactions. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2285–2288.

[55] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.

[56] Jun Zhang, Chen Gao, Depeng Jin, and Yong Li. 2021. Group-buying recommendation for social e-commerce. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 1536–1547.

[57] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 11–20. https://doi.org/10.1145/3404835.3462875

[58] Yu Zheng, Chen Gao, Liang Chen, Depeng Jin, and Yong Li. 2021. DGCN: Diversified Recommendation with Graph Convolutional Networks. In *Proceedings of the Web Conference 2021 (TheWebConf)*. 401–412.

[59] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling User Interest and Conformity for Recommendation with Causal Embedding. In *Proceedings of the Web Conference 2021 (TheWebConf)*. 2980–2991.

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

[61] Ziwei Zhu, Jingu Kim, Trung Nguyen, Aish Fenton, and James Caverlee. 2021. Fairness among New Items in Cold Start Recommender Systems. (2021).

A APPENDIX

A.1 Offline and Online Computation of WTG.

With respect to offline evaluation, WTG can be efficiently computed from the entire data of recommender systems. We provide the pseudocode for calculating WTG in Algorithm 2.

However, in terms of online serving, what recommender systems handle are streams of unstopped logs. Thus WTG is dynamically changing since the arriving new data from the stream influences the mean and standard deviation of watch time in each bin. Fortunately, the mean and standard deviation can be updated in a recursive manner with no need to save all the data [35]. Specifically, we only need three extra variables (μ, σ, and n) to keep track of the mean and standard deviation of watch time, as well as the number of data points in each bin. These three variables are updated dynamically according to the data stream, and Algorithm 3 briefly illustrates such process. It is worthwhile to notice that our provided Algorithm 3 is just a sketch for the online serving of WTG, and there can be more efficient implementations. Nevertheless, our purpose is to show that the proposed metric can be seamlessly integrated into the online recommendation systems in a real-time streaming manner.

Algorithm 2 Offline Computation of WTG

Input: Dataframe D of format $(user, video, watchtime, duration)$

$DG \leftarrow \text{GroupBy}(D.duration)$
$D_{mean} \leftarrow \text{Mean}(DG.watchtime)$
$D_{std} \leftarrow \text{Std}(DG.watchtime)$
$D.mean, D.std \leftarrow \text{Join}(D, D_{mean}, D_{std})$
$D.wtg \leftarrow (D.watchtime - D.mean)/D.std$

Algorithm 3 Online Computation of WTG

Input: Data stream of application logs S
Tracking Variables: Mean of watch time $[\mu_1, \cdots, \mu_m]$, std of watch time $[\sigma_1, \cdots, \sigma_m]$, and the number of data points $[n_1, \cdots, n_m]$ in m different bins

while S is not empty **do**
  $(\text{WT}, d_v) \leftarrow S.pop()$ // Get watch time and duration.
  $b \leftarrow f_b(d_v)$ // Get the corresponding bin.
  $n_b \leftarrow n_b + 1$ // Update the number of data points.
  $\sigma_b^2 \leftarrow \frac{n_b - 1}{n_b^2}(\text{WT} - \mu_b)^2 + \frac{n_b - 1}{n_b}\sigma_b^2$ // Update std.
  $\mu_b \leftarrow \mu_b + \frac{\text{WT} - \mu_b}{n_b}$ // Update mean.
end while

A.2 Details of the Adopted Datasets

We utilize two public real-world datasets, both of which are collected from large short-video platforms. We summarize the statistics of the adopted datasets in Table 3, where we also list the total duration (in seconds) of all records. The details of the two adopted datasets are as follows,

- **Wechat**: This dataset is released by WeChat Big Data Challenge 2021[4] which contains the logs on Wechat Channels within two weeks. We split the data into the first ten days, the middle two days, and the last two days as training, validation, and test set. The adopted input features include *UserID*, *VideoID*, *DeviceID*, *AuthorID*, *BGMSongID*, *BGMSingerID*, *UserActiveness*, and *VideoPopularity*.
- **Kuaishou**: This dataset [39] is released by the Kuaishou Competition in China MM 2018 Conference[5], and we also split the datasets into training, validation, and test sets according to timestamps with the splitting ratio as 8:1:1. The adopted input features include *UserID*, *VideoID*, *UserActiveness*, and *VideoPopularity*.

Table 3: Statistics of two adopted real-world datasets.

Dataset	#Users	#Videos	#Records	Total Duration (s)
Wechat	10,000	639,557	2,672,809	46,785,442
Kuaishou	20,000	96,418	7,310,108	227,955,046

Since we divide all the micro-videos into separate bins according to their duration and compute the mean and standard deviation of watch time, each bin is supposed to have enough data points to guarantee that the computed μ and σ are statistically significant. Therefore, we filter out those bins of too long or too short duration, which only contain a few data points. Specifically, for the Wechat dataset, we reserve the micro-videos with a duration between 5 seconds and 60 seconds, and for the Kuaishou dataset, we keep the micro-videos longer than 5 seconds and shorter than 120 seconds. Micro-videos with duration outside the above range are of low prevalence and they only take less than 0.1% of all the records. After filtering out too long or too short micro-videos, for both datasets, we construct equally wide bins with 1 second per bin. It is worthwhile to note that each bin contains over 10,000 data points which guarantees the statistical significance of the computed mean and standard deviation values.

A.3 Details of Backbone Models

We include the following recommendation backbone models,

- **LibFM** [44]. This is a classical recommendation algorithm which captures feature interaction by taking inner product of each pair of features.
- **Wide&Deep** [11]. It combines linear regression and deep neural networks to learn direct feature matching and high-order feature interaction separately.
- **DeepFM** [23]. This method ensembles multi-layer perceptions (MLP) and LibFM.
- **NFM** [24]. This method extends LibFM with a Bi-Interaction layer.
- **AFM** [53]. It utilizes attention to aggregate different cross features in LibFM.
- **AutoInt** [47]. It utilizes multi-head self-attention to automatically construct complex feature interactions.
- **AFN** [12]. This is the state-of-the-art method which learns arbitrary order of feature interaction with a logarithmic transformation layer.

[4] https://algo.weixin.qq.com/

[5] https://github.com/liyongqi67/ALPINE

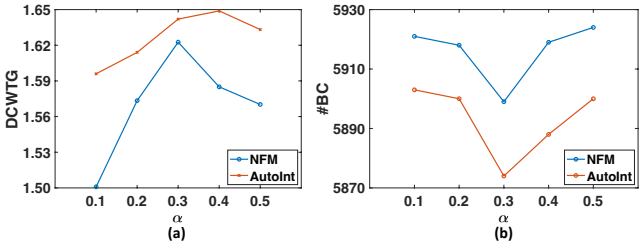

Figure 9: Performance of NFM and AutoInt on Wechat dataset under different values of α with respect to (a) DCWTG (b) Number of bad cases.

Table 4: Ablation study of DVR.

Dataset	Model	None	+DD	+WTG	+ADV
Wechat	NFM	0.3334	+8.59%	+354.28%	+386.68%
Kuaishou	AFN	0.4037	+49.08%	+267.39%	+285.21%

A.4 Implementation Details

We implement all the backbone models, and the proposed DVR with TensorFlow [1]. We use Adam [29] as the optimizer and set the initial learning rate as 0.001. The batch size is set as 512. For a fair comparison, we use three hidden layers and 64 hidden units per layer for all models using DNN. We train the models until convergence and use early stopping to avoid overfitting. For DVR, Ψ is implemented as a 1×1 dense layer, and the optimal α is 0.1. Other hyper-parameters of all these models are tuned carefully on the validation set using grid search, following settings or suggestions of original papers. We have released the code and data at https://github.com/tsinghua-fib-lab/WTG-DVR.

A.5 More Experimental Results

A.5.1 Ablation studies of our DVR model. There are three key strategies in DVR, which are DD (delete duration from input features), WTG (use WTG as the target instead of watch time), and ADV (Adversarial training). We investigate the contribution of each component by adding the three strategies one by one. Table 4 shows the DCWTG@10 of two typical cases. The results of other cases are similar and omitted due to space limitation. We can observe that simply deleting micro-video duration from input features can bring about 10% and 50% improvements on two datasets, respectively. Meanwhile, the largest improvements come from introducing the proposed unbiased WTG as the prediction target. Moreover, adversarial learning can further improve the recommendation performance by about 10%. In fact, the three key strategies eliminate duration bias from three different perspectives, which are input, output, and model itself. Combining the three simple yet effective strategies leads to fair and accurate recommendation of micro-videos.

A.5.2 Hyper-parameter study of DVR.. In the proposed DVR model, we introduce a hyper-parameter α, the loss weight, which controls the intensity of adversarial learning. Figure 9 illustrates the recommendation performance of NFM and AutoInt under different values of α. We can observe that setting α as 0.3-0.4 achieves the best performance with respect to both DCWTG and #BC. On the one hand, low α such as 0.1 imposes too weak adversarial supervision on the recommendation model. In other words, the duration regressor Ψ receives insufficient optimization, which can not provide much help for the recommendation backbone model Φ. As a consequence, the prediction of Φ is still correlated with video duration, which leads to less gain and more bad cases. On the other hand, if α is too high, the auxiliary adversarial task may interfere with the main task. Specifically, the adversarial signals from the duration regressor Ψ becomes dominant of the optimization, which makes the recommendation backbone Φ prone to underfitting.