Invariant Feature Learning for Counterfactual Watch-time Prediction in Video Recommendation

Chenghou Jin[*1], Yixin Ren[*1], Hongxu Ma[*1], Yewei Xia[1], Yi Guan[1], Hao Zhang[2], Jiandong Ding[1], Jihong Guan[†3], Shuigeng Zhou[†1]

[1]College of Computer Science snd Artificial Intelligence, Fudan University, Shanghai, China
[2]SIAT, Chinese Academy of Sciences, Shenzhen, China
[3]Department of Computer Science and Technology, Tongji University, Shanghai, China
{jinch24, yxren21, hxma24, ywxia23, guany23}@m.fudan.edu.cn, {jdding, sgzhou}@fudan.edu.cn, h.zhang10@siat.ac.cn, jhguan@tongji.edu.cn

Abstract

Video recommendation systems heavily rely on user watch time feedback, making accurate watch time prediction a crucial task. However, this task inherently suffers from bias, as recommendation models tend to favor long-duration videos to maximize watch time. This issue, known as duration bias in the watch-time prediction context, can be explained from a causal perspective, where video duration acts as a confounder. Recent works address this bias using backdoor adjustment, isolating the direct effect of content on watch time from observational data. These methods typically discretize video duration into groups, estimate group-wise effects, and then aggregate them via a unified prediction model. However, this aggregation strategy is prone to model misspecification due to feature distribution shift across groups. In this paper, we reinterpret the problem through the lens of invariant learning and propose a novel framework: Duration-Invariant Feature Learning (DIFL). DIFL employs a kernel-based regularization that enforces representation invariance across duration groups, reducing sensitivity to group design and improving generalization. This enables more accurate modeling of the direct causal effect and making counterfactual inference. Extensive experiments on both public and real large-scale production datasets demonstrate the effectiveness of the proposed approach, which achieves SOTA performance.

1 Introduction

With the rapid advancement of multimedia technology, short videos have become a primary source of information and entertainment for many users. To align with users' content preferences and enhance their engagement, it is essential to develop highly accurate and personalized video recommendation systems that effectively deliver relevant and engaging content. In either traditional Video on Demand (VOD) systems (such as YouTube) or streaming media recommendation platforms (such as TikTok and Kuaishou) (Li et al. 2022b; Davidson et al. 2010), video watch time has become a crucial metric for measuring user engagement and experience (Yang et al. 2024; Lin et al. 2025; Malitesta et al. 2025; Ma et al. 2025). Accurately predicting user watch time is vital for platform success, as longer watch time reflects deeper engagement, and enhances user retention and conversion, directly contributing to the growth in Daily Active Users (DAU) and overall revenue (Li et al. 2024; Wu et al. 2018).

Despite its importance, video recommendation is known to suffer from various biases (Zheng et al. 2022; Wu et al. 2022; Wei et al. 2021; Zhao et al. 2023; Liu et al. 2023). Among them, duration bias is a specific challenge in the video recommendation task, as watch time depends not only on how well a video matches a user's interest, but also on the video's inherent duration. Since recommendation systems are designed to maximize user watch time, they naturally favor longer videos, which tend to yield higher absolute watch time — even if they are less relevant to user interest. This bias is further exacerbated by exposure imbalance: longer videos are recommended more often, resulting in larger sample sizes that dominate training and amplify the bias, ultimately skewing the system toward long-duration content while overlooking true user preferences.

From the perspective of causality, this bias can be attributed to that video duration acts as a confounder. To address this, recent studies adopt backdoor adjustment to isolate the direct causal effect of content on watch time (Zhan et al. 2022; Lin et al. 2023; Dong et al. 2024). A common approach is to discretize duration into several groups and train group-specific models to mitigate duration bias. However, maintaining separate models is often impractical due to scalability issues and poor sample efficiency. As a compromise, some works (Zhan et al. 2022; Lin et al. 2023) employ a single unified model that aggregates data across groups. But, this strategy introduces model misspecification due to distribution shift in input features across groups.

To overcome these limitations above, we first revisit the problem from the lens of invariant learning, then propose a novel framework that learns duration-invariant representations by enforcing independence between learned features and video duration. We call this new method **DIFL** — the abbreviation of **D**uration-**I**nvariant **F**eature **L**earning. Specifically, DIFL leverages the Hilbert-Schmidt Independence Criterion (HSIC) — a kernel-based method captures both linear and nonlinear dependencies (Zhang et al. 2018; Gretton et al. 2005; Ren et al. 2023) — to enforce this con-

[*]Co-first authors.
[†]Corresponding authors.

straint. HSIC operates by mapping variables into a high-dimensional kernel space (e.g. via RBF kernel) and measuring their covariance (Ren et al. 2025). To improve efficiency, DIFL adopts random Fourier features (RFF) (Rahimi and Recht 2007) for kernel approximation. By learning features that generalize across different duration groups, DIFL enables a more accurate counterfactual watch-time prediction. Consequently, compared with prior methods, DIFL is more robust to group design, leading to higher performance under various settings.

In summary, our contributions are as follows:

- We revisit existing approaches to handling duration bias, identify their key limitations — especially the aggregation-induced model misspecification — and propose a principled alternative.
- We propose a novel watch-time prediction framework DIFL based on invariant learning, which is the first effort to apply invariant learning to watch-time prediction.
- We develop a method for learning duration-invariant representations by enforcing independence constraints via a kernel-based HSIC regularizer.
- We conduct extensive experiments on two public video recommendation benchmarks and a real large-scale production dataset, which show that our DIFL framework consistently outperforms the existing methods.

2 Related Work

2.1 Watch-time Prediction

Watch time prediction is a crucial task in video recommendation, as it directly impacts user engagement and overall system performance. Given a user's profile, historical interactions, and other relevant information, it aims to predict the watch time for each candidate video. A straightforward approach is Value Regression (VR), which directly predicts the absolute watch time and evaluates performance using Mean Squared Error (MSE). WLR (Covington et al. 2016), originally proposed for YouTube video recommendation, reformulates the task as a weighted logistic regression problem, where each training sample is weighted by its actual watch time. However, this weighting scheme may inadvertently introduce significant duration bias (Lin et al. 2023).

To explicitly address the duration bias, D2Q (Zhan et al. 2022) applies backdoor adjustment by grouping data based on video duration. This is the first work in video recommendation to both identify and mitigate duration bias. TPM (Lin et al. 2023) introduces a tree-based method that decomposes watch time prediction into a sequence of interdependent classification problems. More recently, CWM (Zhao et al. 2024) proposes a counterfactual watch model, which estimates watch time assuming video duration are sufficiently long, reducing bias from truncated sessions. CREAD (Sun et al. 2024) formulates this task as a series of classification problems based on its novel discretization strategy, designed to better balance learning and restoration errors.

Despite existing efforts to address this bias, in this paper we revisit existing debiasing frameworks, identify potential pitfalls inherent with them, and design new schemes to improve the overall performance.

2.2 Invariant Learning

Invariant learning has recently attracted significant attention and has been successfully applied in various domains, including zero-shot visual recognition (Yue et al. 2021), scene graph generation (Min et al. 2023), and out-of-distribution generalization in multimodal large language models (Zhang et al. 2024). It operates under the assumption that data are collected from multiple environments with different distributions, and seeks to learn representations that maintain predictive power across these environments.

A foundational method in this area is Invariant Risk Minimization (IRM) (Arjovsky et al. 2019), which extends the Invariant Causal Prediction (ICP) framework (Peters et al. 2016) by incorporating representation learning. In recommendation systems, invariant learning is particularly useful for separating stable user preferences from context-dependent variations (Wang et al. 2022; Zhang et al. 2023).

Although invariant learning has been explored in various domains, to the best of our knowledge, our work is the first to apply it to addressing video duration bias in watch-time prediction. We highlight the limitations of existing debiasing frameworks and connect them to challenges in invariant learning. Based on this connection, we propose a principled approach for learning duration-invariant representations.

3 Preliminaries

In this paper, we address the problem of watch-time prediction. Given a dataset $\mathcal{D} := \{(u_i, v_i, d_i, w_i)\}_{i=1}^n$, where u_i and v_i are user and video features, d_i is the duration of the video, and $w_i \in \mathbb{R}^+$ is the corresponding ground-truth watch time. The goal is to predict a user's watch time w_i for a given instance based on the combined features (u_i, v_i, d_i). This task is typically formulated as a regression problem, and a straightforward solution is the Value Regression (VR) approach. Formally, let $f(\cdot;\theta)$ be a predictive model parameterized by θ. VR aims to minimize the following loss:

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(f(u_i, v_i, d_i;\theta), w_i\right), \tag{1}$$

where $\mathcal{L}(\cdot,\cdot)$ is a suitable loss function. However, this direct approach often leads to biased predictions due to the presence of duration bias in the data-generating process. Since recommendation models are optimized to maximize total watch time, they tend to preferentially recommend long-duration videos. As a result, the observed dataset $\mathcal{D}$ is not randomly sampled, instead, longer videos have biased exposures (Li et al. 2022a). This induces a spurious correlation between duration and watch time, making Eq. (1) a biased estimator. This issue can be intuitively explained through the lens of causal inference, as discussed in the following.

3.1 Causal Model

We formulate the watch-time prediction problem using a causal graph, which captures the relationships among user features (U), video features (V), video duration (D), and watch time (W). The causal graph is illustrated in Fig. 1. As shown in Fig. 1(a), duration (D) acts as a confounder, influencing watch time (W) through two causal paths: $D \to W$

and $D \to V \to W$. The first path reflects a direct causal effect between duration and watch time, which should be captured by the model. The second path indicates that duration affects video exposure (e.g. longer videos being more likely to be recommended), introducing spurious correlations between duration and watch time. A model trained directly on observational data will inadvertently capture both effects, leading to biased predictions. To address this, we employ do-calculus to isolate the direct causal effect by blocking the confounding influence of duration on video exposure. Specifically, we try to estimate $\mathbb{E}[W \mid do(U, V))]$, with which to remove the spurious edge $D \to V$, as shown in the deconfounded causal graph in Fig. 1(b). This ensures that the learned model captures only the direct relationship between the video and user preference, without interference from recommendation-induced bias.

4 Methodology

In this section, we first revisit the existing frameworks, analyze the potential drawback with them, and then we introduce our solution to overcome the drawback.

4.1 Revisiting the Existing Frameworks

In last section, we establish that the goal of the watch-time prediction task is to model the direct causal effect $\mathbb{E}[W \mid do(U, V)]$. However, due to the presence of the do-operator, estimating this quantity requires access to interventional data, which is generally unavailable in practice. To address this, many approaches (e.g. (Zhan et al. 2022; Lin et al. 2023; Wang et al. 2021)) leverage the backdoor adjustment technique from causal inference, which allows us to express the interventional distribution in terms of the observable distribution under certain assumptions. The following derivation illustrates this process.

Backdoor Adjustment. Let the causal graph in Fig. 1(b) be denoted by CG where the path $D \to V$ is removed, and let $\mathbb{E}_{CG}$ and $\mathbb{P}_{CG}$ represent the expectations and probabilities under the distribution induced by G. Using the backdoor adjustment, the direct causal effect can be computed as

$$\begin{aligned}
&\mathbb{E}[W \mid do(U, V))] = \mathbb{E}_{CG}[W \mid (U, V)] \\
&\overset{(i)}{=} \sum_d \mathbb{P}_{CG}(D = d \mid (U, V)) \, \mathbb{E}_{CG}[W \mid U, V, D = d] \\
&\overset{(ii)}{=} \sum_d \mathbb{P}_{CG}(D = d) \, \mathbb{E}_{CG}[W \mid U, V, D = d] \\
&\overset{(iii)}{=} \sum_d \mathbb{P}(D = d) \, \mathbb{E}[W \mid U, V, D = d],
\end{aligned} \tag{2}$$

where (i) is derived based on the law of total expectation (Weiss, Holmes, and Hardy 2006). (ii) holds because D is independent of (U, V) in graph CG after the edge $D \to V$ is removed by intervention. (iii) follows from the fact that the intervention does not affect the conditional distribution of W given (U, V, D), and the marginal distribution of D remains unchanged. As a result, we can compute the direct causal effect by using quantities $\mathbb{P}(D)$ and $\mathbb{E}[W \mid U, V, D]$ that both can be obtained using the observable distribution.

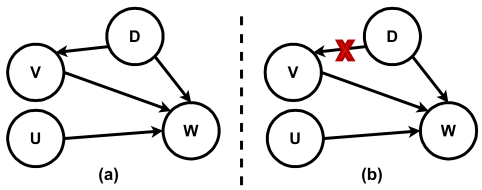

Figure 1: Causal graph of watch-time prediction.

Discretization Trick. To estimate the direct causal effect via the backdoor adjustment in Eq. (2), we need to compute both $\mathbb{P}(D)$ and $\mathbb{E}[W \mid U, V, D]$. A commonly used and easy-to-implement method (Zhan et al. 2022; Lin et al. 2023) for this is the discretization trick. Formally, the continuous variable D (duration) is discretized into non-overlapping intervals $(t_k, t_{k+1}]$, $k = 1, ..., M$. Let $D_k = \{d | d \in (t_k, t_{k+1}]\}$ be the set of durations that fall into the k-th interval. Using this discretization, the integral in Eq. (2) can be approximated by a Riemann sum:

$$\begin{aligned}
&\sum_d \mathbb{P}(D = d) \, \mathbb{E}[W \mid U, V, D = d] \\
&\approx \sum_{k=1}^{M} \mathbb{P}\{d \in D_k\} \, \mathbb{E}[W \mid U, V, d \in D_k].
\end{aligned} \tag{3}$$

This approximation introduces a trade-off. When $M = 1$, all durations fall into a single group, reducing the model to standard value regression without debiasing. As M increases, the Riemann approximation becomes more accurate, reducing the bias from duration. However, this comes at the cost of variance: as groups become smaller, the number of samples in each group decreases, leading to high estimation error. Formally, let $\mathcal{I}_k = \{i | d_i \in D_k\}$ be the set of indices whose durations fall in group k. Then the group-wise estimate of the conditional expectation can be approximated by

$$\mathbb{E}[W \mid U, V, d \in D_k] \approx \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} w_i \, , \tag{4}$$

where $|\mathcal{I}_k|$ denote the number of samples in group k. Hence, when $|\mathcal{I}_k|$ is too small, this estimate becomes unreliable due to high variance. In practical, it's important to balance approximation bias and estimation variance. Moreover, training a separate model for each group is often infeasible in real-world systems due to scalability concern. Therefore, aggregation strategies are typically adopted to share information across groups and produce a unified prediction model.

Aggregation Strategy. As discussed earlier, real-world applications typically require a single unified model that can aggregate information across duration groups while still accounting for the de-biasing objective introduced by backdoor adjustment. To address this, prior work (Zhan et al. 2022; Lin et al. 2023) proposes a strategy that learns a shared quantile prediction model across all duration groups and maps the predicted percentiles back to watch-time values using group-specific empirical distributions. Formally, let $\widehat{Q}_k$ denote the empirical cumulative distribution function

(ECDF) of watch time for duration group k, and let $q(u, v; \theta)$ be a shared model parameterized by θ that predicts the percentile (i.e., quantile) of watch time based on user and video features. The predicted watch time for a sample $i \in \mathcal{I}_k$ is then given by:

$$\widehat{w}_i = \widehat{Q}_k^{-1}[q(u_i, v_i; \theta)], \quad i \in \mathcal{I}_k, \tag{5}$$

where $\widehat{Q}_k^{-1}$ is the inverse of the empirical CDF in group k. And the train loss is designed as

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\big(q(u_i, v_i; \theta),\ \widehat{Q}_{k_i}(w_i)\big), \tag{6}$$

with k_i is the group of sample i, i.e., $d_i \in D_{k_i}$. This approach can be interpreted from two complementary perspectives: (1) From the bias correction perspective, it achieves debiasing by assigning group-specific labels via discretization. (2) From the aggregation view, it enables model sharing by training a single model $q(\cdot)$ to predict percentiles across all groups. Thus, the design conveniently addresses both goals — bias correction and model efficiency — through a unified, group-aware architecture.

The Pitfall. However, a key limitation remains: training with the loss in Eq. (6) can still be significantly affected by distributional differences across duration groups. To analyze this issue, we rewrite the objective as:

$$\frac{1}{n}\sum_{k=1}^{M}\sum_{i\in\mathcal{I}_k} \mathcal{L}\big(q(u_i, v_i; \theta),\ \widehat{Q}_{k_i}(w_i)\big). \tag{7}$$

where the shared model $q(u_i, v_i; \theta)$ is trained to predict stable quantile labels across all groups, while the input features (u_i, v_i) are drawn from a mixture of group-specific distributions. In this setting, the learning algorithm is vulnerable to model misspecification, especially when there is substantial variation in the marginal distribution $\mathbb{P}(U, V)$ across groups. In practice, this issue is exacerbated by the limited fidelity of the group-wise prior modeling and the presence of noise in real-world data. As a result, the approximated solution learned by the model becomes highly sensitive to shifts in $\mathbb{P}(U, V)$, reducing its generalization ability.

This problem is closely related to the challenges addressed in the Invariant Learning literature, which seeks to identify and leverage stable features across varying environments. We elaborate on this connection in the following section.

4.2 Duration-invariant Feature Learning

To address the limitation discussed in the previous section, here we propose an improved approach from the perspective of invariant learning. Specifically, we design a method that encourages the learned representations to be invariant to video duration, thereby reducing potential bias and improving generalization. We begin by briefly introducing the concept of invariant learning. The general framework is illustrated in Fig. 2. In this setting, training data comes from multiple domains $S = \{S_i\}_{i=1}^{N}$, where $S_i = \{(x_k^{(i)}, y_k^{(i)})\}_{k=1}^{n_i}$ is sampled from domain-specific distribution $\mathbb{P}_{XY}^{i}$. Invariant learning typically assumes that $\mathbb{P}_{XY}^{i} \neq \mathbb{P}_{XY}^{j}$ for $i \neq j$, meaning that samples in S are not independently and identically distributed (*i.i.d.*) across domains. Given this multi-domain setting, the goal is to learn a function $f : \mathfrak{B}_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}$ that generalizes well to unseen domains, where $\mathcal{X}$ is the input feature space and $\mathfrak{B}_{\mathcal{X}}$ is the set of probability distributions over $\mathcal{X}$. This is typically achieved by learning a stable conditional distribution $\mathbb{P}(Y|X)$ across domains. In practice, this amounts to learning an invariant representation $B(X)$ of the input X that facilitates cross-domain generalization.

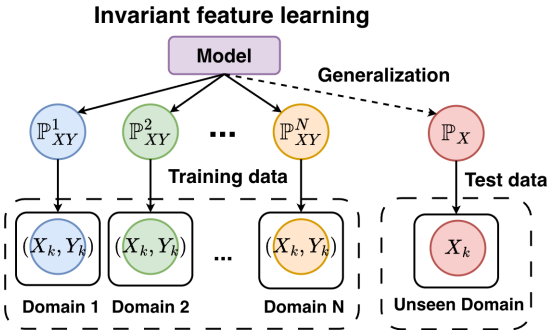

Figure 2: The framework of conventional invariant learning.

We observe a natural correspondence between our specific task with the invariant learning setup. In our case, groups correspond to domains, and the video-user feature pair (U, V) corresponds to the input feature X. The task involves modeling the conditional dependency between percentile labels and features — similar to learning a stable $\mathbb{P}(Y|X)$ across various domains. Inspired by this analogy, we seek to leverage invariant learning techniques to guide the learning of group-invariant features and mitigate aggregation bias. However, standard invariant learning methods generally assume that each domain contains *i.i.d.* samples. In contrast, the grouping in our task is based on duration, which is a continuous variable. As a result, intra-group samples are not strictly *i.i.d.*, and grouping schemes may introduce additional bias. Therefore, instead of applying conventional invariant learning directly, we propose a more suitable approach tailored to this structure.

Key Idea. We aim to learn a representation O of the input (U, V) that is both predictive of the label and invariant to duration. Specifically, O should satisfy two criteria: (1) Predictiveness: O should contain sufficient information to predict the label, which is guided by optimizing the prediction loss in Eq. (7). (2) Invariance: O should be independent of the duration variable D, i.e., $I(O, D) = 0$, where I denotes mutual information. This ensures duration-invariant of the learned representations.

To enforce the second constraint, we employ the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al. 2005), a measure that can detect both linear and nonlinear dependencies between random variables, and is more feasible to estimate than mutual information (Shannon 1948) .

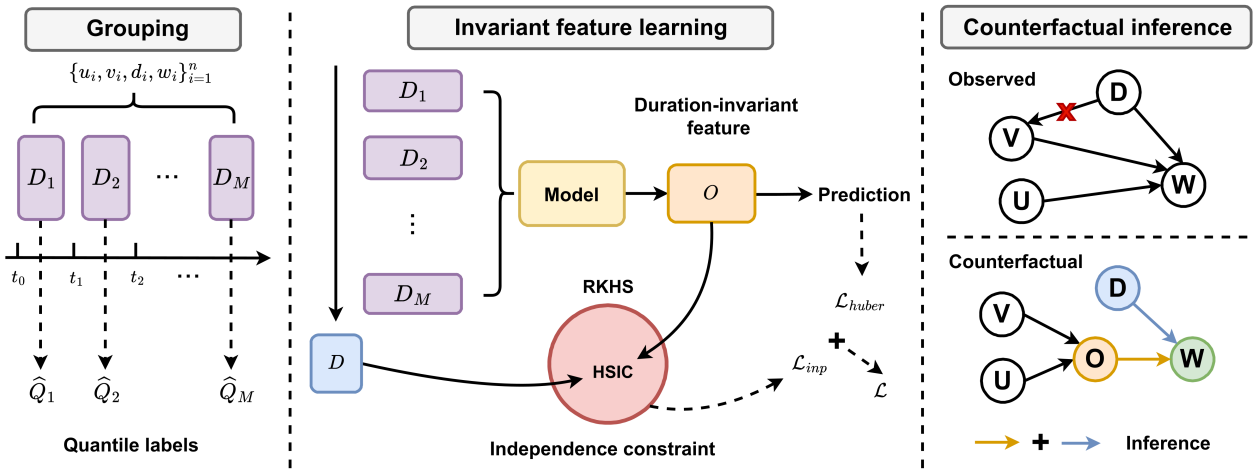

Figure 3: The framework of our Duration-Invariant Feature Learning (DIFL) method.

Duration Dependence Regulazation. HSIC measures the dependence between two variables using kernels in Reproducing Kernel Hilbert Space (RKHS). In our setting, the goal is

$$\text{HSIC}(O, D) = 0 \Longleftrightarrow O \perp\!\!\!\perp D, \tag{8}$$

which enforces that the representation O is independent of the duration D. And based on a minibatch of B i.i.d. samples $(o_i, d_i)_{i=1}^{B}$, the empirical estimate of HSIC is

$$\text{HSIC}_b(O, D) = \frac{1}{n^2}\text{Tr}(\mathbf{KHLH}), \tag{9}$$

where $\mathbf{K}$ is an $B \times B$ kernel matrix with entries $k_{ij} = k(o_i, o_j)$, $\mathbf{L}$ is an $B \times B$ kernel matrix with entries $l_{ij} = l(d_i, d_j)$, $\mathbf{H} = \mathbf{I} - \mathbf{1}\mathbf{1}^T/B$ is the centering matrix, where $\mathbf{1}$ is an $B \times 1$ vector of ones. However, the computation of HSIC has quadratic time complexity, i.e., $O(B^2)$, which can be computationally expensive when the batch size is too large. To address this issue, we adopt Random Fourier Features (RFF) (Rahimi and Recht 2007) to approximate the kernel computation, reducing the time complexity to $O(B \cdot T)$, where T is the dimensionality of the feature mapping (generally $T \ll B$). The RFF-based approximation of HSIC (Ren et al. 2024; Zhang et al. 2018) is given by

$$\text{HSIC}_\omega(O, D) = \frac{1}{B^2}\left\|\Phi_k^\top \mathbf{H}\Phi_l\right\|_F^2, \tag{10}$$

where $\Phi_k := [\phi_k(o_1), \phi_k(o_2), \cdots, \phi_k(o_B)]$ with the term

$$\phi_k(o) := \sqrt{\frac{2}{T}}\left[\cos(\omega_1^T o), ..., \sin(\omega_{D/2}^T o)\right]$$

with each $\omega_i \sim \mathcal{N}(0, \sigma_o^{-2}\mathbf{I})$ drawn independently from a Gaussian distribution. And the definition of Φ_l can be obtained by analogy. This approximation enables efficient estimation of HSIC while preserving its capacity to detect nonlinear dependencies.

In summary, up to now we have outlined the approach to learning duration-invariant representations. In the following, we present the details of our proposed framework.

4.3 The DIFL Framework

Our framework comprises three main components, as illustrated in Fig. 3. We describe each component in detail below.

Grouping. We split the data into M groups with equal frequency based on their durations. Within each group k, we compute the empirical cumulative distribution function (ECDF) $\widehat{Q}_k$ of watch time. For each sample i, the percentile-based label is calculated as

$$\widehat{Q}_k(w_i) = \frac{\sum_{j=1}^{n} \mathbf{1}\{d_j \in D_k \wedge w_j < w_i\}}{\sum_{j=1}^{n} \mathbf{1}\{d_j \in D_k\}}, \quad d_i \in D_k. \tag{11}$$

This within-group watch time percentile serves as the prediction target for our model in the next stage.

Invariant Feature Learning. To extract features that are invariant to video duration, we employ two loss terms: (i) a regression loss to guide the features O to effectively predict the target, and (ii) an independence loss to encourage invariance between the learned features and durations.

The feature representation O is obtained using a standard recommendation model backbone. We use DCN (Wang et al. 2017) in our implementation. A projection layer maps O to the predicted value $\hat{\tau}$, which is trained to approximate $\widehat{Q}_k(w)$. To enhance robustness to noise and outliers — common in recommendation scenarios — we use the Huber loss (Huber 1992) as the prediction objective

$$\mathcal{L}_{huber} = \begin{cases} \frac{1}{2}(\widehat{Q}_k(w) - \hat{\tau})^2 & \text{if } |\widehat{Q}_k(w) - \hat{\tau}| \le \delta, \\ \delta \cdot (|\widehat{Q}_k(w) - \hat{\tau}| - \frac{1}{2}\delta) & \text{otherwise,} \end{cases} \tag{12}$$

where we set the threshold $\delta = 0.2$ in our experiments.

To enforce duration-invariant, we introduce a second loss term that penalizes statistical dependence between the feature O and duration D, formulated using the RFF-approximated HSIC as Eq. (10),

$$\mathcal{L}_{inp} = \text{HSIC}_\omega(O, D). \tag{13}$$

Thus, the total loss is given by

$$\mathcal{L} = \mathcal{L}_{huber} + \lambda \cdot \mathcal{L}_{inp}, \tag{14}$$

where λ is the regularization weight that controls the trade-off between prediction accuracy and invariance.

Counterfactual Inference. Given the duration-invariant representation O, we can perform counterfactual prediction of watch time under different video durations. For each test sample i, the projected value $\hat{\tau}_i$ is shared across all groups, and we identify the corresponding group D_k based on its duration d_i. The final predicted watch time is then recovered by inverting the ECDF:

$$\widehat{w}_i = \widehat{Q}_k^{-1}(\hat{\tau}_i). \tag{15}$$

This approach allows us to predict the watch time under various hypothetical video durations, leveraging the invariant feature.

5 Experiments

In this section, we evaluate the performance of DIFL using real-world datasets through a series of experiments. Our code is available at: https://github.com/jinchenghou123/DIFL.

5.1 Experimental setting

Datasets Two publicly available datasets and one industrial dataset are utilized to evaluate the proposed method. The CIKM dataset is derived from the CIKM16 Cup competition and focuses on predicting user engagement duration during online search sessions. The KuaiRec dataset (Gao et al. 2022) is a real-world collection of video view logs from the Kuaishou App. We also conduct evaluation on a large-scale dataset (referred to as "Product") from a real video platform, which has 400 million daily active users (DAUs) and generates billions of impressions daily.

Compared Methods Following the experimental setups in prior studies (Lin et al. 2023; Sun et al. 2024), we compare DIFL against several state-of-the-art methods, including VR (Value Regression), WLR (Covington et al. 2016), D2Q (Zhan et al. 2022), CWM (Zhao et al. 2024), TPM (Lin et al. 2023) and CREAD (Sun et al. 2024).

Metrics Following the previous works (Lin et al. 2023; Sun et al. 2024; Ma et al. 2024), we use two performance metrics in our experiments: MAE and XAUC (Zhan et al. 2022).

5.2 Evaluation

Performance comparison. We compare our DIFL with six existing state-of-the-art (SOTA) models mentioned above, the results are presented in Tab. 1. DIFL consistently outperforms these baselines on all three datasets in terms of all metrics. Specifically, DIFL achieves a 1.519% decrease in MAE and a 3.968% increase in XAUC compared to the second-best method on the KuaiRec dataset. On the CIKM dataset, DIFL obtains a 3.849% reduction in MAE and a 2.964% improvement in XAUC. For the Product dataset, it surpasses CREAD with a 4.159% decrease in MAE and a 1.61% gain in XAUC. These significant performance gains across all three datasets demonstrate our method's superior ability to model users' true interests.

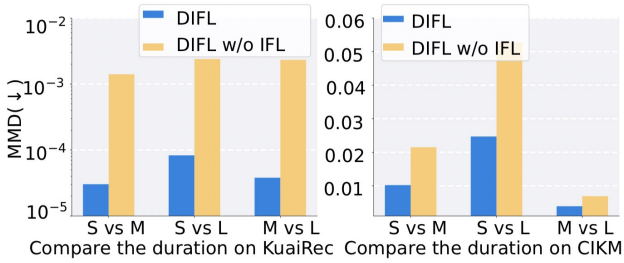

Figure 4: The effect of video duration on user interest distribution, results using maximum mean discrepancy (MMD).

Differences in user interest representations. In the previous section, DIFL demonstrates a markedly superior performance compared to existing SOTA methods. This advantage stems from DIFL's ability to constrain user interest embeddings to be independent of video duration, thereby to more effectively capture users' true preferences. To further illustrate the impact of our proposed Invariant Feature Learning (IFL) on user interest embeddings (i.e., feature representations produced by DCN), we divide all videos into three groups based on duration, i.e., short, medium, and long, and employe the Maximum Mean Discrepancy (MMD) (Borgwardt et al. 2006) to quantify the differences in user interest embeddings across these groups. Note that a larger MMD indicates greater distributional divergence. As shown on the left side of Fig. 4, for the Kuairec dataset, without IFL, the MMD values between the groups are consistently on the order of 10^{-3}. This implies that user interest representations remain implicitly influenced by video duration. However, after applying IFL, the distributional disparities between all the groups are significantly reduced. Specifically, the distributional differences between groups are reduced to between 1/15 and 1/50 of their original values.

On the CIKM dataset, which is inherently a search-based dataset, the item-side information consists of item IDs browsed by users within a session. Consequently, session duration time is affected by the length of the input item list, indirectly serving as an implicit duration feature that influences user interest embeddings, inevitably resulting in relatively higher MMD values. Nevertheless, by incorporating the regularization to control the relationship between user interest embeddings and video duration, the application of IFL reduces the MMD across the three groups by 52%, 53%, and 43%, respectively.

5.3 Ablation Study

In this section, we analyze the contribution of each component in the DIFL framework. Given that backdoor adjustment requires duration grouping, we specifically evaluate the roles of Invariant Feature Learning (IFL) and Counterfactual Inference (CI), with the results presented in Table 2. It can be observed that removing any component leads to

Method	KuaiRec				CIKM				Product			
	MAE	Improv.	XAUC	Improv.	MAE	Improv.	XAUC	Improv.	MAE	Improv.	XAUC	Improv.
VR	7.634	-	0.504	-	1.039	-	0.641	-	49.239	-	0.559	-
WLR	6.047	20.789 %	0.525	4.167 %	0.998	3.946 %	0.672	4.836 %	-	-	-	-
D2Q	3.512	53.995 %	0.569	12.897 %	0.858	17.421 %	0.676	5.460 %	-	-	-	-
CWM	3.452	54.781 %	0.580	15.079 %	0.891	14.244 %	0.662	3.276 %	-	-	-	-
TPM	3.456	54.729 %	0.571	13.294 %	0.850	18.191 %	0.676	5.460 %	42.303	14.086 %	0.565	1.073 %
CREAD	3.307	56.681 %	0.594	17.857 %	0.865	16.747 %	0.678	5.772 %	42.601	13.481 %	0.568	1.610 %
Ours (DIFL)	**3.191**	**58.200 %**	**0.614**	**21.825%**	**0.810**	**22.040 %**	**0.697**	**8.736 %**	**40.553**	**17.640 %**	**0.577**	**3.220 %**

Table 1: Performance comparison among different methods on KuaiRec, CIKM and Product. '-' means data unavailable.

Method		KuaiRec		CIKM	
IFL	CI	MAE	XAUC	MAE	XAUC
✓	-	3.304	0.597	0.835	0.681
-	✓	3.296	0.599	0.831	0.683
✓	✓	**3.191**	**0.614**	**0.810**	**0.697**

Table 2: Ablation study on KuaiRec and CIKM.

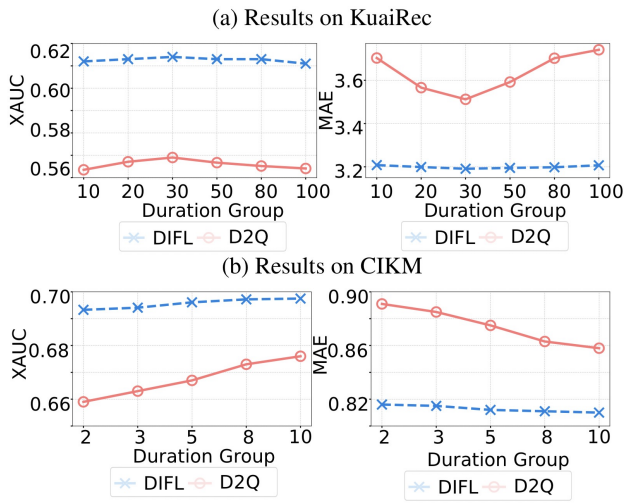

Figure 5: XAUC and MAE of DIFL and D2Q across different duration groups on (a) KuaiRec and (b) CIKM datasets.

a performance drop across all datasets. When IFL is removed, the user interest representations exhibit distribution shifts across duration groups, making it difficult to learn unified representations suitable for counterfactual inference. The strength of the independence constraint—controlled by the regularization weight λ in Eq. (14) — is discussed in detail in Appendix. When CI is removed, the estimation of $\widehat{Q}(w)$ in Eq. (11) is performed using the ECDF computed over the entire training set. However, since the ECDF varies across duration groups, which introduces errors in the recovery of watch time of Eq. (15).

Influence of the duration group number. Fig. 5 shows the impact of varying the number of duration groups on DIFL's performance, compared against the D2Q baseline, to assess the model's sensitivity and robustness to this parameter.

Method	DIFL		D2Q	
	MAE	XAUC	MAE	XAUC
Equal-Freq	**3.191**	**0.614**	3.512	0.569
Equal-Width	3.193	**0.614**	3.576	0.563

Table 3: Effect of different grouping methods.

On the Kuairec dataset, we explore the effect of varying the number of duration groups in Fig. 5(a). DIFL consistently outperforms D2Q across all groups in terms of MAE and XAUC metrics. On the CIKM dataset, duration corresponds to the length of the item list browsed by users, i.e., a discrete value with a maximum of 10. Accordingly, we change the number of duration groups in Fig. 5(b). It can be seen that, by learning invariant representations, DIFL maintains robust performance across diverse configurations.

Effects of different grouping methods. To further validate that our DIFL does not rely on sophisticated grouping strategies, we evaluate it on the Kuairec dataset using two grouping methods: equal-frequency bins and equal-width bins. As shown in Tab. 3, DIFL exhibits strong robustness across both grouping strategies. In contrast, the performance of D2Q is highly sensitive to the choice of grouping method, underscoring that DIFL can effectively capture invariant user interest representations, independent of grouping nuances.

6 Conclusion

In this paper, we revisit existing debiasing frameworks for watch-time prediction and highlight the limitation in their aggregation strategies. To address the limitation, we propose a novel framework named Duration-Invariant Feature Learning (DIFL) that is rooted in invariant learning principles. Extensive experiments on both public and large-scale real-world industrial datasets demonstrate the effectiveness of our approach, which achieves state-of-the-art performance.

Acknowledgements

Jihong Guan was supported by National Natural Science Foundation of China (NSFC) under grants No. 62172300 and No. 62372326. The computations in this research were performed using the CFFF platform of Fudan University.

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Borgwardt, K. M.; Gretton, A.; Rasch, M. J.; Kriegel, H.-P.; Schölkopf, B.; and Smola, A. J. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14): e49–e57.

Covington, P.; et al. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, 191–198.

Davidson, J.; Liebald, B.; Liu, J.; Nandy, P.; Van Vleet, T.; Gargi, U.; Gupta, S.; He, Y.; Lambert, M.; Livingston, B.; et al. 2010. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, 293–296.

Dong, L.; et al. 2024. Not All Videos Become Outdated: Short-Video Recommendation by Learning to Deconfound Release Interval Bias. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 179–188.

Gao, C.; Li, S.; Lei, W.; Chen, J.; Li, B.; Jiang, P.; He, X.; Mao, J.; and Chua, T.-S. 2022. KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 540–550.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, 63–77. Springer.

Huber, P. J. 1992. *Robust Estimation of a Location Parameter*, 492–518. New York, NY: Springer New York.

Li, D.; Li, W.; Lu, B.; Li, H.; Ma, S.; Krishnan, G.; and Wang, J. 2024. Delving Deep into Engagement Prediction of Short Videos. In *European Conference on Computer Vision*, 289–306. Springer.

Li, H.; et al. 2022a. StableDR: Stabilized doubly robust learning for recommendation on data missing not at random. *arXiv preprint arXiv:2205.04701*.

Li, K.; Shao, G.; Yang, N.; Fang, X.; and Song, Y. 2022b. Billion-user customer lifetime value prediction: an industrial-scale solution from Kuaishou. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 3243–3251.

Lin, C.; Wang, C.; Xie, A.; Wang, W.; Zhang, Z.; Ruan, C.; Huang, Y.; and Liu, Y. 2025. AlignPxtr: Aligning Predicted Behavior Distributions for Bias-Free Video Recommendations. *arXiv preprint arXiv:2503.06920*.

Lin, X.; Chen, X.; Song, L.; Liu, J.; Li, B.; and Jiang, P. 2023. Tree based progressive regression model for watch-time prediction in short-video recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4497–4506.

Liu, D.; Qiao, Y.; Tang, X.; Chen, L.; He, X.; and Ming, Z. 2023. Prior-guided accuracy-bias tradeoff learning for CTR prediction in multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 995–1003.

Ma, H.; Tian, K.; Zhang, T.; Zhang, X.; Chen, C.; Li, H.; Guan, J.; and Zhou, S. 2024. Generative Regression Based Watch Time Prediction for Video Recommendation: Model and Performance. *arXiv preprint arXiv:2412.20211*.

Ma, H.; Wang, G.; Yu, F.; Jia, Q.; and Ding, S. 2025. MS-DETR: Towards Effective Video Moment Retrieval and Highlight Detection by Joint Motion-Semantic Learning. *arXiv preprint arXiv:2507.12062*.

Malitesta, D.; Cornacchia, G.; Pomo, C.; Merra, F. A.; Di Noia, T.; and Di Sciascio, E. 2025. Formalizing multimedia recommendation through multimodal deep learning. *ACM Transactions on Recommender Systems*, 3(3): 1–33.

Min, Y.; et al. 2023. Environment-invariant curriculum relation learning for fine-grained scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13296–13307.

Peters, J.; et al. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5): 947–1012.

Rahimi, A.; and Recht, B. 2007. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20.

Ren, Y.; Xia, Y.; Zhang, H.; Guan, J.; and Zhou, S. 2024. Efficiently Learning Significant Fourier Feature Pairs for Statistical Independence Testing. *Advances in Neural Information Processing Systems*, 37: 99800–99835.

Ren, Y.; Zhang, H.; Xia, Y.; Guan, J.; and Zhou, S. 2023. Multi-level wavelet mapping correlation for statistical dependence measurement: methodology and performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 6499–6506.

Ren, Y.; Zhang, J.; Xia, Y.; Wang, R.; Xie, F.; Guan, J.; Zhang, H.; and Zhou, S. 2025. Regression-based Conditional Independence Test with Adaptive Kernels. *Artificial Intelligence*, 104391.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Sun, J.; Ding, Z.; Chen, X.; Chen, Q.; Wang, Y.; Zhan, K.; and Wang, B. 2024. Cread: A classification-restoration framework with error adaptive discretization for watch time prediction in video recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 9027–9034.

Wang, R.; Fu, B.; Fu, G.; and Wang, M. 2017. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*, 1–7.

Wang, W.; Feng, F.; He, X.; Wang, X.; and Chua, T.-S. 2021. Deconfounded recommendation for alleviating bias amplification. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 1717–1725.

Wang, Z.; He, Y.; Liu, J.; Zou, W.; Yu, P. S.; and Cui, P. 2022. Invariant preference learning for general debiasing in recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1969–1978.

Wei, T.; Feng, F.; Chen, J.; Wu, Z.; Yi, J.; and He, X. 2021. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 1791–1800.

Weiss, N. A.; Holmes, P. T.; and Hardy, M. 2006. *A course in probability*. Pearson Addison Wesley Boston, MA, USA:.

Wu, P.; Li, H.; Deng, Y.; Hu, W.; Dai, Q.; Dong, Z.; Sun, J.; Zhang, R.; and Zhou, X.-H. 2022. On the opportunity of causal learning in recommendation systems: Foundation, estimation, prediction and challenges. *arXiv preprint arXiv:2201.06716*.

Wu, S.; et al. 2018. Beyond views: Measuring and predicting engagement in online videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Yang, S.; Yang, H.; Du, L.; Ganesh, A.; Peng, B.; Liu, B.; Li, S.; and Liu, J. 2024. SWaT: Statistical Modeling of Video Watch Time through User Behavior Analysis. *arXiv preprint arXiv:2408.07759*.

Yue, Z.; Wang, T.; Sun, Q.; Hua, X.-S.; and Zhang, H. 2021. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15404–15414.

Zhan, R.; Pei, C.; Su, Q.; Wen, J.; Wang, X.; Mu, G.; Zheng, D.; Jiang, P.; and Gai, K. 2022. Deconfounding duration bias in watch-time prediction for video recommendation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 4472–4481.

Zhang, Q.; Filippi, S.; Gretton, A.; and Sejdinovic, D. 2018. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28: 113–130.

Zhang, X.; Li, J.; Chu, W.; Hai, J.; Xu, R.; Yang, Y.; Guan, S.; Xu, J.; and Cui, P. 2024. On the out-of-distribution generalization of multimodal large language models. *arXiv preprint arXiv:2402.06599*.

Zhang, Z.; Gao, H.; Yang, H.; and Chen, X. 2023. Hierarchical invariant learning for domain generalization recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3470–3479.

Zhao, H.; Cai, G.; Zhu, J.; Dong, Z.; Xu, J.; and Wen, J.-R. 2024. Counteracting Duration Bias in Video Recommendation via Counterfactual Watch Time. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4455–4466.

Zhao, H.; Zhang, L.; Xu, J.; Cai, G.; Dong, Z.; and Wen, J.-R. 2023. Uncovering user interest from biased and noised watch time in video recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*, 528–539.

Zheng, Y.; Gao, C.; Ding, J.; Yi, L.; Jin, D.; Li, Y.; and Wang, M. 2022. Dvr: micro-video recommendation optimizing watch-time-gain under duration bias. In *Proceedings of the 30th ACM International Conference on Multimedia*, 334–345.