DADF: A Distribution-Aware Debiasing Framework for Watch-Time Regression in Recommender Systems

Yiqing Yang*
Kuaishou Technology
Beijing, China
yangyiqing06@kuaishou.com

Xinlong Zhao*
Kuaishou Technology
Beijing, China
zhaoxinlong03@kuaishou.com

Zhao Liu*
Kuaishou Technology
Beijing, China
liuzhao09@kuaishou.com

Xiao Lv†
Kuaishou Technology
Beijing, China
lvxiao03@kuaishou.com

Ruiming Tang†
Kuaishou Technology
Beijing, China
tangruiming@kuaishou.com

Han Li
Kuaishou Technology
Beijing, China
lihan08@kuaishou.com

Kun Gai
Unaffiliated
Beijing, China
gai.kun@qq.com

Abstract

Watch-time prediction is a central regression task in short-video recommender systems, where labels are highly long-tailed and residual errors vary systematically across observed watch-time regions. In practice, a model may appear globally calibrated while still overestimating short views and underestimating long views, because opposite errors cancel out in aggregate. Existing methods mainly improve the first-stage watch-time predictor, but often leave such residual distributional bias insufficiently corrected. We propose DADF, a distribution-aware debiasing framework for watch-time regression. Instead of replacing a deployed predictor, DADF performs second-stage multiplicative residual correction on top of it. DADF combines three complementary designs: a dynamic distribution-aware transformation for stabilizing long-tailed correction targets, a debias-factor-aware module for modeling heterogeneous residual patterns using inference-time observable factors, especially video duration, and a multi-label-aware module that exploits auxiliary prediction signals from engagement heads. We evaluate DADF on public short-video benchmarks and a large-scale industrial ranking system. DADF consistently improves both pointwise accuracy and ranking quality across datasets and backbones. In the industrial setting, it achieves a 1.88 percentage-point WUAUC gain over the production baseline, reduces MAE by 12.57%, and yields a statistically significant 0.347% lift in average time spent per device in online A/B testing. These results demonstrate that DADF effectively mitigates local calibration bias and provides a practical plug-in solution for debiasing long-tailed continuous targets. The source code is available at https://github.com/liuzhao09/DADF.

CCS Concepts

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

Keywords

Short Video Recommendation, Watch Time Prediction, Debiasing

1 Introduction

Watch time is a central continuous target in short-video recommender systems. Compared with binary feedback such as clicks or likes, watch time provides a more fine-grained signal of user consumption depth and is therefore widely used to optimize ranking quality, user engagement, and long-term retention. In industrial ranking systems, inaccurate watch-time prediction not only hurts offline regression quality, but also directly distorts exposure allocation and traffic distribution in downstream ranking.

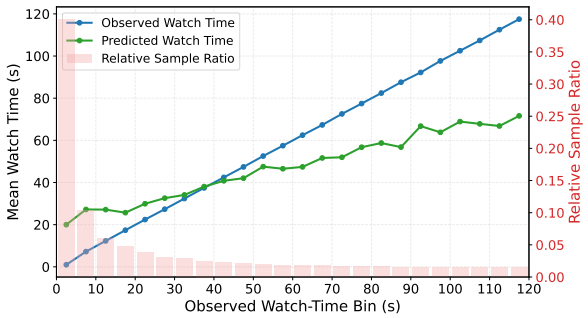
In this work, we focus on the *multiplicative bias factor* of a deployed predictor, defined as the ratio between observed watch time and predicted watch time, i.e., $y/\hat{y}$. Intuitively, this factor measures how much the first-stage predictor underestimates or overestimates the true watch time. A value larger than 1 indicates underestimation, while a value smaller than 1 indicates overestimation. Our goal is not to replace the deployed predictor, but to learn a second-stage correction module that models and removes this systematic bias.

Accurate watch-time prediction is challenging because the label distribution is highly imbalanced: most impressions correspond to short views or quick skips, while only a small fraction lead to long consumption. This produces a watch-time distribution that is both long-tailed [35] and multi-peaked [50]. Under such a distribution, a predictor can achieve seemingly reasonable aggregate performance while still making systematic errors in different watch-time regions. As shown in Figure 1a, the deployed predictor consistently overestimates short views and underestimates long views across observed watch-time buckets. These opposite errors cancel out in aggregate, yielding an apparently well-calibrated global ratio while leaving substantial local bias unresolved. We refer to this phenomenon as *pseudo-balance*, where error cancellation masks systematic miscalibration across watch-time regions.

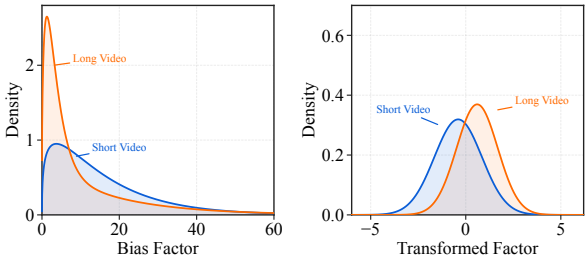
Existing methods mainly improve watch-time prediction from two directions. The first line improves the first-stage predictor itself, including discretization-based methods such as TPM [19] and CREAD [35], duration-aware methods such as D2Q [47] and D2CO [49], and distributional modeling methods such as EGMN [50]. The second line performs post-hoc correction, among which TranSUN [46] is the closest related work. However, these methods do not directly address the setting studied here: a deployed first-stage predictor with systematic residual bias that varies across regimes observable before ranking decisions.

This limitation is especially important in industrial systems, where the first-stage watch-time predictor is tightly coupled with ranking logic and serving infrastructure, and replacing it often requires substantial engineering changes and costly online validation.

*Equal contribution.
†Corresponding author.



(a) Local bias across observed watch-time buckets.



(b) Distribution-aware transformation of heterogeneous bias factors.

Figure 1: Motivation of DADF. (a) Even when the overall prediction-to-observation ratio is close to 1.0, the deployed predictor can still systematically overestimate short views and underestimate long views across observed watch-time buckets. (b) Raw bias-factor distributions under different debias-factor regimes can be highly heterogeneous and skewed; a distribution-aware transformation maps them into a more aligned and stable space for residual correction.

A more practical solution is therefore to keep the deployed predictor fixed and learn a lightweight second-stage module that directly models the multiplicative bias factor. Moreover, bias-factor distributions can differ substantially across debias-factor regimes. As illustrated in Figure 1b, even two representative cases, short videos and long videos, can exhibit markedly different skewed bias distributions in the raw space, while a distribution-aware transformation maps them into a more aligned and stable space for correction. Figure 1b gives the intuition, and Figure 6 later provides a detailed empirical comparison between raw and transformed distributions.

To this end, we propose **DADF**, a **D**istribution-**A**ware **D**ebiasing **F**ramework for watch-time prediction. DADF performs second-stage multiplicative correction on top of an existing first-stage predictor. It uses inference-time observable video duration to define debias-factor groups, applies a dynamic distribution-aware transformation to stabilize long-tailed bias factors, and further incorporates auxiliary prediction signals from engagement heads to improve correction quality. In this way, DADF serves as a minimally intrusive, model-agnostic plug-in for existing watch-time prediction pipelines with limited serving changes.

Our main contributions are summarized as follows.

- We propose **DADF**, a distribution-aware second-stage debiasing framework that models multiplicative bias factors and performs residual correction without replacing the deployed first-stage predictor in production.
- We design three complementary components, including dynamic distribution-aware transformation, debias-factor-aware correction, and multi-label-aware auxiliary representation learning, to improve the stability and expressiveness of second-stage residual correction under long-tailed labels.
- We conduct extensive offline and online experiments on public datasets and a large-scale industrial ranking system, showing consistent gains in MAE and XAUC on public benchmarks, a 1.88 percentage-point WUAUC gain in production offline evaluation, and a statistically significant 0.347% online improvement in average time spent per device.

2 Related Work

2.1 Watch-Time Prediction in Recommender Systems

Watch-time and dwell-time prediction are central problems in modern recommender systems, especially for online video and short-video platforms. Early industrial systems, such as YouTube recommendation [8], already incorporated post-click engagement signals into ranking. Subsequent studies further explored dwell-time personalization [45], engagement-aware video recommendation [43], dwell-time distribution modeling [40], real-time short-video ranking [14], retention-oriented recommendation [7], and passive-negative feedback modeling [28]. Overall, this line of work focuses on improving the prediction of raw watch-time or engagement labels for downstream ranking.

This problem has also been studied on several public benchmarks, including KuaiRec [12] and KuaiRand [13], together with representative modeling methods such as TPM [19], D2Q [47], CREAD [35], D2CO [49], conditional quantile estimation [18], and EGMN [50]. Most of these methods aim to build a stronger first-stage predictor for the target watch-time label itself, rather than a post-hoc residual correction layer after deployment.

Continuous engagement labels have also been modeled with more general distributional approaches, including mixture density networks [4], finite mixture models [22], Gaussian mixtures [53], exponential-family distributions [1], Tweedie models [32], beta regression [10], Poisson inverse Gaussian regression [27], normalizing flows [29], and optimal-distribution learning [42]. DADF is complementary to this line of work: rather than replacing the first-stage distributional predictor, it explicitly corrects residual bias left by the fixed predictor after deployment.

2.2 Duration Bias and Debiasing

Duration bias arises because video duration affects both observed watch time and user feedback. Prior work studies duration and exposure bias through D2Q [47], DVR [51], D2CO [49], counterfactual watch time [48], deconfounded recommendation [41], unbiased implicit-feedback learning [30], and causal embeddings [5]. Our setting is different: we start from a trained first-stage predictor and focus on systematic residual bias correction. Duration is not

only a confounder, but also a debias factor that partitions residual distributions into heterogeneous regimes.

2.3 Transformed Regression and Retransformation Bias

Transformed regression maps a skewed continuous label into a more regular optimization space and then applies an inverse transformation at inference time, as in square-root transformations [2], general transformations [3], Box–Cox transformations [6], and Box–Cox reviews [31]. Nonlinear inverse transformation can introduce retransformation bias, a classical problem studied for lognormal variables [11], transformation bias [26], smearing estimates [9], hydrologic prediction [16], log-transformed regression [25], Bayesian correction [33], a posteriori correction [34], and forecasting models [24]. TranSUN [46] addresses retransformation bias by learning a multiplicative correction factor for transformed regression. DADF is related to TranSUN, but differs in three aspects: it applies the transformation to residual factors rather than only raw labels, it uses distribution-aware transformations and experts rather than only a global correction, and it incorporates multi-label engagement signals to reduce residual uncertainty.

2.4 Auxiliary Signals and Multi-Task Recommendation

Large-scale recommenders commonly share representations across related prediction heads. Multi-task modules such as MMoE [20], PLE [37], ESMM [21], and AITM [44], together with attention [38], DIN [52], DCN [39], and SENet [15], provide useful building blocks for modeling auxiliary behaviors. In DADF, auxiliary objectives that are relevant to watch-time debiasing are selected, and their logits and tower representations are used as side information to estimate the second-stage correction factor.

3 Problem Formulation

We consider post-hoc two-stage correction for watch-time regression, inspired by the multiplicative correction paradigm of TranSUN [46]. For each impression, a fixed first-stage predictor produces a base watch-time prediction $\hat{y}_0 \in \mathbb{R}_+$, while the observed watch time $y \in \mathbb{R}_+$ is used as the training label. Let $s = (x, d, \ell_0, a)$ denote the inference-time correction signals, including ranking features x, video duration d, the raw watch-time output ℓ_0 before the final output mapping, and auxiliary representations a from related engagement heads. Instead of replacing the first-stage predictor, our goal is to learn a lightweight correction function $c_\theta(s) \in \mathbb{R}_+$ that estimates a multiplicative correction factor:

$$\hat{b} = c_\theta(s), \qquad \hat{y} = \hat{y}_0 \cdot \hat{b}, \tag{1}$$

where $\hat{b}$ is the predicted correction factor and $\hat{y}$ is the corrected watch-time prediction used by the downstream ranker.

During training, we define the multiplicative correction label as:

$$b = \frac{y}{\mathrm{sg}(\hat{y}_0) + \epsilon}, \tag{2}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and $\epsilon > 0$ prevents numerical instability when $\hat{y}_0$ is close to zero. The first-stage output mapping ensures $\hat{y}_0 \geq 0$ in practice. At inference time, since y is unavailable,

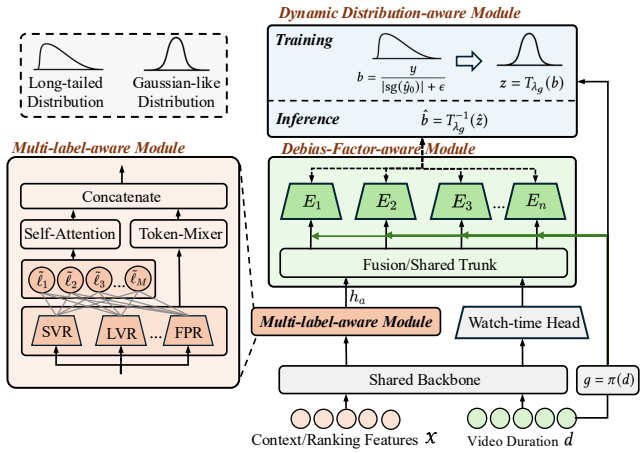


Figure 2: Overview of DADF. The framework corrects an existing watch-time predictor through distribution-aware, factor-aware, and multi-label-aware residual modeling.

DADF estimates $\hat{b} = c_\theta(s)$ from inference-time signals and predicts $\hat{y} = \hat{y}_0 \cdot \hat{b}$. Thus, DADF reduces post-hoc debiasing to estimating a stable multiplicative correction factor on top of a fixed first-stage watch-time predictor during serving.

4 Method

Based on the two-stage correction setting, DADF implements a distribution-aware multiplicative correction network with three modules. The Dynamic Distribution-aware Module maps long-tailed correction labels into a stable transformed space. The Debias-Factor-aware Module uses duration groups to capture residual heterogeneity and predict transformed corrections. The Multi-label-aware Module extracts behavioral representations from first-stage auxiliary heads. During inference, DADF applies the group-specific inverse transformation and multiplies the recovered correction factor with the first-stage prediction, serving as a lightweight plug-in for residual debiasing in ranking systems.

4.1 Dynamic Distribution-aware Module

The multiplicative correction label b is difficult to regress directly. Since watch-time labels are highly long-tailed, the ratio-style correction factor may inherit and even amplify such heavy-tailed fluctuations, especially when the base prediction $\hat{y}_0$ is small. As empirically shown in Appendix A.1, the raw correction factor exhibits clear right-skewness and heavy-tailed behavior. Such a target is poorly suited for direct regression, because extreme values dominate optimization and make residual learning unstable. Therefore, DADF first maps the raw correction factor into a more regular transformed space, where the target becomes more concentrated and closer to symmetric after a Box–Cox-style transformation.

A single global transformation, however, may still be insufficient, because the distribution of b varies across duration groups, with different skewness, dispersion, and residual bias patterns. To account for this heterogeneity, DADF assigns each sample to a duration group according to video duration, i.e., $g = \pi(d)$, where

$\pi(\cdot)$ is the duration bucketing function, and learns a group-specific Box–Cox-style transformation within each group. Specifically,

$$T_{\lambda_g}(b) = \begin{cases} \frac{(b+\epsilon)^{\lambda_g}-1}{\lambda_g}, & \lambda_g \neq 0, \\ \log(b+\epsilon), & \lambda_g = 0, \end{cases} \tag{3}$$

where ϵ is a small positive constant for numerical stability, and λ_g is the learnable transformation parameter optimized jointly with the correction network for duration group g. By learning an independent λ_g for each duration group, DADF adaptively adjusts the transformation strength according to the group-specific correction-factor distribution, avoiding the limitation of forcing all samples to share a single global transformation.

After the correction network predicts $\hat{z}$ in the transformed space, DADF maps it back to the multiplicative-factor space through the corresponding group-specific inverse transformation:

$$T_{\lambda_g}^{-1}(z) = \begin{cases} (\lambda_g z + 1)^{1/\lambda_g} - \epsilon, & \lambda_g \neq 0, \\ \exp(z) - \epsilon, & \lambda_g = 0. \end{cases} \tag{4}$$

Thus, this module transforms heavy-tailed multiplicative correction factors into a more stable optimization space while preserving the semantics of multiplicative correction. Duration-specific parameters further allow the transformation strength to adapt to heterogeneous residual distributions across groups.

4.2 Debias-Factor-aware Module

In the duration-adaptive transformed space defined in Section 4.1, the Debias-Factor-aware Module predicts the transformed correction $\hat{z}$ from inference-time signals. Specifically, DADF fuses the correction signals into a unified representation, including ranking features x, the first-stage raw watch-time output ℓ_0, the duration group $g = \pi(d)$, and the multi-label auxiliary representation h_a:

$$h = \text{Fusion}(x, \ell_0, g, h_a), \tag{5}$$

where h denotes the fused representation for correction prediction, and $\text{Fusion}(\cdot)$ is a learnable feature fusion network.

The motivation for duration-group-conditioned prediction can be understood from an oracle-level perspective. Let U denote the correction target in the transformed space and G denote the duration group. By the law of total variance,

$$\text{Var}(U) = \mathbb{E}[\text{Var}(U \mid G)] + \text{Var}(\mathbb{E}[U \mid G]). \tag{6}$$

Therefore, under squared loss, the oracle optimal risk when conditioning on duration groups satisfies

$$\mathcal{R}_G^* = \mathbb{E}[\text{Var}(U \mid G)] \leq \text{Var}(U) = \mathcal{R}_0^*, \tag{7}$$

where $\mathcal{R}_G^*$ and $\mathcal{R}_0^*$ denote the oracle risks with and without group information, respectively. This suggests that duration-group conditioning can reduce the oracle-level uncertainty caused by mixing heterogeneous residual regimes. A detailed proof is provided in Appendix A.2 for completeness.

Motivated by this observation, DADF adopts hard routing conditioned on g. Each sample selects the expert network associated with its duration group and predicts the transformed correction as:

$$\hat{z} = E_g(h), \tag{8}$$

where E_g denotes the expert associated with duration group g. By dedicating a separate expert to each duration regime, DADF avoids forcing heterogeneous samples to share a single correction function, thereby reducing cross-group interference during residual learning.

4.3 Multi-label-aware Module

The Debias-Factor-aware Module in Section 4.2 uses the multi-label auxiliary representation h_a to enrich correction-factor prediction. This module constructs h_a from first-stage auxiliary heads, whose logits and tower representations are available at inference time. Auxiliary engagement tasks, such as completion, effective view, long view, negative feedback, and thresholded watch labels, are closely related to watch-time bias. Although their observed labels are unavailable at inference time, their predicted signals provide useful conditional information for second-stage correction. To exploit the behavioral semantics encoded by such multi-task heads, e.g., MMoE [20] and PLE [37], DADF extracts correction-relevant signals from auxiliary-task logits and task-specific tower representations for correction-factor estimation.

As illustrated in Figure 2, the Multi-label-aware Module contains three information branches. First, DADF combines the playtime-related auxiliary logit with common ranking feature representation to obtain a basic correction-context representation:

$$h_c = \text{MLP}\left([\ell_{\text{play}}, x_c]\right), \tag{9}$$

where ℓ_{play} denotes the playtime-related auxiliary logit, which is distinct from the raw watch-time output ℓ_0, and x_c denotes the common feature representation shared by the first-stage auxiliary heads. Here, $[\cdot]$ denotes concatenation.

Second, to alleviate scale mismatch among auxiliary logits, DADF applies task-specific nonlinear projections to each auxiliary-task logit:

$$\tilde{\ell}_m = \Phi_m(\ell_m), \quad m = 1, \ldots, M, \tag{10}$$

where ℓ_m denotes the logit of the m-th auxiliary task, $\Phi_m(\cdot)$ denotes the corresponding task-specific nonlinear projection, and $\tilde{\ell}_m$ denotes the projected auxiliary logit. The projected logits are then fed into a self-attention layer to model cross-task dependencies:

$$h_\ell = \text{SelfAttention}\left(\tilde{\ell}_1, \tilde{\ell}_2, \ldots, \tilde{\ell}_M\right), \tag{11}$$

where h_ℓ captures correlations and complementarity among auxiliary behavior signals for correction.

Third, DADF leverages tower representations from auxiliary tasks and treats them as semantic tokens for cross-task interaction:

$$h_r = \text{TokenMixer}\left(r_1, r_2, \ldots, r_M\right), \tag{12}$$

where r_m denotes the tower representation of the m-th auxiliary task, and h_r preserves richer intermediate semantics than logits alone for estimating corrections.

Finally, the three branch outputs are concatenated and passed through an MLP to produce the multi-label-aware representation:

$$h_a = \text{MLP}\left([h_c, h_\ell, h_r]\right), \tag{13}$$

where h_a is fed into the Debias-Factor-aware Module as auxiliary behavioral information. By integrating both logits and tower representations, h_a provides a more comprehensive characterization of user consumption strength and preference signals than isolated auxiliary signals, thereby improving the stability of correction-factor estimation under sparse feedback.

4.4 Training Objective and Inference

After freezing the first-stage watch-time predictor, DADF optimizes the second-stage correction module in both the transformed correction space and the original watch-time space. For a sample with duration group $g = \pi(d)$, the correction network predicts the transformed correction as:

$$\hat{z} = F_\theta(x, \ell_0, g, h_a), \tag{14}$$

where $F_\theta(\cdot)$ denotes the overall second-stage correction network, including feature fusion and the duration-routed expert defined in Section 4.2. The predicted transformed correction is then mapped back to the multiplicative-factor space and applied to the base prediction:

$$\hat{b} = T_{\lambda_g}^{-1}(\hat{z}), \qquad \hat{y} = \hat{y}_0 \cdot \hat{b}, \tag{15}$$

where $\hat{b}$ is the predicted multiplicative correction factor and $\hat{y}$ is the corrected watch-time prediction.

The primary supervision is imposed in the transformed correction space:

$$\mathcal{L}_{\text{trans}} = \ell(z, \hat{z}), \tag{16}$$

where $z = T_{\lambda_g}(b)$ is the group-specific transformed correction target, and $\ell(\cdot,\cdot)$ denotes a regression loss such as MSE. This loss allows DADF to learn correction patterns in a more stable space with reduced long-tailedness and skewness.

To align the correction with the original watch-time target, we further introduce an absolute-space Huber loss:

$$\mathcal{L}_{\text{abs}} = \ell_{\text{Huber}}(y, \hat{y}), \tag{17}$$

where y is the observed watch time. This term directly constrains the final prediction after inverse transformation and multiplicative correction in the original label space.

To stabilize the transformed correction space, DADF regularizes the transformed correction targets within each duration group:

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \left(w_\mu \mu_g^2 + w_\sigma \left(\sigma_g^2 - 1\right)^2 + w_s \left|\text{Skew}_g\right| \right), \tag{18}$$

where $\mathcal{G}$ is the set of duration groups. For each group g, μ_g, σ_g^2, and Skew_g denote the mini-batch mean, variance, and skewness of $z = T_{\lambda_g}(b)$ for samples in group g. The coefficients w_μ, w_σ, and w_s control the strengths of the corresponding moment regularization terms. In practice, groups with insufficient samples are masked to avoid unstable moment estimation. This regularization encourages the transformed targets to be centered, normalized, and less skewed within each duration group, thereby stabilizing second-stage correction learning across groups during optimization.

The final second-stage objective is:

$$\mathcal{L}_{\text{DADF}} = \alpha \mathcal{L}_{\text{trans}} + \beta \mathcal{L}_{\text{abs}} + \eta \mathcal{L}_{\text{reg}}, \tag{19}$$

where α, β, and η balance transformed-space fitting, absolute-space accuracy, and moment regularization.

During inference, DADF uses only inference-time signals and does not rely on observed labels or training-only targets. Given the base prediction $\hat{y}_0$ and inference-time signals x, d, ℓ_0, and h_a, it assigns the sample to $g = \pi(d)$ and produces the final prediction as:

$$\hat{y} = \hat{y}_0 \cdot T_{\lambda_{\pi(d)}}^{-1}\left(F_\theta(x, \ell_0, \pi(d), h_a)\right). \tag{20}$$

Thus, DADF reuses the fixed first-stage prediction $\hat{y}_0$ as the multiplicative reference while learning a lightweight duration-conditioned correction factor for residual debiasing.

5 Experiments

We conduct comprehensive offline and online experiments to evaluate DADF from four perspectives: general offline effectiveness, production impact, component contribution, and bias-oriented analysis. Specifically, we aim to answer the following research questions:

- **RQ1**: Is DADF an effective plug-in correction module across different watch-time prediction backbones and public datasets?
- **RQ2**: Can DADF deliver statistically significant user-engagement gains in a real industrial ranking system?
- **RQ3**: How much does each proposed component contribute to the final performance?
- **RQ4**: Beyond aggregate metrics, does DADF better correct local bias across observed watch-time regions and improve performance on long-duration tail slices?

5.1 Experiment Settings

5.1.1 Datasets. We conduct offline experiments on two public short-video recommendation datasets with watch-time feedback. These datasets cover different data collection protocols and platform characteristics, enabling us to evaluate whether DADF generalizes beyond a single benchmark and sparsity pattern.

- **KuaiRec**[1] [12]: A fully observed short-video recommendation dataset collected from Kuaishou logs. It contains 12,530,806 impressions from 7,176 users and 10,728 videos. Since KuaiRec provides dense user-video interactions, it is suitable for evaluating watch-time prediction under reduced exposure-selection bias.
- **WeChat21**[2]: A large-scale short-video dataset released by the WeChat Big Data Challenge 2021. It contains 7,310,108 interaction logs between 20,000 users and 96,418 videos. Compared with KuaiRec, WeChat21 is larger and sparser, providing a complementary benchmark for testing the robustness of DADF in large-scale short-video recommendation.

5.1.2 Baselines. We compare DADF with two groups of baselines: (i) representative first-stage watch-time prediction models, and (ii) a second-stage multiplicative correction baseline. The first group evaluates whether DADF can improve different types of watch-time predictors, while the second group examines whether the proposed distribution-aware correction is more effective than an existing multiplicative correction strategy.

- **VR**: a direct value-regression baseline that predicts watch time with a pointwise regression loss.
- **WLR** [8]: a weighted logistic regression baseline for expected watch-time prediction, where positive samples are weighted by their observed watch time during training.
- **TPM** [19]: a tree-based progressive model that builds a hierarchy of regressors for watch-time prediction.
- **D2Q** [47]: a duration-aware debiasing method that models watch time under a duration-aware treatment framework.
- **CREAD** [35]: a classification-restoration framework using calibrated discretization and restoration for watch-time prediction.

[1] https://kuairec.com/
[2] https://algo.weixin.qq.com/

- **D2CO** [49]: a debiasing method that learns user interest from biased and noisy watch-time feedback.
- **EGMN** [50]: a distributional watch-time model that represents the target with an exponential-Gaussian mixture network.
- **TranSUN** [46]: a second-stage correction baseline that applies multiplicative calibration to transformed regression.

For each first-stage backbone, we report the reproduced backbone itself, the backbone equipped with TranSUN, and the backbone equipped with DADF. This controlled protocol allows us to evaluate whether DADF provides consistent gains beyond both the original watch-time predictor and an existing multiplicative correction method under the same frozen backbone.

5.1.3 Metrics. Following previous watch-time prediction studies [19, 35, 47, 50], we adopt MAE and XAUC to evaluate offline performance. These two metrics measure pointwise regression accuracy and ranking consistency, respectively.

- **MAE** [23] measures the average absolute error between the predicted watch time $\hat{y}_i$ and the observed watch time y_i, which is defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i|, \tag{21}$$

  where N denotes the number of test samples. A lower MAE indicates better pointwise prediction accuracy.
- **XAUC** [47] measures the consistency between the predicted watch-time order and the observed watch-time order, which is defined as:

$$\mathrm{XAUC} = \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega} \mathbf{1}\left[(\hat{y}_i - \hat{y}_j)(y_i - y_j) > 0\right], \tag{22}$$

  where $\Omega = \{(i,j) \mid y_i \neq y_j\}$ denotes the set of comparable sample pairs, and $\mathbf{1}[\cdot]$ is the indicator function. A higher XAUC indicates better ranking performance.

For the Kwai online A/B test, we further report user-level engagement metrics, including watch time, video plays, retention, completion behavior, and watch-depth-related rates.

5.1.4 Implementation Details. We use a random 8:1:1 split for training, validation, and testing on both public datasets. All first-stage backbones are trained under the same feature preprocessing and comparable model capacity, with 16-dimensional sparse-feature embeddings and MLP layers of hidden dimensions 256, 128, and 64. After each backbone is trained, its predictions are frozen and shared by TranSUN and DADF, ensuring that their comparison reflects only the second-stage correction strategy. For DADF, we use equal-frequency duration buckets in offline experiments, with $K = 4$ for KuaiRec and $K = 3$ for WeChat21; in the online deployment, we use four fixed-duration buckets based on historical traffic statistics and production experience. For the objective in Eq. 19, we set $\alpha = 1.0$, $\beta = 0.8$, and $\eta = 0.10$, while the remaining hyperparameters are selected on the validation set to avoid test-set tuning.

5.2 Main Results

5.2.1 Offline Main Results (RQ1). Table 1 reports the offline comparison on KuaiRec and WeChat21. DADF achieves the best MAE and XAUC on all seven backbones across both datasets, demonstrating that it functions as a general plug-in correction module regardless of the underlying watch-time modeling paradigm—covering direct regression (VR, WLR), duration-aware debiasing (D2Q, D^2CO), discretized modeling (CREAD), and distributional modeling (EGMN, TPM). On average, DADF reduces MAE by 4.33% and improves XAUC by 4.01% over all backbone–dataset combinations. The gains are especially pronounced for VR, TPM, and D^2CO, whose base predictions exhibit larger residual bias, suggesting that these models benefit most from distribution-aware second-stage correction. For already competitive backbones such as D2Q and EGMN, the improvements are smaller yet consistently positive, indicating that DADF can further refine well-calibrated first-stage predictors without degrading them. Compared with TranSUN, which applies a single global multiplicative correction, DADF delivers more consistent gains across backbones and datasets, confirming that explicitly modeling duration-dependent residual heterogeneity and auxiliary prediction signals from engagement heads is more effective than a backbone-agnostic calibration strategy.

Table 1: Overall performance on KuaiRec and WeChat21 by MAE↓ and XAUC↑. Best results within each backbone group are in bold; second-best are underlined. DADF significantly outperforms the strongest baseline within each group (paired t-test, $p < 0.05$).

Backbone	Method	KuaiRec		WeChat21	
		MAE↓	XAUC↑	MAE↓	XAUC↑
VR	Base	4.584	0.5578	18.681	0.6766
	w/ TranSUN	4.478	0.5693	18.571	0.6787
	w/ DADF	**4.235**	**0.6125**	**17.912**	**0.6902**
WLR	Base	4.414	0.5941	18.215	0.6861
	w/ TranSUN	4.364	0.5965	18.133	0.6876
	w/ DADF	**4.172**	**0.6227**	**17.838**	**0.6934**
TPM	Base	4.459	0.5495	19.545	0.6570
	w/ TranSUN	4.361	0.5971	18.529	0.6814
	w/ DADF	**4.166**	**0.6233**	**18.109**	**0.6898**
D2Q	Base	4.123	0.6319	17.544	0.6935
	w/ TranSUN	4.323	0.6082	17.855	0.6925
	w/ DADF	**4.106**	**0.6345**	**17.534**	**0.6946**
CREAD	Base	4.346	0.5927	19.128	0.6679
	w/ TranSUN	4.395	0.5958	18.515	0.6824
	w/ DADF	**4.189**	**0.6211**	**18.164**	**0.6903**
D^2CO	Base	4.613	0.5687	18.558	0.6861
	w/ TranSUN	4.300	0.6097	18.080	0.6868
	w/ DADF	**4.168**	**0.6233**	**17.683**	**0.6952**
EGMN	Base	4.081	0.6245	18.330	0.6896
	w/ TranSUN	4.255	0.6120	18.099	0.6892
	w/ DADF	**4.002**	**0.6257**	**17.955**	**0.6911**

5.2.2 Online A/B Test (RQ2). Before the online A/B test, DADF also shows consistent gains on streaming-updated production logs,

Table 2: Online A/B test lifts on Kwai. Relative improvement over the WLR-based production baseline. ($p < 0.05$ for all metrics.)

Metric	Relative improvement
Total time spent (App)	+0.356%
Avg. time spent per device	+0.347%
Avg. video plays per device	+0.120%
7-day retention (LT7)	+0.054%
Completion rate	+0.215%
Effective-view rate	+0.312%
Long-view rate	+0.351%
Short-view rate	-0.273%

improving WUAUC by 1.88 percentage points and reducing MAE by 12.57% over the WLR-based production baseline. We then deploy DADF as a lightweight debiasing plug-in on top of the production watch-time predictor in Kwai, Kuaishou's international short-video app. The control group uses the WLR-based production baseline, while the treatment group applies DADF as a post-hoc correction before the final ranking formula. We run the experiment for 7 days on real production traffic, following controlled experiment [17] and overlapping experiment infrastructure [36]. As shown in Table 2, DADF consistently improves engagement across all dimensions: total time spent on app and average time spent per device increase by 0.356% and 0.347%, respectively, and 7-day retention improves by 0.054%. On watch-quality metrics, DADF increases completion, effective-view, and long-view rates while reducing the short-view rate, confirming that the debiasing correction shifts the distribution toward higher-quality views and away from quick exits.

5.3 Ablation Study (RQ3)

Although the main results demonstrate that DADF consistently outperforms existing methods, it is still necessary to investigate how each component contributes to the overall improvement. To this end, we construct three ablated variants by removing one module at a time: (1) **w/o Dist.** removes the dynamic distribution-aware module, forcing the model to optimize residuals directly in the original long-tailed space; (2) **w/o Factor** removes the debias-factor-aware module, eliminating duration-conditioned residual heterogeneity modeling; (3) **w/o Aux.** removes the multi-label-aware module, discarding auxiliary prediction signals from engagement heads, such as short-view, completion, and long-view indicators.

Table 3 reports the results. All three variants consistently underperform the full DADF on both MAE and XAUC across both datasets, confirming that each module provides complementary benefits. Removing the distribution-aware module leads to the largest MAE degradation, suggesting that stable transformed-space learning is critical for handling the long-tailed label distribution. Removing the auxiliary labels causes the most notable XAUC drop, indicating that engagement signals are particularly informative for ranking quality. The debias-factor-aware module provides consistent but moderate gains on both metrics, reflecting the value of conditioning residual correction on duration-related factors.

Table 3: Ablation study results of DADF on WLR [8].

Variant	KuaiRec		WeChat21	
	MAE↓	XAUC↑	MAE↓	XAUC↑
Full DADF	4.1723	0.6227	17.8376	0.6934
w/o Dist.	4.1901	0.6210	17.8748	0.6930
w/o Factor	4.1823	0.6212	17.8454	0.6931
w/o Aux.	4.1865	0.6204	17.9137	0.6920

5.4 Bias-oriented Analysis (RQ4)

Beyond aggregate metrics, we further examine whether DADF reduces prediction errors in duration-related residual regimes. This analysis is motivated by the observation that watch-time bias is not uniformly distributed: samples with different video durations often exhibit different residual patterns, and long-duration videos are usually sparser and harder to fit. Therefore, we analyze DADF from two complementary perspectives: fine-grained duration buckets and long-duration tail slices of difficult examples.

5.4.1 Duration-wise Error Reduction. Figure 3 compares the relative MAE reduction of DADF under two bucketization views: *duration-wise* buckets defined by video duration, and *watch-time-wise* buckets defined by observed watch time. Unlike the watch-time buckets in Figure 1, video duration is available at inference time and serves as the debias factor in DADF. Therefore, the duration-wise view directly evaluates whether duration-aware correction is effective under different debias-factor regimes.

DADF achieves positive MAE reduction in all duration-wise buckets, indicating that the improvement is not merely driven by aggregate averaging. The gain is relatively small in short-duration buckets, such as 5.5% in $[0, 20)$ and 7.0% in $[20, 40)$, but becomes much larger as video duration increases, exceeding 15% in most medium- and long-duration buckets, peaking at 23.8% in $[180, 200)$, and remaining 18.2% in $[200, +)$. In contrast, under the watch-time-wise view, the largest gain appears in the short-watch region, especially 20.7% in $[0, 20)$, which is consistent with the strong overestimation of short plays shown in Figure 1. These two views are complementary rather than contradictory: the watch-time-wise view shows *where* correction happens in terms of user consumption, while the duration-wise view shows *under which debias-factor regimes* it is most beneficial. In particular, long-duration videos contain a large number of short-play samples, where the base model often exhibits substantial bias. As a result, improvements on short-watch behaviors are amplified inside long-duration buckets, leading to much larger duration-wise MAE reduction in long-video regions.

Overall, these results provide direct evidence for the effectiveness of the Debias-Factor-aware Module. By conditioning correction on duration groups, DADF adapts to heterogeneous residual regimes more effectively than a single global correction pattern.

5.4.2 Tail-slice Analysis. To further verify whether the above duration-wise improvements translate into stronger performance on difficult long-duration regions, we evaluate DADF on top-duration tail slices, including Tail-20% and Tail-10%. As shown in Figure 4,

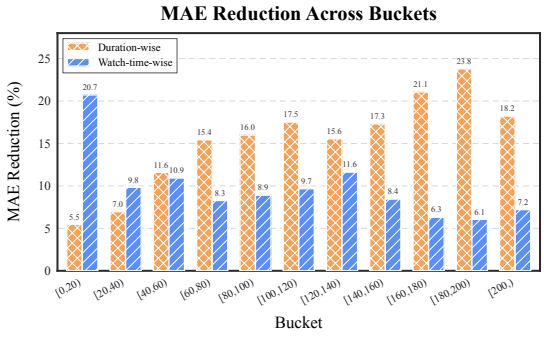


Figure 3: MAE reduction across duration/watch-time buckets.

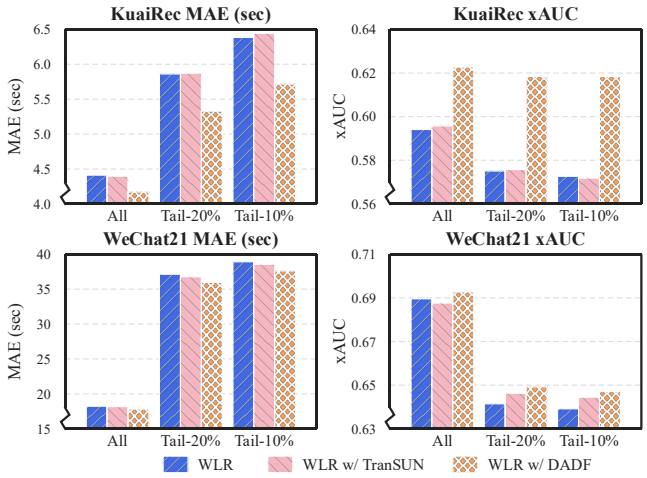


Figure 4: Tail-slice analysis on long-duration videos. DADF brings larger gains in high-duration regions where samples are sparse and duration-related bias is more pronounced.

the relative improvements become larger when the evaluation focuses on longer-duration samples. On KuaiRec, the MAE reduction increases from 5.48% on all samples to 9.10% on Tail-20% and 10.43% on Tail-10%, while the XAUC lift increases from 4.82% to 7.52% and 7.98%. On WeChat21, the corresponding MAE reduction increases from 2.26% to 3.21% and 3.35%, while the XAUC lift increases from 0.45% to 1.21% and 1.24%, respectively.

Together with the duration-wise analysis in Figure 3, these results show that DADF is especially beneficial in regions where duration-related residual bias is more pronounced. The fine-grained bucket analysis demonstrates consistent error reduction across duration regimes, while the tail-slice analysis further confirms that the gains are amplified in sparse long-duration regions. This provides direct evidence for RQ4: DADF not only improves aggregate MAE and XAUC, but also better corrects local bias in difficult duration-related regions and long-duration tail slices.

5.4.3 Bucket Sensitivity. We vary the number of duration buckets $K = |\mathcal{G}|$ to examine the sensitivity of DADF to the granularity of the bucketing function $\pi(\cdot)$. Offline experiments use equal-frequency duration buckets to balance sample sizes across groups, while the online system uses four fixed-duration buckets based on historical traffic statistics and production experience. As shown in Figure 5, DADF remains stable across a moderate range of K, suggesting that the framework is not sensitive to the exact bucket configuration. This stability is desirable in production because bucket boundaries may need to be updated with traffic distribution changes, and the correction module should not depend on a fragile hand-tuned partition. It also indicates that the gains mainly come from duration-conditioned residual modeling rather than from an accidental choice of bucket boundaries. Since overly fine-grained buckets may reduce the number of samples per group and make group-specific correction less stable, we choose the final configuration by balancing validation performance, bucket-level sample size, and deployment complexity in the online system.

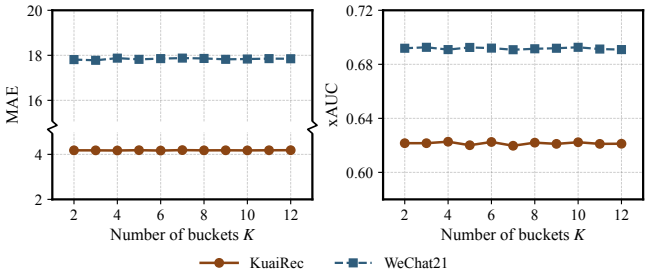


Figure 5: Sensitivity to the number of duration buckets K. DADF remains stable across moderate bucket configurations.

6 Conclusion

In this study, we address the critical challenge of local calibration bias in industrial short-video watch-time prediction by proposing DADF, a distribution-aware debiasing framework. By systematically analyzing residual patterns from real-world production data, we identify pseudo-balance—where systematic overestimation of short views and underestimation of long views cancel out in aggregate—as a key obstacle. To tackle this challenge without replacing the deployed predictor, we perform lightweight second-stage multiplicative correction through an adaptive design, where a dynamic distribution-aware module stabilizes long-tailed correction factors via group-specific transformations, a debias-factor-aware module partitions heterogeneous residuals by video duration into dedicated expert branches, and a multi-label-aware module incorporates auxiliary engagement signals to enrich correction estimation. Extensive offline evaluations on KuaiRec and WeChat21, together with online A/B tests on our large-scale industrial platform, demonstrate that DADF significantly outperforms state-of-the-art baselines, yielding consistent MAE and XAUC improvements across diverse backbones, a 1.88 pp WUAUC gain and 12.57% MAE reduction offline, and a statistically significant 0.347% lift in average time spent per device online. Notably, the success of DADF highlights the importance of explicitly modeling heterogeneous residual distributions through inference-time observable factors to improve long-tailed regression tasks. As a model-agnostic second-stage correction paradigm, DADF may also inform future work on other continuous engagement prediction scenarios beyond watch time.

References

[1] Narayanaswamy Balakrishnan and Asit P. Basu. 1996. *Exponential Distribution: Theory, Methods and Applications.*

[2] Maurice S. Bartlett. 1936. The Square Root Transformation in Analysis of Variance. *Supplement to the Journal of the Royal Statistical Society* 3, 1 (1936), 68–78.

[3] Maurice S. Bartlett. 1947. The Use of Transformations. *Biometrics* 3, 1 (1947), 39–52.

[4] Christopher M. Bishop. 1994. *Mixture Density Networks.* Technical Report. Aston University.

[5] Stephen Bonner and Flavian Vasile. 2018. Causal Embeddings for Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems.* 104–112.

[6] George E. P. Box and David R. Cox. 1964. An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)* 26, 2 (1964), 211–252.

[7] Qingpeng Cai, Shuchang Liu, Xueliang Wang, Tianyou Zuo, Wentao Xie, Bin Yang, Dong Zheng, Peng Jiang, and Kun Gai. 2023. Reinforcing User Retention in a Billion Scale Short Video Recommender System. In *Companion Proceedings of the ACM Web Conference 2023.* 421–426.

[8] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems.* 191–198.

[9] Naihua Duan. 1983. Smearing Estimate: A Nonparametric Retransformation Method. *J. Amer. Statist. Assoc.* 78, 383 (1983), 605–610.

[10] Silvia Ferrari and Francisco Cribari-Neto. 2004. Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics* 31, 7 (2004), 799–815.

[11] D. J. Finney. 1941. On the Distribution of a Variate Whose Logarithm Is Normally Distributed. *Supplement to the Journal of the Royal Statistical Society* 7, 2 (1941), 155–161.

[12] Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. 2022. KuaiRec: A Fully-observed Dataset and Insights for Evaluating Recommender Systems. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management.* 540–550.

[13] Chongming Gao, Shijun Li, Yuan Zhang, Jiawei Chen, Biao Li, Wenqiang Lei, Peng Jiang, and Xiangnan He. 2022. KuaiRand: An Unbiased Sequential Recommendation Dataset with Randomly Exposed Videos. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management.* 3953–3957.

[14] Xudong Gong, Qinlin Feng, Yuan Zhang, Jiangling Qin, Weijie Ding, Biao Li, Peng Jiang, and Kun Gai. 2022. Real-Time Short Video Recommendation on Mobile Devices. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management.* 3103–3112.

[15] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7132–7141.

[16] Roy W. Koch and Gary M. Smillie. 1986. Bias in Hydrologic Prediction Using Log-Transformed Regression Models. *Journal of the American Water Resources Association* 22, 5 (1986), 717–723.

[17] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online Controlled Experiments at Large Scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 1168–1176.

[18] Chengzhi Lin, Shuchang Liu, Chuyuan Wang, and Yongqi Liu. 2024. Conditional Quantile Estimation for Uncertain Watch Time in Short-Video Recommendation. arXiv:2407.12223.

[19] Xiao Lin, Xiaokai Chen, Linfeng Song, Jingwei Liu, Biao Li, and Peng Jiang. 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.

[20] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 1930–1939.

[21] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. 2018. Entire Space Multi-Task Model: An Effective Approach for Estimating Post-Click Conversion Rate. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1137–1140.

[22] Geoffrey J. McLachlan, Sharon X. Lee, and Suren I. Rathnayake. 2019. Finite Mixture Models. *Annual Review of Statistics and Its Application* 6 (2019), 355–378.

[23] Mean Absolute Error. 2016. Mean Absolute Error. Retrieved September 19, 2016, 14.

[24] Sushant More. 2022. Identifying and Overcoming Transformation Bias in Forecasting Models. arXiv:2208.12264.

[25] Michael C. Newman. 1993. Regression Analysis of Log-Transformed Data: Statistical Bias and Its Correction. *Environmental Toxicology and Chemistry* 12, 6 (1993), 1129–1133.

[26] Jerzy Neyman and Elizabeth L. Scott. 1960. Correction for Bias Introduced by a Transformation of Variables. *The Annals of Mathematical Statistics* 31, 3 (1960), 643–655.

[27] Vincent Moshi Ouma, Samuel Musili Mwalili, and Anthony Wanjoya Kiberia. 2016. Poisson Inverse Gaussian Regression Model for Infectious Disease Count Data. *American Journal of Theoretical and Applied Statistics* 5, 5 (2016), 326–333.

[28] Yunzhu Pan, Chen Gao, Jianxin Chang, Yanan Niu, Yang Song, Kun Gai, Depeng Jin, and Yong Li. 2023. Understanding and Modeling Passive-Negative Feedback for Short-Video Sequential Recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems.* 540–550.

[29] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. 2021. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research* 22, 57 (2021), 1–64.

[30] Yuta Saito, Suguru Yaginuma, Yuta Nishino, Kazuhide Nakata, and Keiichi Sakata. 2020. Unbiased Recommender Learning from Missing-Not-at-Random Implicit Feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining.* 501–509.

[31] Remi M. Sakia. 1992. The Box-Cox Transformation Technique: A Review. *Journal of the Royal Statistical Society: Series D (The Statistician)* 41, 2 (1992), 169–178.

[32] Hiroshi Shono. 2008. Application of the Tweedie Distribution to Zero-Catch Data in CPUE Analysis. *Fisheries Research* 93, 1–2 (2008), 154–162.

[33] Craig A. Stow, Kenneth H. Reckhow, and Song S. Qian. 2006. A Bayesian Approach to Retransformation Bias in Transformed Regression. *Ecology* 87, 6 (2006), 1472–1477.

[34] Bogdan M. Strimbu, Alexandru Amarioarei, John Paul McTague, and Mihaela M. Paun. 2018. A Posteriori Bias Correction of Three Models Used for Environmental Reporting. *Forestry: An International Journal of Forest Research* 91, 1 (2018), 49–62.

[35] Jie Sun, Zhaoying Ding, Xiaoshuang Chen, Qi Chen, Yincheng Wang, Kaiqiao Zhan, and Ben Wang. 2024. CREAD: A Classification-Restoration Framework with Error Adaptive Discretization for Watch Time Prediction in Video Recommender Systems. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 8 (2024), 9027–9034.

[36] Diane Tang, Ashish Agarwal, Deirdre O'Brien, and Mike Meyer. 2010. Overlapping Experiment Infrastructure: More, Better, Faster Experimentation. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 17–26.

[37] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems.* 269–278.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30. 5998–6008.

[39] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & Cross Network for Ad Click Predictions. In *Proceedings of the ADKDD'17.* 1–7.

[40] Tianxin Wang, Jingwu Chen, Fuzhen Zhuang, Leyu Lin, Feng Xia, Lihuan Du, and Qing He. 2020. Capturing Attraction Distribution: Sequential Attentive Network for Dwell Time Prediction. In *ECAI 2020.* 529–536.

[41] Wenjie Wang, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2021. Deconfounded Recommendation for Alleviating Bias Amplification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 1717–1725.

[42] Yunpeng Weng, Xing Tang, Zhenhao Xu, Fuyuan Lyu, Dugang Liu, Zexu Sun, and Xiuqiang He. 2024. OptDist: Learning Optimal Distribution for Customer Lifetime Value Prediction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management.* 2523–2533.

[43] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. 2018. Beyond Views: Measuring and Predicting Engagement in Online Videos. *Proceedings of the International AAAI Conference on Web and Social Media* 12, 1 (2018), 434–442.

[44] Dongbo Xi, Zhen Chen, Peng Yan, Yao Zhang, Yongchun Zhu, Fuzhen Zhuang, and Yu Chen. 2021. Modeling the Sequential Dependence among Audience Multi-Step Conversions with Multi-Task Learning in Targeted Display Advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* 3745–3755.

[45] Xing Yi, Liangjie Hong, Erheng Zhong, Nanthan Nan Liu, and Suju Rajan. 2014. Beyond Clicks: Dwell Time for Personalization. In *Proceedings of the 8th ACM Conference on Recommender Systems.* 113–120.

[46] Jiahao Yu, Haozhuang Liu, Yeqiu Yang, Lu Chen, Jian Wu, Yuning Jiang, and Bo Zheng. 2025. TranSUN: A Preemptive Paradigm to Eradicate Retransformation Bias Intrinsically from Regression Models in Recommender Systems. arXiv:2505.13881. NeurIPS 2025 poster.

[47] Ruohan Zhan, Changhua Pei, Qiang Su, Jianfeng Wen, Xueliang Wang, Guanyu Mu, Dong Zheng, Peng Jiang, and Kun Gai. 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.

[48] Haiyuan Zhao, Guohao Cai, Jieming Zhu, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. 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.
[49] Haiyuan Zhao, Lei Zhang, Jun Xu, Guohao Cai, Zhenhua Dong, and Ji-Rong Wen. 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.
[50] Xu Zhao, RuiBo Ma, Jiaqi Chen, Weiqi Zhao, Ping Yang, and Yao Hu. 2025. Multi-Granularity Distribution Modeling for Video Watch Time Prediction via Exponential-Gaussian Mixture Network. In *Proceedings of the 19th ACM Conference on Recommender Systems*. 309–318.
[51] 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*. 334–345.
[52] 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*. 1059–1068.
[53] Xinhua Zhuang, Yan Huang, Kannappan Palaniappan, and Yunxin Zhao. 1996. Gaussian Mixture Density Modeling, Decomposition, and Applications. *IEEE Transactions on Image Processing* 5, 9 (1996), 1293–1302.

A Proofs

A.1 Long-Tailedness Inheritance of the Multiplicative Correction Factor

We provide a simple theoretical argument showing that the ratio-style multiplicative correction factor can inherit the long-tailed property of the original watch-time label. For a given feature value x, let $\bar{F}_Y(t \mid x) := \mathbb{P}(Y > t \mid X = x)$ denote the conditional survival function of the observed watch time Y.

We say that the conditional distribution $F_{Y|X=x}$ belongs to the long-tailed class $\mathcal{L}$, if for any fixed $a \in \mathbb{R}$,

$$\lim_{t\to\infty} \frac{\bar{F}_Y(t + a \mid x)}{\bar{F}_Y(t \mid x)} = 1, \tag{23}$$

assuming the survival function is positive for sufficiently large t.

Let $g : \mathcal{X} \to (0, \infty)$ be a positive measurable function and define the ratio-style correction factor $R := \frac{Y}{g(X)}$. Here $g(X)$ can be viewed as a positive prediction-dependent quantity, such as a stopped-gradient first-stage prediction with numerical stabilization.

We show that, for any fixed x, if $F_{Y|X=x} \in \mathcal{L}$, then $F_{R|X=x} \in \mathcal{L}$. Conditioning on $X = x$, $g(X) = g(x) > 0$ is a positive constant. Therefore,

$$\begin{aligned} \bar{F}_R(t \mid x) &:= \mathbb{P}(R > t \mid X = x) \\ &= \mathbb{P}(Y > tg(x) \mid X = x) \\ &= \bar{F}_Y(tg(x) \mid x). \end{aligned} \tag{24}$$

For any fixed $a \in \mathbb{R}$, we have

$$\frac{\bar{F}_R(t + a \mid x)}{\bar{F}_R(t \mid x)} = \frac{\bar{F}_Y((t + a)g(x) \mid x)}{\bar{F}_Y(tg(x) \mid x)}, \tag{25}$$

where both sides express the ratio of survival probabilities under a shift of size a. Therefore,

$$\frac{\bar{F}_R(t + a \mid x)}{\bar{F}_R(t \mid x)} = \frac{\bar{F}_Y(s + ag(x) \mid x)}{\bar{F}_Y(s \mid x)}, \tag{26}$$

where the equality follows from the substitution $s = tg(x)$. For fixed x, $ag(x)$ is a fixed constant. Since $F_{Y|X=x} \in \mathcal{L}$, it follows that

$$\lim_{s\to\infty} \frac{\bar{F}_Y(s + ag(x) \mid x)}{\bar{F}_Y(s \mid x)} = 1. \tag{27}$$

Consequently,

$$\lim_{t\to\infty} \frac{\bar{F}_R(t + a \mid x)}{\bar{F}_R(t \mid x)} = 1, \tag{28}$$

which implies $F_{R|X=x} \in \mathcal{L}$.

This result shows that the multiplicative correction factor does not automatically remove the tail behavior of the original watch-time label. Instead, under the same conditioning, it can preserve the long-tailedness of Y. This provides a theoretical rationale for applying a distributional transformation to the correction factor before regression.

A.2 Oracle Risk Analysis of Group-Conditioned Modeling

We provide a population-level oracle-risk analysis to justify the use of group-conditioned modeling under the squared loss. The result shows that, when the correction target has a finite second moment, conditioning on a debiasing-related group variable does not increase the oracle optimal prediction risk.

Let U denote the correction target learned by DADF. It can be the raw multiplicative correction factor:

$$F = \frac{y}{|\operatorname{sg}(\hat{y}_0(x))| + \epsilon}, \tag{29}$$

where y is the observed watch time, $\hat{y}_0(x)$ is the first-stage prediction, $\operatorname{sg}(\cdot)$ denotes the stop-gradient operation, and $\epsilon > 0$ is used for numerical stability.

Alternatively, U can be the group-specific transformed correction target:

$$U = T_{\lambda_G}(F), \tag{30}$$

where G denotes a discrete debiasing-related group variable, such as the duration group.

Assume that $G \in \{1, 2, \ldots, K\}$, $\mathbb{E}[U^2] < \infty$, so that all squared risks and variance terms are well-defined. We consider a prediction function that only depends on the group variable G, denoted by $g_\theta(G)$. Its population risk under the squared loss is defined as:

$$\mathcal{R}_G(\theta) = \mathbb{E}\left[(U - g_\theta(G))^2\right], \tag{31}$$

where the expectation is taken over the joint distribution of (U, G).

The global model without group information can be regarded as a special case of the group-conditioned model, since it corresponds to assigning the same prediction value to all groups. Its prediction function is:

$$g^{(0)}_{\theta_0}(G) = \theta_0. \tag{32}$$

The corresponding risk is:

$$\mathcal{R}_0(\theta_0) = \mathbb{E}\left[(U - \theta_0)^2\right], \tag{33}$$

where the expectation is taken over the marginal distribution of U.

Proposition. *Under the squared loss, if the group-conditioned model is allowed to learn an independent prediction value for each group with positive probability, then its optimal population risk is no larger than that of the global model without group information:*

$$\mathcal{R}^\star_G \le \mathcal{R}^\star_0, \tag{34}$$

where $\mathcal{R}^\star_0$ and $\mathcal{R}^\star_G$ are defined as:

$$\mathcal{R}^\star_0 = \min_{\theta_0} \mathbb{E}\left[(U - \theta_0)^2\right], \tag{35}$$

and

$$\mathcal{R}^\star_G = \min_{\theta} \mathbb{E}\left[(U - g_\theta(G))^2\right]. \tag{36}$$

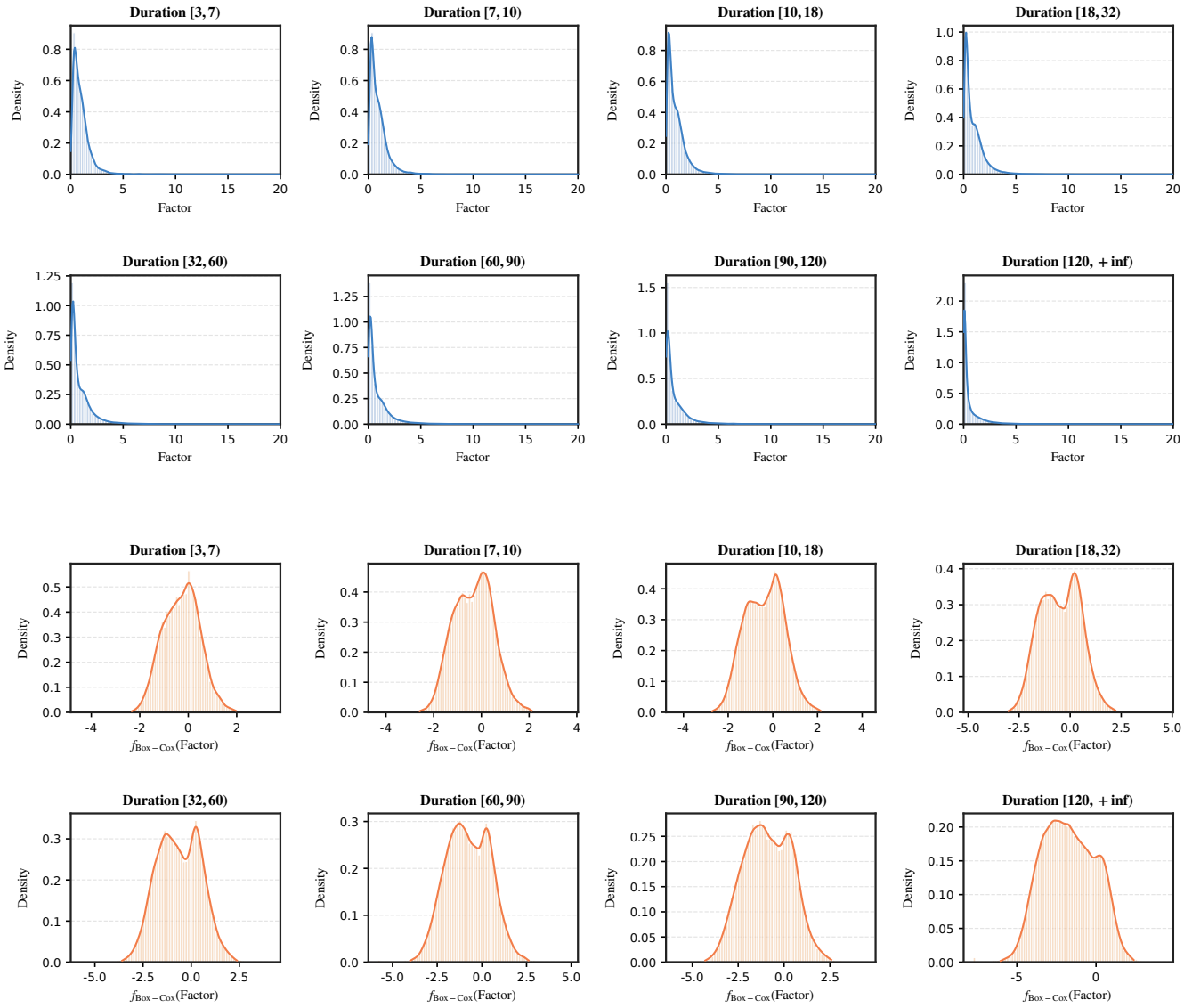


Figure 6: Distribution comparison of the raw multiplicative correction factor (top) and the group-specific Box–Cox transformed correction target (bottom) on Kwai. The transformed target is substantially more compact and less skewed across duration buckets

Proof. We first consider the global model without group information. Let $\mu = \mathbb{E}[U]$. For any global constant θ_0, we have $U - \theta_0 = (U - \mu) + (\mu - \theta_0)$. Therefore,

$$\begin{aligned}\mathcal{R}_0(\theta_0) &= \mathbb{E}\left[(U-\theta_0)^2\right] \\ &= \mathbb{E}\left[(U-\mu)^2\right] + (\mu-\theta_0)^2 + 2(\mu-\theta_0)\mathbb{E}[U-\mu].\end{aligned} \tag{37}$$

Since $\mathbb{E}[U - \mu] = 0$, we obtain:

$$\mathcal{R}_0(\theta_0) = \mathrm{Var}(U) + (\mu - \theta_0)^2. \tag{38}$$

The second term is non-negative, and the minimum is attained if and only if:

$$\theta_0^\star = \mu = \mathbb{E}[U]. \tag{39}$$

Thus, the oracle optimal risk of the global model is:

$$\mathcal{R}_0^\star = \mathrm{Var}(U). \tag{40}$$

We then consider the group-conditioned model. If the model can learn an independent prediction value for each group, its prediction function can be written as:

$$g_\theta(G) = \sum_{g=1}^{K} \theta_g \mathbf{1}(G = g), \tag{41}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function.

By the law of total expectation, the population risk can be decomposed as:

$$\mathcal{R}_G(\theta) = \sum_{g=1}^{K} \mathbb{P}(G=g)\mathbb{E}\left[(U-\theta_g)^2 \mid G=g\right], \tag{42}$$

where $\mathbb{P}(G=g)$ denotes the probability of group g. Groups with zero probability do not affect the risk and can be ignored.

For a fixed group g with $\mathbb{P}(G=g) > 0$, let $\mu_g = \mathbb{E}[U \mid G=g]$. The within-group risk satisfies:

$$\begin{aligned}\mathbb{E}\left[(U-\theta_g)^2 \mid G=g\right] &= \mathbb{E}\left[((U-\mu_g)+(\mu_g-\theta_g))^2 \mid G=g\right] \\ &= \mathbb{E}\left[(U-\mu_g)^2 \mid G=g\right] + (\mu_g-\theta_g)^2 \\ &\quad + 2(\mu_g-\theta_g)\mathbb{E}\left[U-\mu_g \mid G=g\right].\end{aligned} \tag{43}$$

Since $\mathbb{E}\left[U-\mu_g \mid G=g\right] = 0$, we obtain:

$$\mathbb{E}\left[(U-\theta_g)^2 \mid G=g\right] = \mathrm{Var}(U \mid G=g) + (\mu_g-\theta_g)^2. \tag{44}$$

The second term is non-negative, and the minimum is attained if and only if:

$$\theta_g^\star = \mu_g = \mathbb{E}[U \mid G=g]. \tag{45}$$

Therefore, the oracle optimal group-conditioned predictor is:

$$g_{\theta^\star}(G) = \mathbb{E}[U \mid G]. \tag{46}$$

The corresponding oracle optimal risk is:

$$\mathcal{R}_G^\star = \sum_{g=1}^{K} \mathbb{P}(G=g)\,\mathrm{Var}(U \mid G=g) = \mathbb{E}\left[\mathrm{Var}(U \mid G)\right]. \tag{47}$$

By the law of total variance:

$$\mathrm{Var}(U) = \mathbb{E}\left[\mathrm{Var}(U \mid G)\right] + \mathrm{Var}\left(\mathbb{E}[U \mid G]\right), \tag{48}$$

where the second term on the right-hand side is non-negative. Combining this identity with $\mathcal{R}_0^\star = \mathrm{Var}(U)$ and $\mathcal{R}_G^\star = \mathbb{E}\left[\mathrm{Var}(U \mid G)\right]$, we have:

$$\mathcal{R}_0^\star - \mathcal{R}_G^\star = \mathrm{Var}\left(\mathbb{E}[U \mid G]\right) \geq 0. \tag{49}$$

Thus:

$$\mathcal{R}_G^\star \leq \mathcal{R}_0^\star. \tag{50}$$

This completes the proof.

The above result shows that, under the squared loss, conditioning on duration groups can remove the additional mixture variance induced by differences in group-wise conditional means. In other words, if duration groups capture systematic distributional differences in the correction target U, then group-aware correction is better than global correction in terms of oracle optimal risk. This provides a theoretical motivation for introducing duration-group-aware expert branches in DADF.

B Additional Experimental Results

B.1 Duration-aware Distribution Stabilization via Group-specific Box–Cox Transformation

Figure 6 compares the raw multiplicative correction factor and its group-specific Box–Cox transformed counterpart across different duration buckets. The raw correction factor exhibits clear right-skewed and long-tailed patterns in all duration groups, with most samples concentrated around small values and a small fraction extending to large values. This suggests that directly regressing the correction factor in the original space may suffer from distributional skewness and variance amplification. After applying the bucket-wise Box–Cox transformation, the transformed correction targets become substantially more compact and less skewed within each duration bucket, providing a more stable supervision signal for the second-stage correction model.

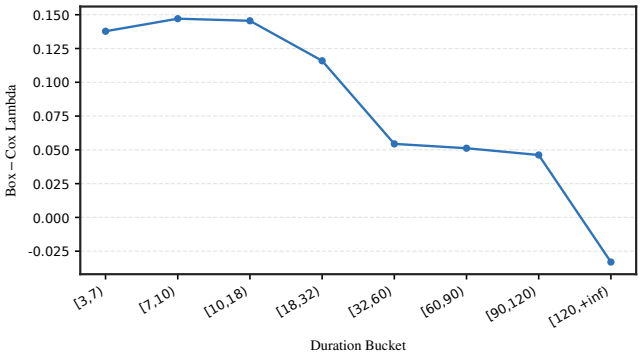


Figure 7: Learned group-specific Box–Cox transformation parameters λ_g across duration buckets on Kwai.

Figure 7 further reports the group-specific Box–Cox transformation parameter λ_g across duration buckets. The parameter is not globally constant, but changes noticeably with video duration and shows an overall decreasing trend as duration increases. In particular, the long-duration buckets require smaller transformation parameters, with the last bucket even falling below zero. This indicates that the correction-factor distribution varies across duration regimes and that a single global transformation may mix heterogeneous residual patterns. These results support the use of a duration-aware group-specific transformation in DADF.