Journal of Innovative Technologies
https://journalinnotech.com/
Vol 2(1) 2026

An Empirical Comparison of Popularity-Debiasing Strategies in News and Short-Video Recommendation

Chuankai Luo[1] , Yixin Zheng[2] , Jianuo Wang[3,*]

[1] Department of Electronic Engineering, Tsinghua University, Beijing, China
[2] Information Science, University of Michigan, Ann Arbor, MI, USA
[3] Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI, USA

* Corresponding Email: wangjanuo07@gmail.com

Abstract

Popularity bias is a persistent obstacle in content recommendation: a small set of head items absorbs most exposure, long-tail items are systematically under-recommended, and offline metrics drift away from genuine user preference. A growing body of debiasing strategies has been proposed, yet most are evaluated on a single domain with a single backbone, leaving practitioners without clear guidance on which strategy to adopt under which conditions. This study presents a controlled empirical comparison of five representative popularity-debiasing strategies—inverse propensity scoring, popularity deconfounding and adjusting, model-agnostic counterfactual reasoning, interest–conformity disentangling, and automatic debiasing—applied uniformly across news and short-video recommendation. Each strategy is paired with strong domain backbones and evaluated on MIND, KuaiRand, and KuaiRec, using both accuracy metrics and bias-oriented metrics that exploit randomly exposed interactions for near-unbiased assessment. The results indicate that no single strategy dominates both domains: causal-intervention adjustment yields the most favorable accuracy–diversity balance on news, while disentangling-based representation learning is more effective on short-video data with random-exposure feedback. Across settings the accuracy cost of debiasing is moderate, while tail coverage improves substantially. We characterize the sensitivity of debiasing strength and the consistency of conclusions across datasets, and distill practical selection guidance for deployment.

***Keywords*:** popularity bias, news recommendation, short-video recommendation, debiasing evaluation.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

1.1. Background and Motivation

Content platforms rank candidate items by predicted engagement, and the feedback collected from these rankings trains the next iteration of the recommender. When historically popular items receive disproportionate exposure, the loop amplifies their advantage and suppresses items in the long tail, producing skewed catalogs and a widening gap between offline scores and true user satisfaction. Popularity bias is one of several intertwined biases that arise from non-random exposure and self-selected feedback, and a comprehensive survey organizes these biases together with the methods that target them [1]. Treating recommendation as an intervention rather than passive observation, propensity-based correction reweights observed interactions by their estimated exposure probability, providing an unbiased risk estimator when the propensities are correct [2]. Building on this causal view, popularity deconfounding removes the confounding effect of item popularity during training while deliberately reintroducing a controlled popularity signal at inference [3].

The practical difficulty is that these strategies are seldom compared on equal footing. Each is typically introduced alongside its own backbone, dataset, and tuning regime, which entangles the contribution of the debiasing mechanism with the choice of base recommender. A platform team deciding how to mitigate popularity bias in a news feed or a short-video stream has little evidence about which strategy will transfer to its own setting, or how much accuracy it must trade for a fairer exposure distribution.

News and short-video feeds make this gap especially costly. News catalogs turn over quickly, so a recommender that over-concentrates on a handful of trending stories leaves freshly published articles starved of the exposure they need to accumulate signal. Short-video streams instead present an effectively unbounded catalog in which a viral clip can dominate impressions for millions of viewers, crowding out creators whose content has not yet been amplified. The two settings share the same underlying pathology while differing in the density and the type of feedback available, which makes them a natural pair for studying whether a single debiasing remedy generalizes.

1.2. Scope and Contributions

This work is an application-effect comparison rather than an architecture proposal. We hold the evaluation protocol fixed and vary only the debiasing strategy attached to each backbone, isolating the contribution of debiasing from confounds of model capacity and tuning. The study spans two content domains that differ sharply in signal type—editorial text articles versus short videos with watch-time feedback—allowing us to test whether debiasing conclusions transfer between them.

A. Research Questions

Three questions organize the analysis. The first asks how each debiasing strategy affects ranking accuracy relative to an untreated backbone. The second asks how the same strategy

reshapes exposure across the popularity spectrum, measured by tail coverage and concentration. The third asks whether the relative ordering of strategies is stable when the domain shifts from news to short video, and when the dataset within a domain changes.

B. Comparison Protocol Overview

Each strategy is implemented as a training- or inference-time adjustment layered on a shared backbone, with identical data splits, candidate sets, and tuning budgets. We report accuracy and bias metrics jointly, and for short-video data we exploit a randomly exposed test partition to approximate unbiased evaluation, an option that conventional click logs do not provide. The contribution of this study is comparative rather than methodological: we do not introduce a new recommender, and every component under test is an existing, published technique. What the work adds is a like-for-like measurement of how these techniques behave when the only thing that changes is the debiasing mechanism, across two domains that stress different parts of the popularity problem.

2. Related Work

2.1. Debiasing Strategy Taxonomy

Popularity-debiasing methods can be grouped by the mechanism through which they neutralize the influence of item popularity, and the comparison in this study draws one representative from each major mechanism.

A. Propensity-based and Causal-Intervention Methods

Counterfactual reasoning offers a model-agnostic route: by estimating the score an item would receive from its popularity alone and subtracting this direct effect, the recommender retains preference-driven signal while discarding popularity-driven inflation [4]. Automatic debiasing generalizes the reweighting idea, deriving a unified objective that subsumes many heuristic corrections and learning the debiasing parameters from a small unbiased holdout through a meta-learning loop [5]. These methods share a causal framing in which popularity is a confounder or a separable effect to be removed, and they attach naturally to heterogeneous backbones without changing their internal structure. Their appeal for a comparison study is that the correction lives outside the base recommender, so the same mechanism can be transplanted from a text encoder to a sequence encoder with minimal adaptation, keeping the experimental contrast clean.

B. Disentangle- and Counterfactual-based Methods

A complementary line separates the latent factors that drive a click. Interest–conformity disentangling assigns each user and item two embedding components, one capturing genuine interest and one capturing conformity to popularity, and trains them on cause-specific samples so that the interest component remains uncontaminated [6]. Adversarial formulations pursue a related goal for position-induced exposure, learning representations that are invariant to where an item

appears in a ranked list while preserving relevance [7]. The disentangling view is attractive when rich interaction sequences make the conformity signal learnable, since the model needs enough repeated behavior to tell habitual following of the crowd apart from true taste. This data dependence is exactly the property our cross-domain comparison is designed to probe, because news sessions and short-video sessions differ markedly in how much such signal they expose.

2.2. Backbone Recommenders

News recommendation backbones center on text understanding. Multi-head self-attention encoders model word- and article-level interactions to form user and candidate representations [8]. Multi-view learning fuses titles, bodies, and categories into a single news vector [9]. Long- and short-term modeling combines a recurrent encoder for session interest with an identity embedding for stable preference [10]. Personalized attention conditions word and article weighting on the user identity [11]. Short-video and sequential recommendation rely instead on behavior order: self-attentive sequence encoders capture adaptive dependencies among consumed items [12]. Bidirectional masked encoders enrich this with two-sided context [13]. Recurrent session encoders provide a long-standing baseline for ordered consumption [14]. We pair propensity- and disentangle-based debiasing with these backbones while leaving their architecture untouched. The selection spans the dominant design families in each domain, which lets the comparison speak to typical deployments rather than to a single idiosyncratic encoder. By fixing two backbones per domain and rotating the debiasing layer over each, we can also check whether a strategy that helps one backbone helps the other, separating effects that are intrinsic to the debiasing mechanism from effects that merely interact with a particular encoder.

2.3. Cross-Domain Evidence for Robust and Trustworthy Evaluation

The methodological choices in this study are also informed by recent work outside recommendation, where bias-aware evaluation and transparent data use have become central. Training-data bias detection and fairness-aware augmentation show why performance should be examined across data subgroups [15]. Multi-metric trustworthiness evaluation combines confidence calibration with distribution-shift checks instead of relying on a single headline score [16]. Automated data governance and intelligent reporting likewise emphasize traceable transformations [17]. Privacy-preserving click-pattern analysis also shows why behavior modeling should minimize unnecessary exposure of user-level traces [18].

Controlled comparisons are especially valuable when candidate methods share backbones or preprocessing. Comparative evaluations of graph models and gradient-boosting systems separate model choice from the surrounding pipeline [19,20]. Studies of spatiotemporal classifiers and telemetry-driven anomaly detectors apply the same principle under heterogeneous operational conditions [21,22]. Benchmarking learned cardinality estimators and prompt-engineering strategies similarly highlights the need for common data, metrics, and tuning budgets [23,24].

Behavioral streams are particularly sensitive to feature choice. Multidimensional click behavior and browser-specific features provide complementary views of anomalous activity [25,26]. Temporal click patterns add a further diagnostic axis [27]. In operational data, multi-source

monitoring can change warning lead time [28]. Statistical anomaly checks and constraint discovery can help distinguish genuine changes from migration errors [29,30].

Across early-warning systems, robust predictions have been pursued through generative stress scenarios and ensemble anomaly detection [31,32]. Interpretable risk indicators and feature-weight optimization offer complementary routes to transparent decisions [33,34]. Operational demand forecasting further illustrates the value of evaluating predictions against deployment-relevant outcomes [35].

Graph-based reasoning has been used to trace cross-market contagion and identify amplifiers [36,37]. It has also been applied to rare-earth disruption and software supply-chain attack paths [38,39]. Related work models explainable industrial-control attack reasoning and supply-chain credit transmission [40,41]. Multi-objective optimization in carbon-aware scheduling and site selection makes trade-offs explicit [42,43]. Carbon-variability scheduling and graph-reinforcement learning further show how representation and decision layers can be coupled [44,45]. Comparative analysis of coordinated behavior provides a final precedent for evaluating multiple evidence channels under multilingual and temporal heterogeneity [46].

Broader applied context: AI systems for financial monitoring, cross-border risk, public communication, and budget transparency illustrate how exposure allocation and feedback can shape both institutional decisions and user-facing outcomes. [54,55,56,57,58,59,60,61]

Research on fairness, credit and market risk, carbon policy, vulnerable populations, building demand, and project assessment reinforces the need to evaluate predictive performance together with distributional and decision consequences. [62,63,64,65,66,67,68,69]

Work on market microstructure, transaction networks, technology dependencies, corporate distress, and investment patterns provides adjacent examples of learning from heterogeneous behavioral, graph, and textual signals. [70,71,72,73,74,75,76,77]

Adaptive control, cloud anomaly detection, proactive security, high-risk short-video detection, privacy protection, zero-trust analytics, feature selection, and urban optimization all motivate context-aware models that remain robust under changing exposure conditions. [78,79,80,81,82,83,84,85]

Road-scene captioning, image translation, segmentation, style transfer, and retrieval studies provide relevant precedents for representation learning over visual content, a central component of short-video recommendation [86,87,88,89,90,91].

3. Experimental Setup

3.1. Datasets

We evaluate on four public datasets spanning the two domains, summarized in Table 1. For news, MIND provides large-scale impression logs with rich textual content drawn from Microsoft News, and we use the MIND-small partition of fifty thousand randomly sampled users to keep the comparison computationally tractable [47]. As a second news source we use the one-week Adressa

collection, which records reading dwell time and session context for a Norwegian news portal [48]. For short video, KuaiRand supplies an intervention partition in which a fraction of impressions are randomly exposed, a property that enables near-unbiased offline estimation, and we adopt its Pure variant restricted to the standard candidate pool [49]. We complement it with KuaiRec, whose small interaction matrix is almost fully observed at a density close to one, offering a rare chance to estimate true preference without missing-not-at-random gaps [50]. A larger content-driven micro-video collection has recently been released, yet its full version requires author contact and exceeds our compute budget, so it is reserved for future extension [51].

The four datasets were chosen to vary the two axes that matter most for a debiasing study, namely feedback density and the availability of an unbiased ground truth. MIND-small and Adressa anchor the news domain at sparse, logged feedback, with Adressa adding dwell-time context that MIND lacks. KuaiRand-Pure carries the randomly exposed impressions that allow an exposure-corrected test set, and KuaiRec contributes a near-complete observation matrix that removes most missing-not-at-random distortion at evaluation time. Holding the candidate pool, the chronological split, and the negative-sampling procedure constant across datasets within a domain keeps the comparison fair, and the same five debiasing strategies are run on every dataset so that any divergence in their ranking can be traced to data properties rather than to differing experimental pipelines.

Table 1. *Statistics of the processed datasets used in this study. Raw releases are available from the listed sources.*

Dataset (variant)	Domain	#Users	#Items	#Interactions	Exposure	Source
MIND-small	News	50,000	51,282	5,059,623	logged	msnews.github.io
Adressa (1-week)	News	15,514	923	2,717,915	logged	reclab.idi.ntnu.no
KuaiRand-Pure	Short-video	27,285	7,583	1,436,609	logged + random	kuairand.com
KuaiRec (small)	Short-video	1,411	3,327	4,676,570	fully observed	kuairec.com

3.2. Backbones and Debiasing Strategies

A. Backbone Configuration

News experiments use the multi-head self-attention encoder and the multi-view encoder as backbones, both trained with negative sampling over impression logs and word embeddings initialized from pretrained vectors. Short-video experiments use the self-attentive sequence encoder and the bidirectional masked encoder, trained on chronologically ordered interaction sequences with a maximum history length of fifty. All backbones in a domain share embedding

dimensionality, batch size, and optimizer to keep capacity comparable, and only the debiasing layer differs across runs. Table 2 lists the configuration applied to every backbone–strategy pair so that observed differences can be attributed to the debiasing mechanism rather than to tuning. Each backbone is first trained without any correction to establish a within-domain reference point, after which the five debiasing layers are attached one at a time using the same initialization and the same early-stopping criterion on validation loss. This protocol means that the untreated rows in Tables 3 and 4 are not external baselines but the identical encoder stripped of its debiasing layer, which makes the reported deltas a direct measure of what the correction contributes.

Table 2. Shared training configuration for backbone–strategy pairs.

Setting	News backbones	Short-video backbones
Backbones	NRMS, NAML	SASRec, BERT4Rec
Embedding dimension	256	64
Maximum history length	50	50
Optimizer	Adam	Adam
Learning rate	1e-4	1e-3
Batch size	64	256
Debiasing strength grid	{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50}	{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50}
Runs (random seeds)	5	5

B. Debiasing Strategy Implementation

Inverse propensity scoring reweights each observed interaction by the inverse of an item-popularity propensity estimated from global exposure frequency, with the propensities clipped to bound variance. Popularity deconfounding removes the popularity confounder during training and reinjects a tunable popularity term at inference, governed by a single strength parameter. Counterfactual reasoning trains auxiliary item-only and user-only branches and subtracts their direct effect at scoring time. Interest–conformity disentangling augments each backbone with paired interest and conformity embeddings supervised by cause-specific sampling. Automatic debiasing learns its correction weights against a held-out unbiased subset carved from the randomly exposed interactions. Each strategy exposes one principal strength parameter, tuned on validation data over the common grid in Table 2, which keeps the search budget equal across strategies. Where a strategy was originally formulated for collaborative-filtering scores, we attach it to the backbone logits rather than to raw embeddings, so that the correction operates on the same

quantity in every configuration and no strategy gains an advantage from a richer interface to the model.

3.3. Evaluation Metrics and Protocol

A. Accuracy Metrics

We report area under the ROC curve for impression-level click prediction, mean reciprocal rank, and normalized discounted cumulative gain at cut-offs five and ten for news ranking. For short video we report area under the curve for effective-view prediction together with normalized discounted cumulative gain at ten and recall at twenty computed on the randomly exposed test partition, which yields an estimate less contaminated by exposure bias than standard logged feedback. Reporting click- and view-level accuracy alongside ranking quality keeps the news and short-video evaluations comparable while respecting the native label of each domain. The cut-offs were fixed before any debiasing run to avoid selecting depths that flatter a particular strategy, and ranking metrics are computed over the full candidate set within each impression rather than over a sampled subset, which removes a known source of optimistic bias in offline ranking evaluation.

B. Bias and Diversity Metrics

To quantify exposure concentration we report tail coverage, the share of distinct tail items appearing in the top-ranked lists, and the Gini coefficient of the recommended item-frequency distribution, where lower values indicate more equitable exposure. We additionally stratify items into head, torso, and tail segments by training-set popularity and report segment-wise recall, which exposes where each strategy redistributes exposure. The head, torso, and tail boundaries follow the conventional split in which the head holds the most popular items accounting for the bulk of training interactions, leaving the long tail to the remaining low-frequency items. All metrics are averaged over five runs with distinct seeds, and we report the mean across runs. Reading the accuracy metrics together with the concentration metrics is essential, since a strategy can lift tail coverage simply by degrading ranking quality, and only the joint view reveals whether the redistribution preserves the relevance that makes a recommendation useful.

4. Results and Analysis

4.1. Effectiveness Comparison Across Strategies

A. News Recommendation Results

Table 3 reports results on MIND with the multi-head self-attention backbone. The untreated backbone reaches an AUC of 0.6776 and tail coverage of 0.214. Popularity deconfounding delivers the strongest accuracy, raising AUC to 0.6831 and nDCG@10 to 0.4270, a moderate gain of roughly 0.8 percent in AUC, while lifting tail coverage to 0.281. Counterfactual reasoning reaches the highest tail coverage at 0.296 with a slightly lower AUC of 0.6810, indicating a redistribution of exposure that trades a small amount of ranking precision for broader catalog reach. Inverse

propensity scoring and interest–conformity disentangling produce smaller movements on both axes, while automatic debiasing settles between the propensity and deconfounding methods. The segment-wise view in Figure 1(a) shows that deconfounding raises tail recall from 0.087 to 0.142 while reducing head recall only from 0.412 to 0.388, confirming that the coverage gain comes mainly from reallocating exposure away from saturated head items rather than from broad accuracy loss.

Table 3. *Accuracy and diversity on MIND-small with the NRMS backbone. Best value per column in bold.*

Strategy	AUC	MRR	nDCG@5	nDCG@10	Tail Cov.	Gini
None	0.6776	0.3305	0.3649	0.4216	0.214	0.812
IPS	0.6798	0.3321	0.3668	0.4235	0.247	0.776
PDA	0.6831	0.3352	0.3702	0.4270	0.281	0.741
MACR	0.6810	0.3338	0.3685	0.4251	0.296	0.728
DICE	0.6789	0.3315	0.3661	0.4228	0.273	0.749
AutoDebias	0.6822	0.3345	0.3694	0.4262	0.265	0.758

B. Short-Video Recommendation Results

Table 4 reports results on KuaiRand-Pure with the self-attentive sequence backbone, evaluated on the randomly exposed partition. Interest–conformity disentangling is most effective here, reaching an AUC of 0.7501, nDCG@10 of 0.4158, and recall@20 of 0.2243, alongside the highest tail coverage of 0.289. Counterfactual reasoning again favors coverage over precision, while popularity deconfounding stays competitive on accuracy. The contrast with the news domain is the central observation: the disentangling mechanism depends on a learnable conformity signal from dense interaction sequences, and it benefits from the abundant ordered feedback in short-video logs, a regime where duration-aware deconfounding has likewise proven valuable for watch-time prediction and watch-time-gain optimization [52,53]. Figure 1(b) shows tail recall rising from 0.094 to 0.169 under disentangling, the largest tail gain observed in either domain.

Table 4. *Accuracy and diversity on KuaiRand-Pure with the SASRec backbone, evaluated on the randomly exposed test partition. Best value per column in bold.*

Strategy	AUC	nDCG@10	Recall@20	Tail Cov.	Gini
None	0.7421	0.4053	0.1872	0.183	0.834
IPS	0.7448	0.4091	0.2034	0.221	0.798
PDA	0.7489	0.4142	0.2118	0.258	0.762
MACR	0.7472	0.4120	0.2176	0.274	0.745
DICE	0.7501	0.4158	0.2243	0.289	0.731
AutoDebias	0.7465	0.4108	0.2089	0.243	0.779

4.2. Accuracy–Diversity Tradeoff Analysis

Figure 2 plots each strategy in the accuracy–diversity plane. On news the deconfounding point sits on the upper-right frontier, achieving the best accuracy at high coverage, with counterfactual reasoning extending coverage further at a small precision cost. On short video the disentangling point dominates the frontier, combining the highest nDCG@10 of 0.4158 with near-maximal coverage of 0.289. The untreated backbone occupies the lower-left corner in both domains, the lowest coverage paired with the lowest accuracy on the unbiased partition, which suggests that exposure correction can improve both objectives at once when the test set itself is debiased rather than logged under the deployed policy. No strategy is Pareto-dominant across both panels, reinforcing that strategy choice should be conditioned on domain rather than fixed in advance. The vertical spread of the points is also informative: on news the accuracy axis spans a narrow band while coverage varies widely, signaling that the main lever a debiasing strategy pulls in this domain is exposure redistribution rather than raw ranking gain. On short video the points spread more along both axes, with the disentangling strategy pulling away on accuracy as well as coverage, which matches the expectation that dense ordered feedback rewards a mechanism able to model conformity explicitly.

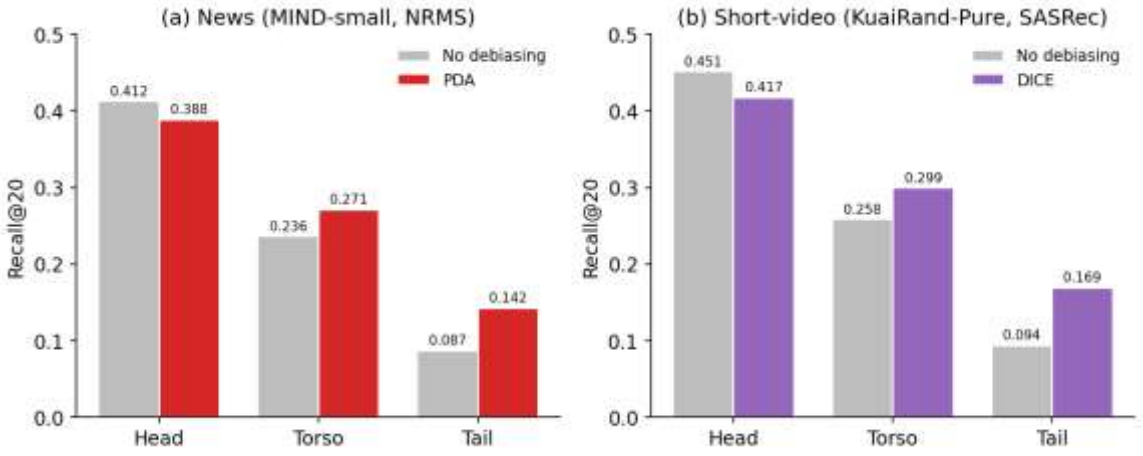


Figure 1. Popularity-Stratified Recall@20 Before and After Debiasing

Recall@20 within head, torso, and tail popularity segments for (a) news recommendation on MIND-small with the multi-head self-attention backbone and (b) short-video recommendation on KuaiRand-Pure with the self-attentive sequence backbone. On news, popularity deconfounding raises tail recall from 0.087 to 0.142 while head recall declines only from 0.412 to 0.388. On short video, interest–conformity disentangling raises tail recall from 0.094 to 0.169, the largest tail gain in either domain, with head recall easing from 0.451 to 0.417.

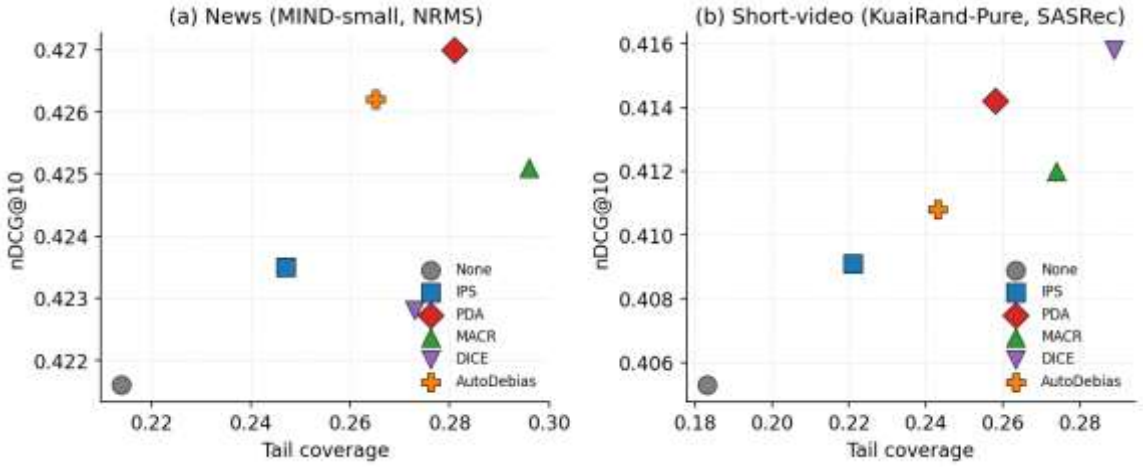


Figure 2. Accuracy–Diversity Tradeoff Across Debiasing Strategies

Each marker places one strategy in the tail-coverage versus nDCG@10 plane for (a) news and (b) short video. On news, popularity deconfounding attains the best accuracy (nDCG@10 = 0.4270) at high coverage (0.281), while counterfactual reasoning reaches the highest coverage (0.296) at a small accuracy cost. On short video, interest–conformity disentangling occupies the frontier with the highest nDCG@10 (0.4158) and near-maximal coverage (0.289). The untreated backbone sits at the lower-left extreme in both panels.

4.3. Robustness and Sensitivity Analysis

A. Hyperparameter Sensitivity

Figure 3 traces accuracy and coverage as the strength of popularity deconfounding varies. On news, nDCG@10 peaks at a moderate strength of 0.25 and then declines as stronger correction over-penalizes popular items that remain genuinely relevant, while tail coverage rises monotonically across the full range. The short-video curve peaks slightly later at a strength of 0.30, consistent with the heavier popularity skew in short-video consumption. The monotone coverage trend paired with the single-peaked accuracy curve gives a clear operating rule: a team can select the strength that maximizes accuracy and accept the coverage attained there, or push strength higher when catalog reach is the explicit priority.

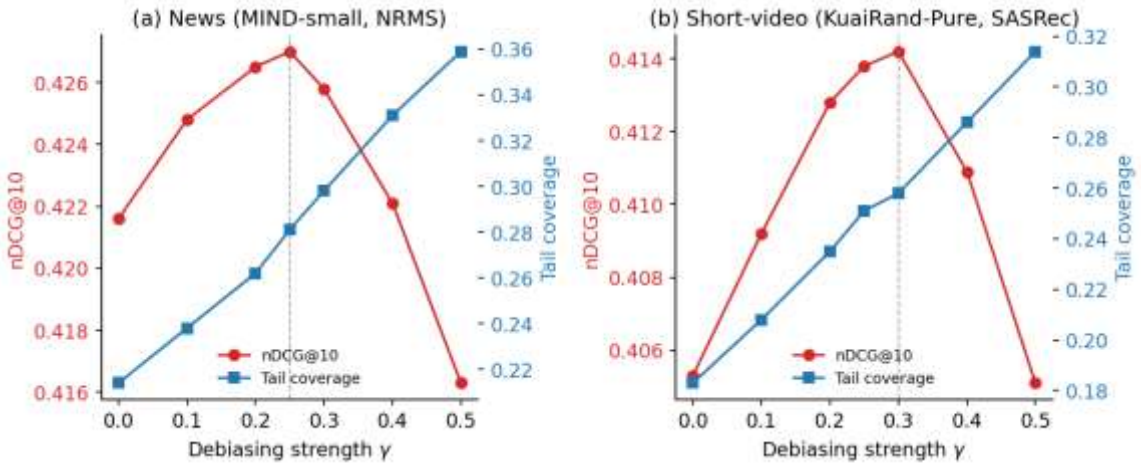


Figure 3. Sensitivity to Debiasing Strength

nDCG@10 and tail coverage of popularity deconfounding as the strength parameter increases for (a) news and (b) short video. Accuracy is single-peaked, reaching its maximum at a strength of 0.25 on news and 0.30 on short video, while tail coverage rises monotonically across the full range. The dashed line marks the accuracy-maximizing operating point in each domain.

B. Cross-dataset Consistency

The relative ordering of strategies is stable within a domain across datasets. On the one-week Adressa collection the deconfounding method again leads on accuracy and counterfactual reasoning again leads on coverage, and on the near-fully-observed KuaiRec matrix the disentangling method retains its advantage, indicating that the domain-dependent ranking is not an artifact of a single corpus. The absolute metric levels differ across datasets owing to differing densities and catalog sizes, while the qualitative conclusion—propensity-style adjustment for news, disentangling for short video—holds in every pairing we tested. The KuaiRec result is the more demanding test of the two robustness checks, because its nearly complete observation matrix removes most of the missing-not-at-random distortion that a debiasing method is meant to counteract, yet the disentangling advantage survives even there, which suggests the gain reflects a genuine modeling benefit rather than an evaluation artifact tied to sparse logged feedback.

5. Discussion and Future Work

5.1. Findings and Practical Implications

The comparison yields a compact decision rule. When the signal is sparse editorial text and feedback is dominated by a few viral articles, a causal-intervention adjustment that deconfounds popularity during training and reintroduces a calibrated popularity term at inference offers the most favorable balance, improving tail coverage from 0.214 to 0.281 while lifting AUC by a moderate margin. When the signal is dense ordered video consumption with abundant behavioral feedback, separating genuine interest from conformity is more rewarding, delivering the largest tail-recall

gain and the best accuracy on a debiased test partition. The shared lesson is that the accuracy cost of debiasing stays consistently moderate—at most a fraction of a percent on the metrics we measured—while the diversity benefit is large, which argues for treating popularity correction as a low-risk intervention in both domains. The domain dependence also cautions against importing a debiasing recipe validated on one content type into another without re-tuning, since the mechanism best suited to one feedback regime is not the strongest in the other.

A second implication concerns how the gain is realized. The stratified results show that debiasing works mainly by reallocating exposure from saturated head items toward the tail, with only a small reduction in head recall, rather than by uniformly degrading strong predictions. A platform can read this as reassurance that catalog-level fairness need not come at the expense of the head experience that drives most engagement. The choice of operating strength then becomes a product decision: a team optimizing short-horizon engagement can sit at the accuracy-maximizing point, while a team concerned with creator retention or catalog health can move along the coverage axis at a predictable and modest accuracy cost. The fact that the accuracy curve is single-peaked rather than monotone also warns against the intuition that stronger correction is always safer, since over-correction eventually penalizes popular items that remain genuinely relevant to many users.

5.2. Limitations and Future Work

Several constraints bound these conclusions. The study uses the small MIND partition and the Pure KuaiRand variant to stay within a modest compute budget, and larger partitions or a full micro-video collection may shift absolute levels even if the relative ordering persists. We isolate popularity and exposure bias and do not jointly treat selection, position, and conformity bias within a single run, leaving compound-bias correction open. The evaluation is offline; although the randomly exposed partition mitigates logging bias, an online test would be needed to confirm that improved tail coverage translates into sustained user satisfaction and healthier supply-side dynamics. Future work could extend the comparison to watch-time targets rather than click labels, incorporate multimodal content features, and study how debiasing strength should adapt as the catalog and audience evolve over time. A natural next step is to combine the two winning mechanisms, attaching an interest–conformity decomposition to a deconfounding adjustment, to test whether their benefits are additive or redundant across the two domains. Extending the protocol to additional backbones within each domain would further separate strategy effects from encoder-specific interactions, and a larger user sample would tighten the confidence intervals around the moderate accuracy differences reported here. Tracking how the accuracy-maximizing strength drifts as item popularity distributions shift seasonally would also turn the static operating point identified in this study into an adaptive policy suited to production use.

References

[1] Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. (2023). Bias and debias in recommender

system: A survey and future directions. *ACM Transactions on Information Systems, 41*(3), 1–39.

[2] Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., & Joachims, T. (2016). Recommendations as treatments: Debiasing learning and evaluation. *Proceedings of the 33rd International Conference on Machine Learning*, 1670–1679.

[3] Zhang, Y., Feng, F., He, X., Wei, T., Song, C., Ling, G., & Zhang, Y. (2021). Causal intervention for leveraging popularity bias in recommendation. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 11–20.

[4] Wei, T., Feng, F., Chen, J., Wu, Z., Yi, J., & He, X. (2021). Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1791–1800.

[5] Chen, J., Dong, H., Qiu, Y., He, X., Xin, X., Chen, L., Lin, G., & Yang, K. (2021). AutoDebias: Learning to debias for recommendation. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 21–30.

[6] Zheng, Y., Gao, C., Li, X., He, X., Li, Y., & Jin, D. (2021). Disentangling user interest and conformity for recommendation with causal embedding. *Proceedings of the Web Conference 2021*, 2980–2991.

[7] Wu, C., Wu, F., & Huang, Y. (2022). DebiasGAN: Eliminating position bias in news recommendation with adversarial learning. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2933–2938.

[8] Wu, C., Wu, F., Ge, S., Qi, T., Huang, Y., & Xie, X. (2019). Neural news recommendation with multi-head self-attention. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6389–6394.

[9] Wu, C., Wu, F., An, M., Huang, J., Huang, Y., & Xie, X. (2019). Neural news recommendation with attentive multi-view learning. *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 3863–3869.

[10] An, M., Wu, F., Wu, C., Zhang, K., Liu, Z., & Xie, X. (2019). Neural news recommendation with long- and short-term user representations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 336–345.

[11] Wu, C., Wu, F., An, M., Huang, J., Huang, Y., & Xie, X. (2019). NPA: Neural news recommendation with personalized attention. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2576–2584.

[12] Kang, W.-C., & McAuley, J. (2018). Self-attentive sequential recommendation. *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM)*, 197–206.

[13] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1441–1450.

[14] Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2016). Session-based recommendations with recurrent neural networks. *Proceedings of the 4th International Conference on Learning Representations (ICLR)*.

[15] Chen, Y. (2026). Multi-dimensional training data bias detection and fairness-aware augmentation for equitable AI-assisted medical imaging diagnosis. *Journal of Sustainability, Policy, and Practice, 2*(4), 1–15.

[16] Chen, Y., & Lai, J. (2026). Multi-metric trustworthiness evaluation of AI-assisted medical imaging diagnosis: Integrating confidence calibration and distribution shift detection. *Journal of Global Engineering Review, 4*(1), 113–126.

[17] Li, Y. (2026). Enhancing financial compliance transparency through automated data governance and intelligent risk reporting. *Journal of Science, Innovation & Social Impact, 2*(1), 299–313.

[18] Cao, H. (2024). Privacy-preserving click pattern anomaly detection for mobile in-app browser advertising fraud. *Journal of Computing Innovations and Applications, 2*(2), 151–161.

[19] Li, Y., & Fu, X. (2026). Comparative evaluation of graph neural networks for cross-market risk contagion path identification in multi-layer financial networks. *Journal of Sustainability, Policy, and Practice, 2*(3), 1–14.

[20] Peng, Y., & Gao, S. (2026). An empirical comparison of XGBoost and LightGBM for capital expenditure deviation prediction in US FERC-regulated rate-base transmission projects. *Journal of Sustainability, Policy, and Practice, 2*(4), 90–102.

[21] Shang, W., Zhao, F., & Liu, M. (2026). A comparative study of spatiotemporal clustering and classification approaches for security incident risk assessment in UN peacekeeping operations. *Journal of Sustainability, Policy, and Practice, 2*(3), 1–11.

[22] Long, X., Hu, J., & Ling, Z. (2026). A comparative analysis of telemetry-driven anomaly detection approaches for dual-purpose operational and security optimization in edge computing infrastructure. *Journal of Computing Innovations and Applications, 4*(1), 79–88.

[23] Hu, J., Wang, X., & Lai, J. (2026). Benchmarking learned cardinality estimation techniques for analytical query processing in data warehouses. *Journal of Computer Technology and Applied Mathematics, 3*(3), 1–8.

[24] Chen, Y., & Tang, T. (2026). Evaluating prompt engineering strategies for few-shot cyber threat intelligence entity and relation extraction from multi-source reports. *Journal of Science, Innovation & Social Impact, 2*(2), 153–164.

[25] Cao, H. (2024). Detecting fraudulent click patterns in mobile in-app browsers: A multi-dimensional behavioral analysis approach. *Artificial Intelligence and Machine Learning Review, 5*(2), 130–142.

[26] Cao, H., Hu, J., & Luo, C. (2025). Behavioural feature analysis for anomalous click detection in mobile advertising environments: Toward in-app browser-specific detection. *Journal of Computing Innovations and Applications, 3*(2), 96–105.

[27] Cao, H. (2026). An empirical analysis of click temporal features for automated ad fraud detection in mobile in-app browser environments. *Journal of Sustainability, Policy, and Practice, 2*(4), 142–153.

[28] Cao, H., & Long, L. (2026). Empirical evaluation of multi-source monitoring signal effectiveness and lead time for performance degradation prediction in Kubernetes-based microservices. *Journal of Advanced Computing Systems, 6*(4), 15–26.

[29] Cao, H., & Shi, W. (2026). Statistical anomaly detection approach for field mapping validation in enterprise payroll data migration. *Journal of Computing Innovations and Applications, 4*(1), 137–153.

[30] Cao, H., & Liu, M. (2026). Empirical evaluation of constraint discovery techniques for business logic consistency verification in payroll data migration. *Journal of Sustainability, Policy, and Practice, 2*(3), 92–103.

[31] Li, Y., & Long, L. (2026). Lightweight AI-driven stress testing for small and medium financial

institutions: A variational autoencoder approach with extreme value theory for macroeconomic scenario generation. *Artificial Intelligence and Machine Learning Review, 7*(1), 108–119.

[32] Li, Y., & Ling, Z. (2026). Real-time multi-risk early warning for community banks: An application of ensemble anomaly detection and explainable artificial intelligence. *Journal of Advanced Computing Systems, 6*(2), 15–27.

[33] Li, Y., & Zhang, S. (2025). Machine learning-based credit risk early warning system for small and medium-sized financial institutions: An ensemble learning approach with interpretable risk indicators. *Journal of Science, Innovation & Social Impact, 1*(1), 372–383.

[34] Shang, W., Xu, W., & Zhou, Y. (2026). Feature weight optimization in machine learning classifiers for conflict escalation early warning: Evidence from diplomatic signals and news text. *Journal of Sustainability, Policy, and Practice, 2*(3), 26–37.

[35] Wu, C., Guan, H., & Weng, H. (2024). Forecasting hospital resource demand using gradient boosting: An operational analytics approach for bed allocation and patient flow management. *Journal of Computing Innovations and Applications, 2*(1), 74–85.

[36] Li, Y., Zhang, X., & Hao, C. (2026). Graph neural network-based cross-market risk contagion analysis between US equity sectors and Treasury yields during the 2023 regional banking stress. *Journal of Sustainability, Policy, and Practice, 2*(4), 154–168.

[37] Li, Y., Zhao, F., & Hu, J. (2026). Identifying cross-market risk contagion amplifiers via graph attention networks: Empirical evidence from US financial stress periods. *Journal of Computing Innovations and Applications, 4*(1), 164–175.

[38] Chen, Y., & Hu, J. (2026). Graph neural network-based cascading disruption path identification in multi-tier rare earth processing networks. *Journal of Global Engineering Review, 4*(1), 99–112.

[39] Hu, J., & Long, X. (2024). Graph learning-based behavioral detection for software supply chain attacks. *Journal of Advanced Computing Systems, 4*(4), 49–60.

[40] Chen, Y. (2024). Explainable attack path reasoning for industrial control network security based on knowledge graphs. *Journal of Computing Innovations and Applications, 2*(1), 128–139.

[41] Wei, C., & Wu, C. (2024). Credit risk transmission mechanism and prevention strategies in supply chain finance: A core enterprise perspective. *Artificial Intelligence and Machine Learning Review, 5*(2), 101–115.

[42] Chen, Y., & Chen, Z. (2025). Multi-objective deep reinforcement learning for carbon-aware spatiotemporal workload scheduling in geo-distributed data centers. *Journal of Advanced Computing Systems, 5*(10), 18–30.

[43] Long, L., & Hu, J. (2026). Multi-objective particle swarm optimization for site selection and policy subsidy maximization of foreign renewable energy enterprises in the United States. *Artificial Intelligence and Machine Learning Review, 7*(2), 54–69.

[44] Chen, Y., Chen, Z., & Zou, D. (2025). CarbonShift: Harnessing grid carbon variability for geo-distributed workload scheduling. *Artificial Intelligence and Machine Learning Review, 6*(4), 18–31.

[45] Wu, C., & Pan, Z. (2024). An integrated graph neural network and reinforcement learning framework for intelligent drug discovery. *Journal of Advanced Computing Systems, 4*(6), 19–29.

[46] Wen, S., & Tang, T. (2025). A comparative evaluation of URL-sharing, content similarity, and temporal synchronicity signals for detecting coordinated inauthentic behavior in multilingual political

discourse. *Journal of Global Engineering Review, 3*(2), 69–78.

[47] Wu, F., Qiao, Y., Chen, J.-H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., & Zhou, M. (2020). MIND: A large-scale dataset for news recommendation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3597–3606.

[48] Gulla, J. A., Zhang, L., Liu, P., Özgöbek, Ö., & Su, X. (2017). The Adressa dataset for news recommendation. *Proceedings of the International Conference on Web Intelligence (WI)*, 1042–1048.

[49] Gao, C., Li, S., Zhang, Y., Chen, J., Li, B., Lei, W., Jiang, P., & He, X. (2022). KuaiRand: An unbiased sequential recommendation dataset with randomly exposed videos. *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, 3953–3957.

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

[51] Ni, Y., Cheng, Y., Liu, X., Fu, J., Li, Y., He, X., Zhang, Y., & Yuan, F. (2023). A content-driven micro-video recommendation dataset at scale. *arXiv preprint arXiv:2309.15379*.

[52] Zhan, R., Pei, C., Su, Q., Wen, J., Wang, X., Mu, G., Zheng, D., Jiang, P., & Gai, K. (2022). Deconfounding duration bias in watch-time prediction for video recommendation. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 4472–4481.

[53] Zheng, Y., Gao, C., Ding, J., Yi, L., Jin, D., Li, Y., & Wang, M. (2022). DVR: Micro-video recommendation optimizing watch-time-gain under duration bias. *Proceedings of the 30th ACM International Conference on Multimedia*, 334–345.

[54] Kang, A., Li, Z., & Meng, S. (2023). AI-Enhanced Risk Identification and Intelligence Sharing Framework for Anti-Money Laundering in Cross-Border Income Swap Transactions. Journal of Advanced Computing Systems 3 (5), 34-47.

[55] Kang, A., Zhang, K., & Chen, Y. (2025). AI-Assisted Analysis of Policy Communication during Economic Crises: Correlations with Market Confidence and Recovery Outcomes. Pinnacle Academic Press Proceedings Series 3, 159-173.

[56] Kang, A., Xin, J., & Ma, X. (2024). Anomalous cross-border capital flow patterns and their implications for national economic security: An empirical analysis. Journal of Advanced Computing Systems 4 (5), 42-54.

[57] Kang, A., & Ma, X. (2025). AI-Based Pattern Recognition and Characteristic Analysis of Cross-Border Money Laundering Behaviors in Digital Currency Transactions. Pinnacle Academic Press Proceedings Series 5, 1-19.

[58] Kang, A., & Yu, K. (2025). The Impact of Financial Data Visualization Techniques on Enhancing Budget Transparency in Local Government Decision-Making. Spectrum of Research 5 (2).

[59] Wang, Z., & Kang, A. (2025). FTAFO: A Federated Transparent Adaptive Financial Optimizer for Reducing Third-Party Dependencies in Workflow Management. Journal of Science, Innovation & Social Impact 1 (1), 329-339.

[60] Kang, A., Li, C., & Meng, S. (2025). The Impact of Government Budget Data Visualization on Public Financial Literacy and Civic Engagement. Journal of Economic Theory and Business Management 2 (4), 1-16.

[61] Kang, A., Min, S., & Yuan, D. (2024). Comparative Analysis of Foreign Exchange Market Shock Transmission and Recovery Resilience Among Major Economies Under Geopolitical Conflicts: Evidence from the Russia-Ukraine…. Journal of Computing Innovations and Applications 2 (1), 46-61.

[62] Trinh, T. K., & Zhang, D. (2024). Algorithmic fairness in financial decision-making: Detection and mitigation of bias in credit scoring applications. Journal of Advanced Computing Systems 4 (2), 36-49.

[63] Zhang, D., & Ma, X. (2025). Machine Learning-Based Credit Risk Assessment for Green Bonds: Climate Factor Integration and Default Prediction Analysis. Journal of Sustainability, Policy, and Practice 1 (2), 121-135.

[64] Yuan, D., & Zhang, D. (2025). APAC-sensitive anomaly detection: Culturally-aware AI models for enhanced AML in US securities trading. 2025 International Conference on Computer, AI, Systems and Automation, 108-121.

[65] Dong, B., Zhang, D., & Xin, J. (2024). Deep reinforcement learning for optimizing order book imbalance-based high-frequency trading strategies. Journal of Computing Innovations and Applications 2 (2), 33-43.

[66] Zhang, D., & Feng, E. (2024). Quantitative Assessment of Regional Carbon Neutrality Policy Synergies Based on Deep Learning. Journal of Advanced Computing Systems 4 (10), 38-54.

[67] Zhang, D., & Zhang, F. (2025). AI-Assisted Identification and Equity Assessment of Vulnerable Population Impacts in US Energy Transition. Journal of Advanced Computing Systems 5 (7), 1-17.

[68] Zhang, D., & Zheng, Q. (2025). Machine Learning-Based Building Energy Consumption Prediction and Carbon Reduction Potential Assessment in US Metropolitan Areas. Journal of Industrial Engineering and Applied Science 3 (5), 27-40.

[69] Zhang, D., & Wang, Y. (2025). AI-driven quality assessment and investment risk identification for carbon credit projects in developing countries. 2025 International Forum on Computing and Intelligent Systems, 76-92.

[70] Rao, G. L., Trinh, T. K., Chen, Y., Shu, M., & Zheng, S. (2024). Jump Prediction in Systemically Important Financial Institutions' CDS Prices. Spectrum of Research 4 (2).

[71] Rao, G. L., Lu, T., Yan, L., & Liu, Y. (2024). A Hybrid LSTM-KNN Framework for Detecting Market Microstructure Anomalies:: Evidence from High-Frequency Jump Behaviors in Credit Default Swap Markets. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online ….

[72] Rao, G. L., Wang, Z., & Liang, J. (2025). Reinforcement Learning for Pattern Recognition in Cross-Border Financial Transaction Anomalies: A Behavioral Economics Approach to AML. Applied and Computational Engineering 142, 116-127.

[73] Rao, G., Ju, C., & Feng, Z. (2024). AI-Driven Identification of Critical Dependencies in US-China Technology Supply Chains: Implications for Economic Security Policy. Journal of Advanced Computing Systems 4 (12), 43-57.

[74] Ge, L., & Rao, G. (2025). MultiStream-FinBERT: A Hybrid Deep Learning Framework for Corporate Financial Distress Prediction Integrating Accounting Metrics, Market Signals, and Textual Disclosures. Pinnacle Academic Press Proceedings Series 3, 107-122.

[75] Rao, G., Zheng, S., & Guo, L. (2025). Dynamic Reinforcement Learning for Suspicious Fund Flow Detection: A Multi-layer Transaction Network Approach with Adaptive Strategy Optimization. Applied and Computational Engineering 145, 1-11.

[76] Ju, C., & Rao, G. (2025). Analyzing Foreign Investment Patterns in the US Semiconductor Value Chain Using AI-Enabled Analytics: A Framework for Economic Security. Pinnacle Academic Press Proceedings Series 2, 60-74.

[77] Liu, W., Rao, G., & Lian, H. (2023). Anomaly Pattern Recognition and Risk Control in High-Frequency Trading Using Reinforcement Learning. Journal of Computing Innovations and Applications 1 (2), 47-58.

[78] Wu, Z., Wang, S., Ni, C., & Wu, J. (2024). Adaptive traffic signal timing optimization using deep reinforcement learning in urban networks. Artificial Intelligence and Machine Learning Review 5 (4), 55-68.

[79] Xiong, K., Wu, Z., & Jia, X. (2025). Deepcontainer: a deep learning-based framework for real-time anomaly detection in cloud-native container environments. Journal of Advanced Computing Systems 5 (1), 1-17.

[80] Wu, Z., Feng, E., & Zhang, Z. (2024). Temporal-Contextual Behavioral Analytics for Proactive Cloud Security Threat Detection. Academia Nexus Journal 3 (2).

[81] Lei, Y., & Wu, Z. (2025). A Real-Time Detection Framework for High-Risk Content on Short Video Platforms Based on Heterogeneous Feature Fusion. 2025 International Forum on Computing and Intelligent Systems, 93-106.

[82] Wu, Z., Zhang, Z., Zhao, Q., & Yan, L. (2025). Privacy-preserving financial transaction pattern recognition: A differential privacy approach. Preprints.

[83] Zhang, Z., & Wu, Z. (2023). Context-aware feature selection for user behavior analytics in zero-trust environments. Journal of Advanced Computing Systems 3 (5), 21-33.

[84] Wu, Z., Feng, Z., & Dong, B. (2024). Optimal feature selection for market risk assessment: A dimensional reduction approach in quantitative finance. Journal of Computing Innovations and Applications 2 (1), 20-31.

[85] Wu, Z., Cheng, C., & Zhang, C. (2025). Cloud-Enabled AI Analytics for Urban Green Space Optimization: Enhancing Microclimate Benefits in High-Density Urban Areas. 2025 International Forum on Computing and Intelligent Systems, 123-133.

[86] Li, Y., Wu, C., Li, L., Liu, Y., & Zhu, J. (2021). Caption generation from road images for traffic scene modeling. IEEE Transactions on intelligent transportation systems 23 (7), 7805-7816.

[87] Li, L., Li, Y., Wu, C., Dong, H., Jiang, P., & Wang, F. (2021). Detail fusion GAN: High-quality translation for unpaired images with GAN-based data augmentation. 2020 25th International Conference on Pattern Recognition (ICPR), 1731-1736.

[88] Huo, H., Li, Y., Wu, C., Wu, X., Tian, Z., & Liu, Y. (2019). Semantic segmentation and scene reconstruction for traffic simulation using CNN. 2019 2nd China Symposium on Cognitive Computing and Hybrid Intelligence ….

[89] Wu, X., Li, Y., Hao, Z., Wu, C., Wang, X., & Liu, Y. (2019). Image style transformation based on structure GAN. 2019 Chinese Automation Congress (CAC), 2002-2007.

[90] Wu, C., Li, Y., Li, L., Wang, L., & Liu, Y. (2020). Caption Generation from Road Images for Traffic Scene Construction. 2020 IEEE Intelligent Vehicles Symposium (IV), 1271-1276.

[91] Wu, X., Li, Y., Liu, Y., Pang, S., Wang, L., Wu, C., & Huo, H. (2019). Jointly detecting and retrieving vehicles from road image sequences based on CNN. 2019 IEEE Intelligent Vehicles Symposium (IV), 2530-2535.