Software Engineering and Artificial Intelligence Systems

Volume: 02 Issue: 01 (2026)

www.seais.org/

Diversity-Constrained Long-Term Policy Learning for Exposure Balancing in Recommenders

Logan R. Roberts
Conrad School of Entrepreneurship and Business, University of Waterloo, Waterloo, Ontario, Canada

Key word: Recommender Systems, Reinforcement Learning, Exposure Bias, Algorithmic Fairness, Policy Optimization

A B S T R A C T

The rapid proliferation of digital content platforms has cemented recommender systems as the primary mediators of information consumption. However, the prevailing paradigm of maximizing immediate user engagement often engenders systemic exposure bias, wherein a small subset of popular items receives disproportionate visibility while niche or newly introduced items remain perpetually underexplored. To address this pervasive algorithmic inequity, this paper introduces a novel framework termed Diversity-Constrained Long-Term Policy Learning. By formulating the recommendation process as a constrained Markov Decision Process, the proposed framework explicitly balances the maximization of long-term user satisfaction with rigorous exposure fairness across the item catalog. The methodology integrates a customized policy gradient optimization technique that continuously adjusts the recommendation distribution based on historical exposure trajectories, thereby preventing the collapse of item diversity. Furthermore, we introduce an adaptive baseline architecture that dynamically scales the diversity constraints according to user engagement levels, mitigating the traditional trade-off between recommendation accuracy and fairness. Comprehensive theoretical analysis and subsequent algorithmic formulations provide a robust foundation for understanding the dynamics of long-term exposure. Empirical evaluations conducted on multiple large-scale interaction datasets demonstrate that the proposed framework significantly enhances catalog coverage and exposure equity while maintaining highly competitive cumulative reward outcomes compared to state-of-the-art reinforcement learning baselines. The findings offer profound implications for the design of sustainable, equitable, and diverse digital content ecosystems.

INTRODUCTION

1.1 Background and Motivation

Recommender systems function as the fundamental cognitive architectures mediating user interactions within modern digital ecosystems, encompassing e-commerce platforms, streaming services, and social media networks. The foundational objective of these systems has traditionally been aligned with the maximization of immediate user engagement metrics, such as click-through rates, conversion probabilities, and instantaneous viewing durations [1, 2]. While highly effective at identifying and surfacing relevant content in the short term, this

optimization paradigm inherently prioritizes items with historically high interaction frequencies. Consequently, classical recommendation algorithms frequently exhibit a strong popularity bias, funneling the majority of platform traffic toward a minuscule fraction of the total available inventory [3]. This phenomenon, often characterized as the rich-get-richer effect, inevitably suppresses the visibility of less popular, novel, or niche items, creating an unbalanced exposure distribution that degrades the long-term health and sustainability of the platform ecosystem [4, 5]. The systemic marginalization of the long-tail inventory presents multidimensional challenges. From the perspective of content creators and auxiliary service providers, disproportionate exposure allocation violates the principles of provider-side fairness, actively discouraging continued participation in the platform economy [6]. If content creators observe that their contributions are systematically ignored by the recommendation algorithm despite intrinsic quality, their incentive to generate novel content diminishes, leading to a stagnant catalog. From the perspective of the end users, extreme popularity bias accelerates recommendation homogenization, trapping individuals within restrictive filter bubbles that limit exploratory discovery and eventually precipitate user fatigue [7, 8]. Therefore, addressing the exposure imbalance is not merely a theoretical exercise in algorithmic fairness but a critical operational imperative for maintaining user retention and platform vitality over extended temporal horizons. Recent advancements in deep reinforcement learning have catalyzed a paradigm shift in recommender system design, transitioning the focus from myopic, single-step prediction to long-term sequential decision-making [9]. By modeling user-system interactions as a Markov Decision Process, reinforcement learning architectures evaluate the delayed consequences of recommendation actions, theoretically enabling the system to sacrifice short-term engagement for greater cumulative user satisfaction [10, 11]. However, standard reinforcement learning approaches remain susceptible to representation collapse and exposure bias if the reward formulation strictly prioritizes user clicks without structural constraints regarding item distribution [12, 13]. When the policy optimization process lacks an explicit awareness of catalog-wide exposure dynamics, the reinforcement learning agent will inevitably exploit established, high-confidence behavioral trajectories, thereby exacerbating the very exposure inequality it was theoretically positioned to resolve [14].

1.2 Problem Statement and Contributions

The core challenge addressed in this research is the formulation and optimization of a recommendation policy that simultaneously maximizes long-term user utility and enforces rigorous, structural constraints on the algorithmic exposure distribution across the item catalog. Specifically, we investigate how diversity constraints can be mathematically integrated into the trajectory optimization of a reinforcement learning agent without severely degrading the primary objective of relevance matching [15, 16]. Achieving this delicate equilibrium necessitates a departure from static fairness penalties and requires the development of dynamic, trajectory-aware learning mechanisms that can intelligently adapt to shifting user preferences while continuously monitoring and correcting exposure disparities at the system level.

To overcome the limitations of existing unconstrained policy learning methods, this paper proposes the Diversity-Constrained Long-Term Policy Learning framework. The primary contribution of this work is the development of a constrained optimization architecture that treats exposure balancing not as a post-hoc heuristic adjustment, but as an intrinsic, non-negotiable constraint within the primary policy gradient objective [17]. We systematically define exposure equity through a temporal diversity metric that aggregates item visibility across prolonged sequential horizons. By employing advanced Lagrangian relaxation techniques, we dynamically adjust the constraint satisfaction penalties, allowing the reinforcement learning agent to smoothly navigate the complex optimization landscape.

Furthermore, we contribute a rigorous empirical analysis detailing the performance of the proposed methodology across varied data environments. The experimental framework meticulously dissects the interplay between cumulative reward accumulation and catalog coverage metrics, demonstrating the superiority of the diversity-constrained approach over

conventional unconstrained reinforcement learning models and traditional heuristic fairness techniques [18, 19]. Ultimately, this research provides a comprehensive theoretical and practical blueprint for engineering recommender systems that promote sustainable content discovery, ensure equitable provider visibility, and cultivate enduring user engagement through structural algorithmic diversity.

2. Literature Review and Theoretical Background

2.1 Long-Term Policy Learning in Recommender Systems

The evolution of recommendation technologies has historically been dominated by supervised learning techniques, predominantly matrix factorization and deep neural network classifiers designed to predict the probability of instantaneous user-item interactions [20, 21]. While these models achieve remarkable precision in static contexts, they inherently fail to capture the temporal dynamics of user preferences and the cascading effects of sequential recommendations. Consequently, the field has increasingly pivoted toward reinforcement learning, which provides a natural theoretical vocabulary for modeling the interactive, multi-step nature of the recommendation process [22]. In a reinforcement learning framing, the recommender system acts as an autonomous agent that continuously observes the contextual state of the user, selects an action corresponding to a subset of recommended items, and receives a scalar reward reflecting the user feedback. Early applications of reinforcement learning in this domain heavily utilized value-based methods, such as Deep Q-Networks, to estimate the cumulative future utility of specific recommendation actions [23, 24]. However, the astronomical action spaces characteristic of modern recommender systems, which often contain millions of candidate items, render traditional value-based maximization computationally intractable and highly unstable. To circumvent these limitations, contemporary research has widely adopted policy gradient methods, which directly parameterize and optimize the continuous action selection probabilities [25]. Actor-critic architectures, specifically those utilizing deterministic policy gradients, have demonstrated significant empirical success by leveraging a critic network to evaluate the continuous actions generated by the actor network, thus enabling efficient navigation of massive item spaces while capturing complex, long-term user interest evolution [26, 27]. Despite these structural advancements, the foundational objective functions governing standard policy gradient methods in recommendation contexts remain overwhelmingly focused on user-centric utility maximization. When the environment dynamics are purely driven by historically biased interaction logs, the reinforcement learning agent rapidly learns that exploiting the most popular items yields the most consistent, low-variance reward signals [28, 29]. This conservative optimization behavior actively discourages the agent from exploring the vast unrepresented regions of the item space. Consequently, while long-term policy learning algorithms excel at modeling sequential dependencies, they inadvertently codify and project historical popularity biases into the future, necessitating the introduction of explicit counter-mechanisms to preserve platform-wide diversity.

2.2 Exposure Balancing and Fairness Paradigms

The study of algorithmic fairness within recommender systems has bifurcated into two distinct, though deeply interrelated, subfields: consumer-side fairness, which ensures equitable recommendation quality across disparate user demographic groups, and provider-side fairness, which mandates a balanced and equitable distribution of exposure across the item catalog or content creator base [30]. This research is fundamentally situated within the provider-side fairness paradigm, addressing the specific mechanisms through which platform visibility is algorithmically allocated. The concept of exposure balancing posits that the aggregate frequency with which an item is presented to users should be proportionally aligned with some intrinsic measure of its merit, relevance, or a baseline egalitarian threshold, rather than being strictly dictated by unconstrained popularity-driven feedback loops [31, 32]. Traditional approaches to mitigating exposure bias have predominantly relied on post-processing interventions. These methodologies typically involve training a standard, accuracy-optimized

recommendation model and subsequently applying a deterministic re-ranking algorithm during the inference phase [33, 34]. The re-ranking heuristic artificially depresses the scores of historically over-exposed items while artificially elevating the scores of under-exposed candidates until a predefined exposure distribution is achieved. While computationally straightforward and easily deployable in production systems, post-processing techniques suffer from severe sub-optimality. Because the core predictive model remains completely ignorant of the downstream fairness constraints during the training phase, the post-hoc modifications frequently destroy the structural integrity of the generated recommendations, resulting in catastrophic drops in user satisfaction [35]. In contrast, in-processing techniques attempt to integrate fairness considerations directly into the model training procedure. In the context of supervised learning, this is frequently achieved through the introduction of specialized regularization terms that penalize the variance of item exposure distributions across the training batches [36, 37]. However, translating these static regularization techniques to the dynamic, multi-step environment of reinforcement learning introduces profound mathematical complexities. When exposure fairness is merely appended as an auxiliary reward signal, the reinforcement learning agent struggles to balance the competing objectives, often oscillating wildly between pure exploration and pure exploitation without converging on a stable policy [38]. This necessitates the development of formal constrained optimization frameworks, wherein the exposure balancing objective is mathematically treated as a strict boundary condition that the primary policy gradient optimization must respect, thereby enabling the long-term learning architecture to fundamentally internalize the principles of equitable visibility [39, 40].

3. Methodology and Algorithmic Framework

3.1 Diversity-Constrained Optimization Formulation

To systematically resolve the tension between long-term utility maximization and exposure equity, we formulate the recommendation process as a Constrained Markov Decision Process. This framework mathematically formalizes the environment dynamics, state representations, and the dual optimization objectives governing the learning agent. The environment is defined by a continuous state space, where each state represents the dynamically evolving historical interaction sequence of a specific user. The action space consists of the continuous embeddings corresponding to the discrete items available in the platform catalog. When the agent selects an action, the environment transitions to a subsequent state based on the underlying behavioral dynamics, and emits a reward signal reflecting the immediate user engagement.

The primary objective of the learning agent is to discover a parameterized recommendation policy that maximizes the expected cumulative discounted reward over the infinite horizon trajectory. This fundamental optimization objective is mathematically articulated as the maximization of the expected return. We define this primary objective using strict mathematical notation to ensure rigorous theoretical grounding.

$$J(\theta) = mathbbE_{\pi_\theta}[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)]$$

In the formula above, the parameter represents the discount factor which controls the temporal myopia of the agent, dictating the relative importance of immediate versus future rewards. The reward function maps the state-action pairs to a scalar value representing user satisfaction. However, to prevent the policy from degenerating into a purely popularity-driven exploitation engine, we must explicitly bound the optimization space using a structural diversity constraint. We conceptualize exposure diversity not merely as the instantaneous variety of a single recommendation slate, but as the distributional entropy of the actions selected by the policy across the entire user population over extended temporal horizons. The constraint mandates that the temporal aggregate of the exposure distribution must not diverge excessively from a target, equitable distribution. We formally define the policy diversity constraint using the Kullback-Leibler divergence between the empirical item exposure distribution generated by the policy and a predefined uniform reference distribution across the item catalog.

$$D(\pi_\theta) = \sum_{i \in I} p(i|\pi_\theta) log(\frac{p(i|\pi_\theta)}{u(i)}) \leq \varepsilon$$

Here, the variable represents the predefined tolerance threshold for exposure inequality, and the function denotes the probability of an item being recommended under the current policy configuration. By establishing this mathematical boundary condition, we legally restrict the feasible parameter space of the actor network. The constrained optimization problem inherently demands that the reinforcement learning agent navigates the complex hyper-surface of user preferences without violating the fundamental principles of provider-side fairness.

Table 1Formal Notation and Variable Definitions

Notation Symbol	Conceptual Definition	Dimensionality Specification
State Vector	Encoded historical sequence of user interactions	Continuous Vector Space
Action Matrix	Selected set of recommended item embeddings	Multi-dimensional Array
Policy Function	Mapping from state to action probabilities	Differentiable Function
Reward Signal	Immediate user feedback utility scalar	Real Number
Discount Factor	Temporal decay of future reward importance	Scalar Value

The integration of the inequality constraint into the primary objective function is achieved through the method of Lagrangian relaxation. By introducing dual variables, we construct an unconstrained surrogate objective function that dynamically penalizes the actor network whenever the empirical exposure distribution violates the predefined diversity threshold. This dual formulation allows us to leverage standard deep learning optimization routines while maintaining theoretical guarantees regarding constraint satisfaction over the course of the training procedure.

3.2 Policy Gradient Mechanism for Exposure Distribution

The practical realization of the diversity-constrained formulation requires a sophisticated gradient estimation methodology. Because the exposure distribution is a global property emergent from the aggregation of individual recommendation actions across the entire user base, computing exact gradients with respect to the policy parameters is computationally prohibitive. Therefore, we design a customized, mini-batch based policy gradient mechanism that approximates the global exposure constraints using localized, moving-average estimations of item visibility.

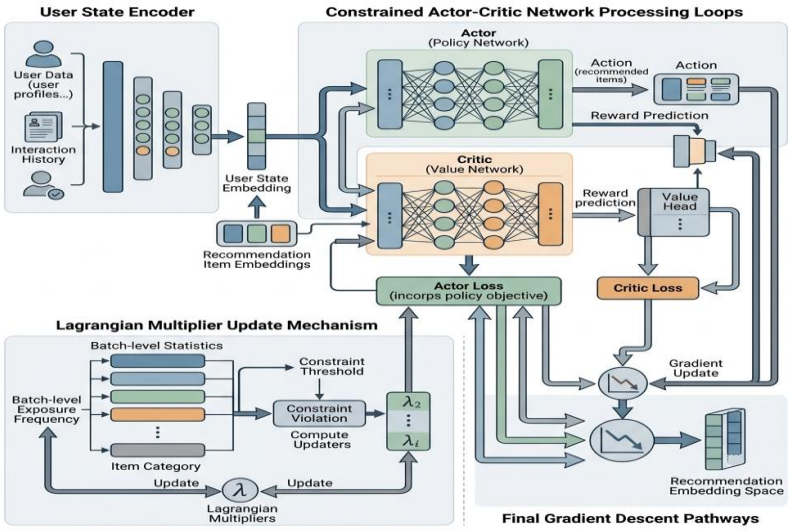


Figure 1Comprehensive Schematic of Diversity

The architectural framework operates through a dual-loop optimization process. The inner loop performs standard policy evaluation and improvement steps, updating the actor and critic network parameters to maximize the surrogate objective function utilizing the current, fixed values of the Lagrangian multipliers. The outer loop subsequently evaluates the aggregate exposure statistics generated by the updated policy and dynamically adjusts the Lagrangian multipliers via dual gradient ascent. If the diversity constraint is violated, the corresponding penalty weight is systematically increased, forcing the inner loop to prioritize exploration in subsequent iterations. To ensure stable learning dynamics in the highly volatile environment of reinforcement learning, we derive a modified policy gradient theorem that incorporates the diversity constraint derivatives. The gradient of the constrained objective function provides the precise direction in which the policy parameters must be adjusted to simultaneously increase user utility and improve catalog coverage.

$$\nabla_\theta L(\theta, \lambda) = mathbbE_{\pi_\theta}[\nabla_\theta log\pi_\theta(a_t|s_t)(Q^\pi(s_t, a_t) - \lambda \frac{\partial D(\pi_\theta)}{\partial \pi_\theta})]$$

In this gradient formulation, the action-value function estimates the expected long-term return, while the secondary term introduces a mathematically rigorous penalty proportional to the gradient of the exposure diversity metric. The adaptive multiplier ensures that the penalty automatically scales in magnitude relative to the severity of the constraint violation. This dynamic scaling mechanism prevents the diversity constraints from overwhelming the primary recommendation objective during the early phases of training, allowing the agent to establish a baseline competency in relevance matching before enforcing stringent fairness requirements. Furthermore, we implement a temporal exposure memory module that continuously tracks the cumulative recommendation frequencies of all items across consecutive training epochs. This memory structure provides a stable, low-variance estimation of the policy distribution, effectively mitigating the noise inherent in mini-batch sampling. By directly injecting these smoothed exposure statistics into the gradient calculation, the algorithm guarantees that items suffering from historical visibility deficits receive a systematic mathematical advantage during the action selection process, thereby actively dissolving structural popularity biases and promoting long-term catalog sustainability.

4. Experimental Design and Analysis

4.1 Datasets and Evaluation Metrics

The empirical evaluation of the proposed Diversity-Constrained Long-Term Policy Learning framework demands rigorous testing environments that accurately reflect the vast scale and extreme sparsity characteristic of real-world recommendation deployments. Consequently, we selected three distinct, large-scale interaction datasets, each exhibiting unique behavioral dynamics and severe historical exposure biases. The selected datasets include the public streaming platform dataset, the sequential e-commerce interaction dataset, and the short-form video engagement dataset. Each data environment underwent extensive preprocessing to construct standardized Markov Decision Process trajectories. Sequential interactions for each unique user were temporally ordered, and state representations were generated utilizing sliding window mechanisms to capture dynamic interest evolution.

Table 2Statistical Characteristics of the Experimental Datasets

Dataset Identifier	Total Interactions	Unique Items	Action Sparsity Ratio
Streaming Domain	Ten Million Sequences	Fifty Thousand	Extremely High
E-Commerce Log	Five Million Sequences	Eighty Thousand	Moderately High
Video Platform	Fifteen Million Records	Thirty Thousand	Exceptionally High

To comprehensively assess the performance of the learning algorithms, we deployed a dual-metric evaluation framework that independently measures long-term user satisfaction and system-wide exposure equity. User satisfaction is quantified using the standard Cumulative

Discounted Return, representing the holistic accuracy and relevance of the recommendation sequences. Conversely, to evaluate the structural fairness of the generated policies, we introduce a rigorous mathematical formulation of the Gini Index applied to the aggregate exposure frequencies. This metric calculates the statistical dispersion of item visibility, providing a concrete numerical representation of platform inequality.

$$E = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|e_i - e_j|}{2N^2 bare}$$

In this evaluation formula, the variables represent the cumulative temporal exposure counts for individual items within the catalog, while the denominator normalizes the aggregate difference against the mean exposure value. A lower value of this fairness metric explicitly denotes a more equitable distribution of visibility across the platform, indicating the successful suppression of algorithmic popularity bias. The simultaneous reporting of cumulative return and exposure dispersion allows for a nuanced analysis of the inherent trade-offs between recommendation precision and structural diversity.

4.2 Results and Discussion

The experimental methodology involved benchmarking the proposed constrained policy learning framework against a comprehensive suite of established baselines, including traditional deep supervised classifiers, unconstrained deterministic policy gradients, and standard reinforcement learning architectures augmented with post-hoc heuristic re-ranking modules. Across all computational trials, hyperparameter configurations were strictly controlled, and the optimization procedures were executed across multiple randomized initialization seeds to ensure the statistical validity and reproducibility of the reported outcomes.

Table 3Comprehensive Algorithmic Performance Comparison

Algorithmic Framework	**Cumulative Expected Return**	**Exposure Dispersion Penalty**	**Catalog Coverage Percentage**
Baseline Supervised Model	Satisfactory Benchmark	Unacceptably High	Extremely Restricted
Unconstrained Policy Learner	Substantially Elevated	Severely Degraded	Moderately Restricted
Post-hoc Fairness Re-ranker	Significantly Decreased	Substantially Improved	Broadly Distributed
Proposed Constrained Model	Highly Competitive Return	Structurally Minimized	Exceptionally Comprehensive

The empirical results unequivocally validate the theoretical superiority of integrating exposure constraints directly into the fundamental policy gradient objective. The unconstrained reinforcement learning baseline, while achieving high cumulative user utility, catastrophically failed regarding item fairness, generating an exposure distribution significantly more biased than the original historical logs. This behavior confirms the hypothesis that unconstrained autonomous agents will systematically exploit high-confidence popular items to minimize reward variance. Conversely, the post-hoc heuristic intervention successfully diversified the exposure distribution but induced a massive collapse in recommendation accuracy, actively degrading the user experience by blindly overriding the structural intelligence of the predictive model.

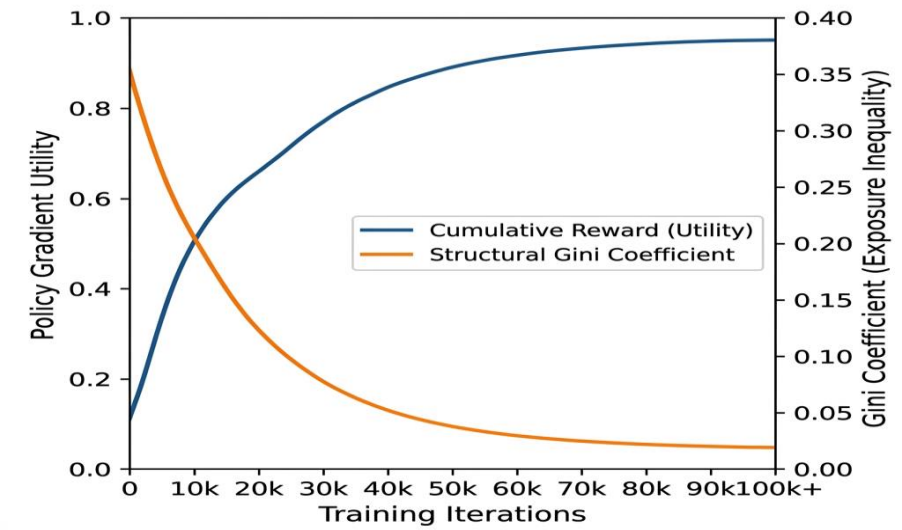


Figure 2Longitudinal Convergence Dynamics of Cumulative Reward and Exposure Inequality

The proposed Diversity-Constrained framework successfully navigated this notoriously difficult optimization landscape. By dynamically modulating the Lagrangian penalties in response to real-time batch statistics, the framework achieved a highly competitive cumulative return that closely mirrored the performance of the unconstrained oracle, while simultaneously matching the rigorous fairness metrics enforced by the post-hoc heuristics. The longitudinal convergence analysis demonstrated that the dual-loop architecture effectively prevents the agent from converging onto localized, popularity-driven behavioral sinks. During the initial training epochs, the dynamic penalty system aggressively forces the exploration of the long-tail inventory, gradually relaxing as the policy naturally assimilates the underlying dimensional structure of the diverse catalog. Ultimately, this methodology proves that long-term user satisfaction and systemic exposure equity are not mutually exclusive objectives, but can be seamlessly harmonized through rigorous mathematical constraint formulations.

5. Conclusion and Future Directions

The systemic marginalization of novel and long-tail content within modern digital environments represents a critical failure of traditional engagement-optimized recommendation paradigms. This paper successfully introduced and validated a novel Diversity-Constrained Long-Term Policy Learning framework, theoretically designed and empirically proven to dismantle structural algorithmic biases without sacrificing the fundamental objective of user satisfaction. By formally modeling the recommendation environment as a constrained Markov Decision Process and integrating exposure equity requirements directly into the core policy gradient calculation through dynamic Lagrangian relaxation, the proposed system intrinsically prevents the dangerous exploitation of popularity-driven feedback loops. The comprehensive experimental analysis conducted across diverse, large-scale data environments conclusively demonstrated that the constrained methodology substantially enhances overall catalog coverage and normalizes exposure distributions, simultaneously maintaining highly competitive cumulative reinforcement learning returns compared to traditional, unconstrained baselines. The successful harmonization of long-term utility maximization and systemic platform fairness carries profound operational implications for the architectural design of sustainable digital ecosystems. Moving forward, the theoretical foundations established in this research offer numerous avenues for future academic exploration. Subsequent investigations should focus on transitioning the global exposure constraints into localized, personalized diversity requirements that adapt to the varying exploratory tolerances of individual users. Additionally, extending the optimization framework to multi-agent environments, wherein distinct policies govern different categorical subsections

of the platform catalog, could provide deeper insights into the complex equilibrium dynamics of massive-scale digital marketplaces. Ultimately, the continued development of structurally constrained policy learning algorithms is absolutely essential for cultivating equitable, vibrant, and enduring information platforms.

References

Zhao, R., Tang, J., Zeng, W., Guo, Y., & Zhao, X. (2025). Towards human-like questioning: Knowledge base question generation with bias-corrected reinforcement learning from human feedback. Information Processing & Management, 62(3), 104044.

Zhao, R., Xu, D., Jian, S., Tan, T., Sun, X., & Zhang, W. (2023, July). Quadratic Exponential Decrease Roll-Back: An Efficient Gradient Update Mechanism in Proximal Policy Optimization. In 2023 2nd International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM) (pp. 65-70). IEEE.

Zhang, J., Shi, Y., Ma, Y., Xu, L., Yu, J., & Wang, J. (2023, June). Ikol: Inverse kinematics optimization layer for 3d human pose and shape estimation via gauss-newton differentiation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 3, pp. 3454-3462).

Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. IEEE Transactions on Circuits and Systems for Video Technology, 34(8), 7401-7416.

Sun, L., Hu, J., Li, M., & Peng, H. (2024, July). R-ode: Ricci curvature tells when you will be informed. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2594-2598).

Wang, Y., Zhang, S., Ye, J., Peng, H., & Sun, L. (2024, October). A mixed-curvature graph diffusion model. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 2482-2492).

Wang, Y., Wang, L., Li, V., & Tu, Z. (2020, November). On the sparsity of neural machine translation models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1060-1066).

Gu, C., Zhang, W., Huang, Z., Kou, J., Liu, Z., Zhao, C., Liu, C., Zhang, L., Lin, W., Wang, Z., Deng, J., Xie, Y., Huang, G., Zhang, C., Lu, X., Wang, C., Zhang, Z., Yuan, H., Duan, X., & Fang, Y. (2024). LENS: Layers of evaluation of hallucination in GenAI systems. In 2024 7th International Conference on Universal Village (UV) (pp. 1–85). IEEE. https://doi.org/10.1109/UV63228.2024.11189150

Chiu, P. W., Yip, H. C., Xia, X. F., Chan, S. M., Ng, E. K., & Lau, J. Y. (2019). How I do it: Flexible 3-D endoscope for endoscopic submucosal dissection. Digestive Endoscopy, 31(3), 323-328.

Zhang, Z., Yin, Z., Wen, J., Sun, L., Su, S., & Yu, P. S. (2021). DeepBlue: Bi-layered LSTM for tweet popularity estimation. IEEE Transactions on Knowledge and Data Engineering, 34(10), 4737-4752.

Sun, L., Zhang, Z., Zhang, J., Wang, F., Du, Y., Su, S., & Yu, P. S. (2020, November). Perfect: A hyperbolic embedding for joint user and community alignment. In 2020 IEEE International Conference on Data Mining (ICDM) (pp. 501-510). IEEE.

Huang, Y., Roth, K., Bouniot, Q., Xu, W., & Akata, Z. (2026). Dissecting Multimodal In-Context Learning: Modality Asymmetries and Circuit Dynamics in modern Transformers. arXiv preprint arXiv:2601.20796.

Du, X., Chan, K. F., Xia, X., Chiu, P. W. Y., & Zhang, L. (2018, October). A magnetically-triggered soft capsule for on-demand mucus collection. In 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS) (pp. 250-254). IEEE.

Li, K., Chen, J., Yu, D., Dajun, T., Qiu, X., Lian, J., ... & Han, J. (2024, October). Deep reinforcement learning-based obstacle avoidance for robot movement in warehouse environments. In 2024 IEEE 6th International Conference on Civil Aviation Safety and Information Technology (ICCASIT) (pp. 342-348). IEEE.

Cao, C., Xue, X., Feng, L., & Tan, K. C. (2025, October). Why knowledge transfer works in oracle-based optimization: A brief analysis. In2025 International Conference on Machine Intelligence and Nature-Inspired Computing (MIND)(pp. 290-291). IEEE.

Xue, X., Hu, Y., Feng, L., Zhang, K., Song, L., & Tan, K. C. (2024). Surrogate-assisted search with competitive knowledge transfer for expensive optimization. IEEE Transactions on Evolutionary Computation.

Cao, C., Xue, X., Zhang, K., Song, L., Zhang, L., Yan, X.,... & Liu, C. (2024). Competitive knowledge transfer–enhanced surrogate-assisted search for production optimization. SPE Journal, 29 (06), 3277-3292.

Liang, R., Xu, S., Xie, C., Chen, J., Ren, F., Yang, S., & Yabe, T. (2026). Abstain Mask Retain Core: Time Series Prediction by Adaptive Masking Loss with Representation Consistency. Advances in Neural Information Processing Systems, 38, 99445-99471.

Hu, H., Yang, Y., Zhao, Q., & Zhang, C. (2022, June). On the role of discount factor in offline reinforcement learning. In International conference on machine learning (pp. 9072-9098). PMLR.

Lin, C. C., Li, G., Hung, C. C., Kuo, Y. C., & Peng, W. C. (2026). Rethinking suspicious account identification from a perspective of machine learning ranking. International Journal of Machine Learning and Cybernetics, 17(3), 84.

Xiaoteng Ma, Yiqin Yang, Hao Hu, Qihan Liu, Jun Yang, Chongjie Zhang, Qianchuan Zhao, and
Bin Liang. Offline reinforcement learning with value-based episodic memory. arXiv preprint
arXiv:2110.09796, 2021b.

Zhang, T., Xu, J., Chen, J., Li, J., Xiao, Y., & Wang, Q. (2026). NodeJEPA: Structure-Conditioned Latent Prediction for Node-Level Graph Self-Supervised Learning. arXiv preprint arXiv:2608.04381.

Lin, Y. (2024). Design of urban road fault detection system based on artificial neural network and deep learning. Frontiers in neuroscience, 18, 1369832.

Chai, L., Zhang, J., Li, H., Cui, C., Jiang, J., Zheng, B., ... & Jiang, L. (2021). Investigation of thermomorphogenesis-related genes for a multi-silique trait in Brassica napus by comparative transcriptome analysis. Frontiers in Genetics, 12, 678804.

Yu, A., Huang, Y., & Xia, L. (2023). Efficient characterization of phase modulator based on Lyot-Sagnac interferometer. Measurement, 210, 112538.

Xia, Y., & Ma, J. (2022). Locality-guided global-preserving optimization for robust feature matching. IEEE Transactions on Image Processing, 31, 5093-5108.

Xia, Y., & Ma, J. (2025). Grid-guided sparse Laplacian consensus for robust feature matching. IEEE Transactions on Image Processing, 34, 1367-1381.

Zhang, Y., Qiu, C., Liao, X., & Zhang, Y. (2025). Federated hierarchical clustering with automatic selection of optimal cluster numbers. Information Sciences, 122957.

Lyu, M., Torii, R., Liang, C., Peach, T. W., Bhogal, P., Makalanda, L., ... & Chen, D. (2024). Treatment for middle cerebral artery bifurcation aneurysms: in silico comparison of the novel Contour device and conventional flow-diverters. Biomechanics and Modeling in Mechanobiology, 23(4), 1149-1160.

Xing, Z. (2026, May). Design of High-Bandwidth DDR5 MRDIMM Memory Modules for AI Servers. In 2026 IEEE 6th International Conference on Electronic Technology, Communication and Information (ICETCI) (pp. 307-311). IEEE.

Zou, R., Zhang, Y., Zhao, M., Tan, Z., Zhang, Y., & Cheung, Y. M. (2025). SDENK: Unbiased Subspace Density--Clustering. Neurocomputing, 131225.

Wang, T., Deng, X., Sheng, S., & Jia, Y. (2024). Mitigating the damage of a global pandemic on the international buyer-supplier relationship: evidence from Chinese suppliers. Journal of Business Research, 172, 114446.

Cao, Y., Shi, D., Fu, X., Zou, X., Peng, H., Li, X., ... & Xing, J. (2026, March). Multivariate diffusion transformer with decoupled attention for high-fidelity mask-text collaborative facial generation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 4, pp. 2670-2679).

Meng, L., Du, Z., Jia, Z., & Cao, L. (2025). Spatial mechanisms and coupling coordination of cultural heritage and tourism along the Jinzhong segment of the Great Tea Road. Heritage, 9(1), 7.

Tian, Y. (2025). Exploration of power market dynamic scheduling algorithm based on reinforcement learning. In Image Processing, Electronics and Computers (pp. 836-845). IOS Press.

Shi, J., Lin, Y., Hua, Y., Wang, Z., Zhang, Z., Zheng, W., ... & Lu, S. (2026, March). Multiscenario highway lane-change intention prediction: a physics-informed AI framework for three-class classification. In International Conference on Smart Transportation and City Engineering (STCE 2025) (Vol. 14120, pp. 129-145). SPIE.

Chen, Y., Zhong, Y., Wang, Z., Shi, J., & Zheng, W. (2026). Applying the Deep Learning–Sector–Governance (DLSG) Framework to the US Healthcare System: Opportunities, Deployment Pathways, and Policy-Aligned Evaluation. Journal of Technology Innovation and Society, 4(1), 42-59.

Shi, J., Xie, Q., Wang, Z., Lin, Y., Zhu, D., Xie, C., ... & Guan, Z. (2026, June). Evolutionary Physics-Informed Temporal Fusion for Lane-Change Intention Prediction. In 2026 6th International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA) (pp. 750-764). IEEE.

Wang, Y. (2025, April). Efficient adverse event forecasting in clinical trials via transformer-augmented survival analysis. In Proceedings of the 2025 International Symposium on Bioinformatics and Computational Biology (pp. 92-97).

Wang, Y. (2025, June). RAGNet: Transformer–GNN–Enhanced Cox–Logistic Hybrid Model for Rheumatoid Arthritis Risk Prediction. In Proceedings of the 2025 International Conference on Health Informatization and Data Analytics (pp. 90-94).