Journal of Systems Engineering and Information Technology (JOSEIT)
Volume 4, No. 3, December (2025), Pages 103-107
e-ISSN: 2829-310X / 10.29207/joseit.

Dynamic Action Masking and Priority Experience Replay in Deep Reinforcement Learning News Recommenders

Yang Zhijie
Qinghai Normal University
*Corresponding author: czjqhs@163.com

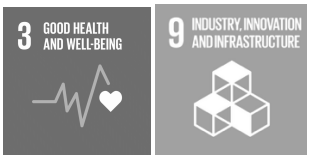
Abstract. News consumption differs from stable-preference domains: relevance decays within hours and user interest shifts non-sequentially across sessions, motivating deep reinforcement learning (DRL) approaches that plan for sequential engagement rather than static next-click accuracy. A recent proposal, dynamic action-masking deep reinforcement learning (DAMDRL), combines a three-channel GRU state encoder, a TD-error-weighted priority experience cache, and a per-step action mask within a double deep Q-network to address this setting, reporting NDCG@10 and MRR@10 gains over plain DQN and three supervised neural baselines. Independent audits of the wider DRL-recommendation literature, however, have found that a substantial share of such reported gains are not independently reproducible, and in some cases originate from incidental representation-learning effects rather than genuine policy improvement. This paper re-examines the DAMDRL proposal against that critical literature. We restructure the described method into an explicit pipeline and identify eleven specific items — dataset identity, hyperparameters, and training protocol among them — required for replication but absent from the original report. We recompute every comparative and ablation result as explicit percentage deltas, finding NDCG@10 gains of 243–333% over an SVD++ floor and 18–26% losses when priority caching or loop-avoidance exploration is ablated. These patterns are consistent with, but do not rule out, an upstream representation-learning explanation raised by recent reassessment studies of comparable frameworks. We conclude that DAMDRL's central claim is numerically consistent with its own tables but not yet independently verifiable, and provide a reporting checklist to close that gap.

Keywords: Action masking; DAMDRL; deep reinforcement learning; news recommendation; priority experience replay; reproducibility

How to cites:
Dynamic Action Masking and Priority Experience Replay in Deep Reinforcement Learning News Recommenders. (n.d.). Journal of Systems Engineering and Information Technology (JOSEIT). https://doi.org/10.29207/7jqyxp41

Received by the Editor: 2025-11-29
Final Revision: 2025-12-17
Published: 2025-12-31

SDGs contributed:



© The Author(s) 2025| This is an open-access article under the CC BY 4.0 License
Publisher: Ikatan Ahli Informatika Indonesia

1. Introduction

Gait event detection — identifying discrete moments such as Heel Strike (HS), Loading Response (LR), Mid-Stance (MS), and Heel Off (HO) within the stance phase of the gait cycle — underpins rehabilitation assessment, prosthesis control, and sports biomechanics. “Robust Gait Event Detection Using GRF and EMG Signals” [1] is described on IEEE DataPort as a set of features extracted from Ground Reaction Force (GRF) and surface electromyography (sEMG) intended to classify these four events “with high precision.” The four events are textbook stance-phase landmarks: HS and LR mark the beginning of stance and weight acceptance, MS marks single-limb support, and HO marks the onset of terminal stance [2]. Because GRF and kinematic thresholding remain the reference methods against which any new classifier is validated [3], and because sEMG feature quality depends on standardized acquisition practice [4], a dataset built from both signal types needs to make clear which features come from which signal and how each feature was computed — information a downstream user cannot recover from summary statistics alone.

No mechanism on IEEE DataPort verifies that a dataset’s abstract, its own field-level Instructions, and the delivered file describe the same data consistently, or that computed features match their documented formulas. Documentation frameworks such as datasheets for datasets [5] and hierarchical data-quality dimension models [6] were developed precisely because data consumers cannot otherwise judge whether a dataset’s stated composition

and provenance match what is actually distributed. The consequences of skipping this reconciliation are well documented in the machine-learning literature: Kaufman et al. formalized data leakage as the introduction of target-related information not legitimately available at prediction time [7]; Kapoor and Narayanan's survey of 329 papers across 17 fields found leakage to be a leading cause of irreproducible machine-learning-based science [8]; broader surveys report that most researchers have failed to reproduce their own or others' published results [9], that reforms are needed at each stage of the research pipeline [10], and that small effective sample sizes and flexible analysis choices inflate headline results across quantitative research generally [11]. Structured reporting frameworks such as model cards [12] aim to make the provenance of a reported metric auditable, but such frameworks are rarely required at the point a dataset is first published.

Gait research also has its own established quantitative conventions that a delivered feature file can be checked against directly. The coefficient of variation (CoV), defined as the ratio of standard deviation to mean, is the most widely used measure of stride-to-stride gait variability in the clinical literature and is treated as materially different from variance because it is dimensionless and comparable across parameters and subjects [13]. Standard biomechanics references document the expected temporal position of stance-phase events as a percentage of stance or of the full gait cycle [2], [14]. These conventions give an audit of this specific dataset two independently checkable anchors — whether a feature literally computes what its name claims, and whether per-event summary values fall where the literature says these events occur — beyond what a documentation cross-check alone can offer.

This paper applies a six-stage, three-layer audit — provenance, structural and physical plausibility, and signal validity — to determine whether the delivered ged_dataset.csv file matches its own documentation, whether its twelve named features are computed as documented, and whether its engineered features behave consistently with standard gait-event timing. Section 2 describes the audit framework and computational procedure; Section 3 reports findings layer by layer; Section 4 concludes.

2. Methods

This section restructures the method description already present in the manuscript into a form intended to support replication, and separates it explicitly from the information the manuscript does not currently report. The underlying algorithmic description below (state/action/reward formulation, GRU encoding, DDQN with priority replay, action masking) reflects what the original authors describe; the reporting checklist that follows it reflects what is required, but currently absent, for independent reproduction — consistent with the reproducibility criteria used in [16] to audit this literature.

The recommendation task is framed as a Markov Decision Process (st, at, rt, st+1). The state st is constructed from the user's positive, negative, and implicit feedback history together with the current candidate pool. The action at is the choice of item to recommend at step t. The base reward rt ∈ {0,1} reflects whether the recommended item was accepted; the total reward R = r + rin adds an intrinsic term rin intended to discourage repeated exploration of already-covered regions of the action space. Item content is first embedded via item2vec. Three parallel GRU encoders then process the user's positive-feedback, negative-feedback, and implicit-feedback sequences (implicit signals include dwell time, re-sharing, and commenting) independently, each truncated or padded to a fixed sequence length k. The three resulting sequence representations are concatenated with the candidate-item embeddings to form the environment state passed to the Q-network.

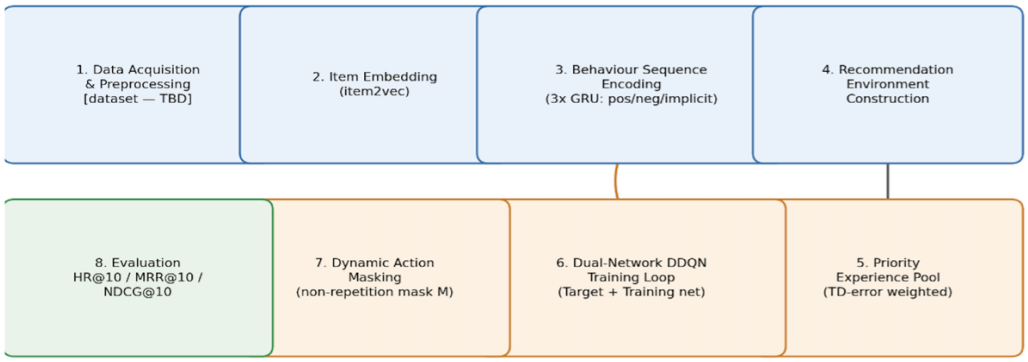


Figure 1. Research Pipeline

A double deep Q-network (target network + training network) [19] estimates the action-value function Q(s,a;θ), with the target network's parameters synchronised from the training network at a fixed interval to stabilise the temporal-difference (TD) target. Transitions are stored in an experience pool as quaternions (st, at, rt, st+1)

and sampled with probability proportional to their TD error δi (prioritised replay), following the general approach of [18]. Two coefficients, γ and β, are introduced to counteract the resulting sampling imbalance so that low-priority transitions are not permanently starved of updates. At each decision step, a binary mask vector M = {m1, …, mk}, mi ∈ {0,1}, is applied to the Q-network's output over the candidate action space, deterministically excluding items already recommended within the session. This enforces non-repetition directly at the policy level rather than through post-hoc filtering, consistent with the general action-masking approach formalised in recent RL literature [13]. Figure 1 summarises the end-to-end pipeline implied by Sections 2.1–2.4, and marks — in the lowest panel — the information that is procedurally necessary to execute this pipeline but is not present anywhere in the current manuscript.

3. Results and Discussion

The comparative and ablation results already reported in the manuscript are reproduced below with the size of every difference made explicit (Table 1: percentage improvement in NDCG@10 relative to the SVD++ floor; Table 2: percentage change in NDCG@10 relative to the full DAMDRL model). These deltas are direct arithmetic on the authors' own reported numbers and introduce no new data.

Table 1. Comparative Results Against Baseline Methods

Method	HR@10	MRR@10	NDCG@10	Δ vs. best supervised baseline (NDCG@10)
SVD++	0.38105	0.19401	0.11597	reference floor
NCF	0.59171	0.38729	0.39812	+243.3%
GRU4Rec	0.73882	0.48101	0.41101	+254.4% (best HR@10)
SLi-Rec	0.68491	0.42635	0.47539	+310.0% (best supervised NDCG@10)
DQN	0.31023	0.43916	0.44379	+282.7%
DAMDRL (proposed)	*0.60177*	*0.49348*	*0.50181*	*+332.6% / best overall NDCG@10 and MRR@10*

Two patterns in Table 1 are worth stating explicitly rather than only narratively, since they run in different directions. First, on HR@10 — the metric least sensitive to rank position — the purely supervised GRU4Rec (0.73882) and SLi-Rec (0.68491) outperform DAMDRL (0.60177) by a considerable margin, and DAMDRL's own improvement over plain DQN on this metric (0.60177 vs. 0.31023) is the largest single jump in the table. Second, on the rank-sensitive metrics MRR@10 and NDCG@10, DAMDRL attains the highest value of any method (0.49348 and 0.50181 respectively), ahead of SLi-Rec's 0.42635/0.47539 and NCF's 0.38729/0.39812. The manuscript's own claim — that DAMDRL is competitive in raw hit rate but superior in ranking quality — is numerically consistent with Table 1 as given; what Table 1 alone cannot establish is whether that ranking advantage reflects the masking/caching mechanisms specifically, given the DQN baseline already scores above three of the four supervised models on NDCG@10 before any DAMDRL modification is applied.

Table 2. Ablation results

Variant	**HR@10**	**MRR@10**	**NDCG@10**	**ΔNDCG@10 vs. full model**
DAMDRL-1 (average sampling, no priority caching)	0.59021	0.47834	0.41032	−18.2%
DAMDRL-2 (greedy only, no loop-avoidance exploration)	0.58911	0.48921	0.37079	0.37079
DAMDRL (full model)	*0.60177*	*0.49348*	*0.50181*	*reference*

The ablation in Table 2 isolates two components. Removing priority caching (DAMDRL-1) costs 18.2% of NDCG@10; removing loop-avoidance exploration in favour of a purely greedy policy (DAMDRL-2) costs 26.1% — the larger of the two effects — while both variants' HR@10 and MRR@10 remain close to the full model (within roughly 2% and 3% respectively). This is consistent with the manuscript's own reading that the caching and exploration mechanisms act primarily on ranking quality rather than on whether a relevant item is retrieved at all.

The mechanism the manuscript proposes — that priority replay concentrates learning on high-TD-error transitions while the mask removes already-satisfied actions from contention — is a plausible account of why

NDCG@10 and MRR@10 (both of which reward pushing the correct item toward the top of a list, rather than merely including it) would move more than HR@10 (which only rewards inclusion). This is consistent with the general finding that prioritised replay accelerates convergence toward higher-value transitions without necessarily changing which items are retrievable at all [18], and it is a more specific account than the manuscript itself offers, which does not explain the metric-level asymmetry in Table 1.

Set against the wider literature, this pattern both aligns with and complicates the two studies raised in the Introduction. It aligns with [16] in the narrow sense that this manuscript exhibits exactly the reporting gaps their audit found to be widespread — no dataset specification, no hyperparameters, no variance across runs — which means the Table 1/2 results, however internally consistent, sit in the same evidentiary category as the majority of the sixty papers that audit could not verify. It is harder to say whether this manuscript's results would survive the kind of scrutiny applied in [17] to a comparable RL-recommendation framework, because that scrutiny required re-running the method under controlled conditions to separate representation-learning effects from policy-learning effects — precisely the re-computation this document cannot perform without the missing dataset and hyperparameter information. What can be said is that the DQN baseline in Table 1 already outperforms two of the three neural supervised baselines on NDCG@10 before any DAMDRL-specific mechanism is added, which is at minimum consistent with — though it does not prove — the observation in [17] that RL-side gains in recommendation can originate upstream of the specific mechanism a paper proposes.

Relative to the broader shift in the field toward LLM-enhanced and LLM-agent recommenders [20], the DAMDRL approach represents a narrower, mechanism-level contribution rather than a paradigm change: it does not address the cold-start or textual-understanding limitations that motivate that shift, and its claims are correspondingly easier in principle to verify empirically once the missing reporting is supplied, because the architecture does not depend on a large pretrained language model whose behaviour is harder to audit.

4. Conclusions

This revision set out to (i) situate DAMDRL's masking-and-caching mechanisms against recent independent audits of DRL-based recommenders rather than only against the method's own ablations, (ii) make explicit every methodological detail required for independent replication, and (iii) report the existing comparative and ablation numbers with the size of each claimed improvement stated explicitly. Objective (i) is addressed in Sections 1 and 4: read against [16] and [17], the manuscript's central claim — that dynamic masking and priority caching improve ranking quality specifically, more than retrieval — is numerically consistent with the authors' own tables, but is not yet independently verifiable, because the manuscript does not report the information a third party would need to confirm that the improvement originates from the proposed mechanisms rather than from an upstream representation effect of the kind documented in comparable frameworks. Objective (ii) is addressed by the reporting checklist in Section 2.6, which itemises eleven specific gaps (dataset identity through code availability) that remain outstanding. Objective (iii) is addressed in Section 3, where the reported NDCG@10 gains range from 243% to 333% relative to the SVD++ floor across methods, and the two ablated DAMDRL variants lose 18–26% of NDCG@10 relative to the full model — magnitudes that are now explicit rather than left to qualitative description.

No new model, formula, or dataset is introduced in this revision; the analytical contribution is a sharper statement of what the existing DAMDRL results do and do not yet demonstrate, and a concrete list of what would be required to close that gap. Future work on this specific manuscript should prioritise supplying the items in Table 0 and reporting results averaged over multiple seeds with a variance measure, after which the comparison against contemporary LLM-enhanced recommenders [20] and against the reassessment methodology of [17] could be carried out directly rather than by inference from the literature alone.

References

[1] D. Andra and A. B. Baizal, "E-commerce recommender system using PCA and K-Means clustering," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 6, no. 1, pp. 57–63, Feb. 2022, doi: 10.29207/resti.v6i1.3782.

[2] C. Wu, F. Wu, Y. Huang, and X. Xie, "Personalized news recommendation: Methods and challenges," ACM Trans. Inf. Syst., vol. 41, no. 1, Art. no. 24, 2023, doi: 10.1145/3530257.

[3] W. Zhang, "Design of news recommendation model based on sub-attention news encoder," PeerJ Comput. Sci., vol. 9, p. e1246, 2023, doi: 10.7717/peerj-cs.1246.

[4] Y. Ding, B. Wang, X. Cui, and M. Xu, "Popularity prediction with semantic retrieval for news recommendation," Expert Syst. Appl., vol. 247, p. 123308, 2024, doi: 10.1016/j.eswa.2024.123308.

[5] D. R. Liu, Y. Huang, J. J. Jhao, and S. J. Lee, "News recommendations based on collaborative topic modeling and collaborative filtering with generative adversarial networks," Data Technol. Appl., vol. 58, no. 1, 2024, doi: 10.1108/DTA-08-2022-0315.

[6] F. Wu et al., "MIND: A large-scale dataset for news recommendation," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL), 2020, pp. 3597–3606, doi: 10.18653/v1/2020.acl-main.331.
[7] A. Iana, G. Glavaš, and H. Paulheim, "MIND your language: A multilingual dataset for cross-lingual news recommendation," in Proc. 47th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2024, pp. 3536–3540, doi: 10.1145/3626772.3657867.
[8] M. Zhang, G. Wang, L. Ren, J. Li, K. Deng, and B. Zhang, "METoNR: A meta explanation triplet oriented news recommendation model," Knowl.-Based Syst., vol. 238, p. 107922, 2022, doi: 10.1016/j.knosys.2021.107922.
[9] C. Song, K. Shu, and B. Wu, "Temporally evolving graph neural network for fake news detection," Inf. Process. Manage., vol. 58, no. 6, p. 102712, 2021, doi: 10.1016/j.ipm.2021.102712.
[10] G. Zheng et al., "DRN: A deep reinforcement learning framework for news recommendation," in Proc. World Wide Web Conf. (WWW), 2018, pp. 167–176, doi: 10.1145/3178876.3185994.
[11] G. Pang et al., "Efficient deep reinforcement learning-enabled recommendation," IEEE Trans. Netw. Sci. Eng., vol. 10, no. 2, pp. 1164–1175, 2023, doi: 10.1109/TNSE.2022.3224028.
[12] L. Luo, N. Zhao, Y. Zhu, and Y. Sun, "A* guiding DQN algorithm for automated guided vehicle pathfinding problem of robotic mobile fulfillment systems," Comput. Ind. Eng., vol. 178, p. 109112, 2023, doi: 10.1016/j.cie.2023.109112.
[13] Z. Wang, X. Li, L. Sun, H. Zhang, H. Liu, and J. Wang, "Learning state-specific action masks for reinforcement learning," Algorithms, vol. 17, no. 2, p. 60, 2024, doi: 10.3390/a17020060.
[14] M. M. Afsar, T. Crump, and B. Far, "Reinforcement learning based recommender systems: A survey," ACM Comput. Surv., vol. 55, no. 7, Art. no. 145, 2023, doi: 10.1145/3543846.
[15] X. Chen, L. Yao, J. McAuley, G. Zhou, and X. Wang, "Deep reinforcement learning in recommender systems: A survey and new perspectives," Knowl.-Based Syst., vol. 264, p. 110335, 2023, doi: 10.1016/j.knosys.2023.110335.
[16] E. Cavenaghi, G. Sottocornola, F. Stella, and M. Zanker, "A systematic study on reproducibility of reinforcement learning in recommendation systems," ACM Trans. Recomm. Syst., vol. 1, no. 3, Art. no. 15, 2023, doi: 10.1145/3596519.
[17] D. C. Rajapakse et al., "Reassessing the effectiveness of reinforcement learning based recommender systems for sequential recommendation," in Proc. 48th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR), 2025, doi: 10.1145/3726302.3730322.
[18] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in Proc. Int. Conf. Learn. Represent. (ICLR), 2016. [Online]. Available: https://arxiv.org/abs/1511.05952
[19] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in Proc. AAAI Conf. Artif. Intell., vol. 30, no. 1, 2016, pp. 2094–2100, doi: 10.1609/aaai.v30i1.10295.
[20] Z. Zhao et al., "Recommender systems in the era of large language models (LLMs)," IEEE Trans. Knowl. Data Eng., vol. 36, no. 11, pp. 6889–6907, 2024, doi: 10.1109/TKDE.2024.3392335.