CoFARL: Copula-Conditioned Feature Aggregation for Sequential Recommendation with Deep Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

Sequential recommendation requires models to combine heterogeneous information while balancing ranking quality and cumulative reward. We introduce CoFARL, a framework for copula-conditioned feature aggregation with deep reinforcement learning. A causal encoder summarizes interaction history, while a copula dependence descriptor conditions the weighting of candidate feature groups. The resulting representation supports temporal ranking, ranking pretraining and continued ranking supervision during sequential decision learning. Our theoretical analysis characterizes fixed-descriptor parameterization, the decision value of dependence information and ranking stability under descriptor perturbations. Experiments span MovieLens recommendation, the RL4RS native simulator and expanding-window cryptocurrency asset selection. In a matched twenty-seed MovieLens study, aligned copula conditioning improves ranking over attention and all descriptor controls, with reward gains over attention, permuted copula and sequence-only models surviving multiplicity correction. Comparisons with side-information fusion methods demonstrate higher simulated reward while retaining most of the strongest comparator's mean ranking quality. RL4RS and financial experiments further characterize the framework across recommendation environments and chronological out-of-sample decisions. Together, these results support dependence-conditioned aggregation as a useful component of sequential recommendation systems.

1 Introduction

Sequential recommendation combines next-item relevance with decisions that produce useful interactions over time. Ranking quality and cumulative reward describe complementary objectives: optimizing one need not improve the other.

CoFARL connects temporal ranking to sequential control through copula-conditioned feature aggregation. A causal encoder summarizes interaction history. A training-fitted copula describes dependence among feature-evidence summaries after marginal normalization, and a learned gate combines this descriptor with candidate and history embeddings. The resulting scores support ranking supervision and learner-specific policy and value representations.

Our contributions are threefold. First, we specify this architecture and its integration with deep RL. Second, we characterize fixed-descriptor equivalence, dependence-based decision distinctions, ranking stability and conditional value-to-control guarantees. Third, we evaluate ranking and sequential utility through matched aggregation controls and comparisons with recommendation systems, complemented by cryptocurrency and native-simulator evaluations.

MovieLens provides the primary study of learner choice, aggregation and continued ranking supervision. Its 20-seed TD3 comparison isolates attention, identity, raw-correlation and permuted-descriptor alternatives. RL4RS evaluates the same central representation question with PPO and native action masks; the financial extension evaluates expanding-window out-of-sample Sharpe. We report mean operating points and uncertainty with each study's data access, action interface and metric scope explicitly distinguished.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2 Related Work

Sequential recommendation and fusion. SASRec learns causal attention over interaction histories (Kang & McAuley, 2018); BERT4Rec uses masked-item prediction (Sun et al., 2019), and S^3-Rec combines attribute and sequence supervision (Zhou et al., 2020). DIF-SR separates item and attribute attention (Xie et al., 2022), while TASIF integrates temporal information and adaptive fusion (Luo et al., 2025). CoFARL conditions feature-group weights on a dependence descriptor alongside history and candidate embeddings. Attention, identity and raw-correlation controls distinguish this input from ordinary fusion and group-specific offsets.

Recommendation and sequential control. Xin et al. (2020) combine recommendation supervision with Q-learning and actor–critic heads. Their Self-Supervised Actor-Critic differs from Soft Actor-Critic (Haarnoja et al., 2018). Next-item accuracy and long-term reward can favour different policies (Silva et al., 2024), motivating their joint evaluation. SlateQ (Ie et al., 2019) and generative slate representations (Deffayet et al., 2023) study slate-level decisions. CoFARL connects a temporally informed item scorer to learner-specific action and value heads.

Dependence and representations. Gaussian copulas have been used to sample dependent feature selections (Peng et al., 2023), model inter-agent actions (Wang et al., 2021a), and combine recommendation relevance and diversity (Coppolillo et al., 2024). Here they describe feature-evidence dependence before aggregation. Actor–critic representation studies (Pinto et al., 2018; Garcin et al., 2025; Baisero & Amato, 2022) motivate separating representation and learning effects. Our fixed-descriptor analysis and matched controls make this distinction explicit. Descriptor interventions evaluate policy sensitivity directly, complementing attention diagnostics (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Appendix F discusses conditional computation, structured state representations, decision-focused estimation and non-stationarity.

3 Sequence-Conditioned Feature Aggregation

At a prediction time t, H_t contains at most 20 strictly preceding positive interactions. A two-layer causal Transformer with two heads and width 64 produces $q_t = f_\psi(H_t)$ from item and valid-position embeddings. Padding is masked and a beginning-of-sequence token handles empty histories. Candidate features $x_{t,i,g}$ form three groups: content, history affinity and training-popularity context. LN_g denotes layer normalization. Group embeddings are $e_{t,i,g} = W_g\,\mathrm{LN}_g(x_{t,i,g}) + b_g$.

Dependence estimation. Let v_i be normalized item metadata, c_i its training positive count, and $w_{t,r}$ normalized exponential recency weights. For L_t valid history events, $\overline{v}_{H_t} = L_t^{-1}\sum_{r=1}^{L_t} v_{i_r}$, with the zero vector for an empty history. For $L_t > 0$, $w_{t,r} = \exp[-0.15(L_t - r)]/\sum_{j=1}^{L_t}\exp[-0.15(L_t - j)]$; empty-history repeat evidence is zero. Recency counts valid event positions rather than elapsed time. The scalar evidence is

$$y_{t,i} = \left(v_i^T \overline{v}_{H_t},\ \log(1 + c_i),\ \sum_r w_{t,r}\mathbf{1}\{i_r = i\} \right). \tag{1}$$

The N training reference rows pair pre-event histories with their observed positive targets. Frozen empirical mid-CDFs give $u_{n,g} = \mathrm{clip}(\widehat{F}_g(y_{n,g}), 1/(2N), 1 - 1/(2N))$ and normal scores $z_{n,g} = \Phi^{-1}(u_{n,g})$. Constant coordinates map to $u = 1/2$. We use $R = 0.9\,\mathrm{Corr}(z) + 0.1I$, with numerical positive-definiteness safeguards. The Gaussian copula is $C_R(u) = \Phi_R(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_3))$. For discrete or tied margins, this construction defines a latent Gaussian dependence convention. The aggregation module uses R as its dependence input.

Gate and score. The archived experiments use evidence in affinity/popularity/repeat order and aggregation groups in content/history/popularity order. Their positional descriptor mapping evaluates fixed conditioning; a semantically aligned group-dependence mechanism requires separately

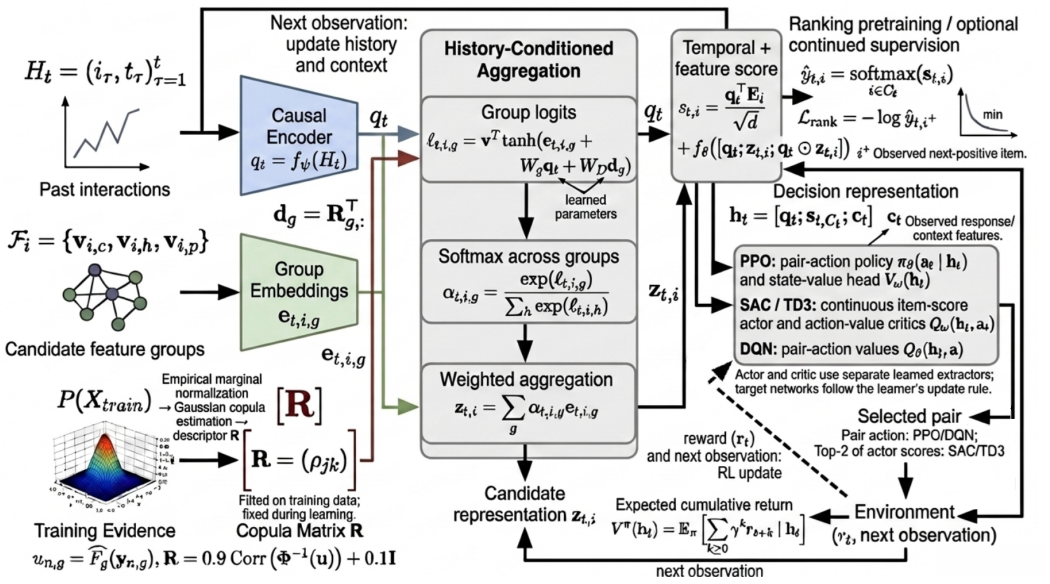

Figure 1: CoFARL architecture. Ranking probabilities are the softmax of candidate scores. A causal history encoder, candidate feature embeddings and a training-fitted copula descriptor determine history-conditioned aggregation and item scores. Ranking supervision alternates with native RL updates through the online actor scorer, or the online Q scorer for DQN. PPO and DQN select candidate pairs; SAC and TD3 map continuous scores to a Top-K slate ($K = 2$ in the current experiments). Separate value and target branches follow each learner's update rule; c_t contains observed decision context.

versioned experiments (Appendix B). Writing $d_g = R^T_{g,:}$, the gate and score are

$$\ell_{t,i,g} = a^T \tanh(e_{t,i,g} + W_q q_t + W_D d_g),$$
$$\alpha_{t,i,g} = \mathrm{softmax}_g(\ell_{t,i,g}), \qquad z_{t,i} = \sum_g \alpha_{t,i,g} e_{t,i,g}, \tag{2}$$

$$s_{t,i} = q_t^T E_i / \sqrt{d} + f_\theta([q_t; z_{t,i}; q_t \odot z_{t,i}]). \tag{3}$$

Here $a \in \mathbb{R}^{64}$ is the learned gate projection, $E_i \in \mathbb{R}^{64}$ is the item embedding shared with the temporal encoder, $d = 64$, and $\odot$ denotes elementwise multiplication. The feature network maps $192 \to 64 \to 1$, with GELU between its two affine layers. Gates vary with history and candidate features; R remains fixed within each fitted MovieLens model. The history and candidate inputs allow the gate to change across decisions with this fixed descriptor.

Learning and matched controls. The supervised objective is

$$\mathcal{L}_{\mathrm{rank}} = -\mathbb{E} \log \frac{\exp(s_{t,i^+})}{\sum_{j \in C_t} \exp(s_{t,j})}. \tag{4}$$

For ranking supervision, C_t contains the observed target and sampled comparison items; the RL auxiliary updates use 32 candidates. This sampled-candidate classification objective treats unobserved items as comparison alternatives. Appendix B specifies its sampling and evaluation conventions. The reported internal controls are attention, which removes the dependence term, and identity, which replaces R with I. Identity and copula retain the same descriptor projection and parameter count.

Reinforcement learning and temporal integration. Each RL branch constructs $h_t = [q_t; s_{t,C_t}; c_t]$, where c_t contains observed response-history summaries and the episode clock. PPO maps h_t to a distribution over candidate pairs and uses a separate temporal value representation. SAC and TD3 output actions in $[-1, 1]^{32}$. Their critics evaluate this continuous score vector directly, while the environment deploys its two highest-scoring candidates; the action-to-slate map is

discrete and many-to-one. They fit separate temporal action-value networks. DQN maps the representation to pair action values. The four learners use their native policy/value updates (Schulman et al., 2017; Haarnoja et al., 2018; Fujimoto et al., 2018), with gradients reaching the history encoder and dependence projection.

Ranking pretraining initializes the scorer in the online branches and target copies before RL. Continued next-item supervision then alternates with RL updates through the online actor optimizer, or the online Q optimizer for DQN. This implements a multi-objective training procedure with ranking weight λ; the auxiliary steps update the scorer while each learner maintains its own value targets. The internal MovieLens comparison uses ranking pretraining followed by RL, with or without continued ranking supervision. The external CoFARL configurations use continued supervision. A score-initialized action head preserves the pretrained scorer's Top-2 choice before RL begins. Appendix B specifies the MovieLens training and evaluation protocol.

4 Dependence, Representation and Decision Error

We first characterize the effect of a fixed descriptor on model capacity. We then consider dependence information omitted by a compressed state and derive bounds relating descriptor and value errors to decisions.

Proposition 1 (Fixed-descriptor equivalence)**.** *For fixed symmetric invertible R and unrestricted W_D, the gates in equation 2 with descriptor R and descriptor I have identical function classes, with all other parameters unchanged.*

The proof is in Appendix A.1. The equivalence holds for any temporal encoder. Performance differences between the two parameterizations can therefore arise from optimization or regularization, despite equal expressive capacity. This result motivates the identity control in the MovieLens study.

Proposition 2 (Dependence can distinguish aliased decisions)**.** *Consider equally likely regimes $b \in \{-1, +1\}$ with identical supplied base representation. Conditional on b, future (U, V) are jointly bivariate normal with zero mean, unit variances and correlation $b\rho$, where $0 < \rho < 1$. An action $a \in \{-1, +1\}$ receives $r = \mathbf{1}\{aUV > 0\}$. A policy that observes only the base representation attains expected reward $1/2$. A policy additionally observing the regime's true copula correlation attains $1/2 + \arcsin(\rho)/\pi$.*

The proof is in Appendix A.2. The example assumes that the base representation loses regime information that the descriptor retains. For a base model with access to the full history determining R_t, the descriptor provides an explicit summary of existing information. The proposition quantifies its value when the supplied state compresses away that distinction. If a regime estimate has symmetric error probability η and that error is independent of future outcomes conditional on regime, the gain becomes $(1 - 2\eta) \arcsin(\rho)/\pi$.

Proposition 3 (Descriptor error and ranking stability)**.** *Fix a decision time t, history representation q_t and candidate group embeddings, with $\|e_{t,i,g}\|_2 \le B$. Suppress t in the scores below. Suppose $g_{q_t}(z) = f_\theta([q_t; z; q_t \odot z])$ is $L_f(q_t)$-Lipschitz in z and $\|\widehat{R} - R\|_F \le \delta$. Then*

$$\max_i |s_i(\widehat{R}) - s_i(R)| \le \varepsilon_s := 2BL_f(q_t)\|a\|_2\|W_D\|_2\delta. \tag{5}$$

If the reference score gap between ranks K and $K + 1$ exceeds $2\varepsilon_s$, the Top-K set is unchanged. Maximizing the perturbed additive score sum incurs reference-score regret at most $2K\varepsilon_s$.

The proof is in Appendix A.3. The Lipschitz constant is conditional on the fixed history. If f_θ is L_θ-Lipschitz in its concatenated input, then $L_f(q_t) \le L_\theta\sqrt{1 + \|q_t\|_\infty^2}$. A bound uniform over histories additionally requires a uniform bound on their norms. The bound holds for a fixed scorer under descriptor perturbation. Top-K set preservation protects hit/recall membership; it does not preserve NDCG or MAP if items reorder inside the set. Preserving all relevant pairwise orderings requires corresponding score margins. The regret is measured in the scorer's units; interpreting it as user utility requires utility-calibrated scores.

Proposition 4 (Conditional value-to-control guarantee)**.** *In a discounted fully observed MDP with bounded rewards and $\gamma < 1$, suppose $\|\widehat{Q} - Q^*\|_\infty \le \varepsilon_Q$. A policy greedy with respect to $\widehat{Q}$ satisfies $\|V^* - V^{\widehat{\pi}}\|_\infty \le 2\varepsilon_Q/(1 - \gamma)$. If instead $\|\widehat{Q} - \mathcal{T}^*\widehat{Q}\|_\infty \le \beta$, the bound is $2\beta/(1 - \gamma)^2$.*

The proof is in Appendix A.4. This standard approximate-greedy bound connects uniform value accuracy to policy quality under a Markov state. It provides the control-side motivation for evaluating how dependence-conditioned representations affect value estimation. The MovieLens study evaluates ranking and simulated return separately; the uniform value-error assumptions of this guarantee are theoretical conditions, not measured properties of the fitted agents.

Implications for evaluation. The results motivate separate measurements of descriptor error, score margins, selected actions and value error. An improvement in dependence estimation can leave recommendations unchanged when it is too small to alter score orderings. Likewise, an improvement in ranking accuracy can coexist with unchanged sequential return. These distinctions guide the MovieLens evaluation of ranking and return. The reported experiments characterize operating points; the bounds supply conditions under which representation accuracy would preserve decisions.

5 EXPERIMENTS

5.1 MOVIELENS: RANKING AND SEQUENTIAL REWARD

We study how CoFARL combines held-out ranking quality with cumulative reward in sequential recommendation. The external comparison evaluates the complete architecture with four RL learners. A separate internal study compares copula conditioning with attention and an identity descriptor, and examines continued ranking supervision.

Evaluation protocol. The external comparison uses the same 128 evaluation users and three training seeds per method. We evaluate ranking and simulated return from the same selected checkpoint. CoFARL and the selected external methods, SASRec, SQN and Self-Supervised Actor-Critic, train on filtered logged transitions. The reported external results use no additional simulator-training interactions; the simulator is used to evaluate sequential return. Matched internal controls address the contribution of aggregation separately. Validation NDCG@10 selects the CoFARL checkpoint from the evaluated training checkpoints.

Metrics. Sequential utility is the mean cumulative reward over a 20-step simulated interaction. Full-catalog ranking evaluates each method's native scorer against the held-out next-positive item. For target rank r_u, Recall@K averages $\mathbf{1}\{r_u \le K\}$, NDCG@K averages $\mathbf{1}\{r_u \le K\}/\log_2(r_u + 1)$, and MAP@$K$ averages $\mathbf{1}\{r_u \le K\}/r_u$. With one relevant item, MAP is truncated reciprocal rank. Table 1 reports $K = 10$; the experiment archive also contains $K = 5, 20$ and deployed Top-2 recall. Native-scorer ranking and deployed action relevance are distinct endpoints.

Table 1: MovieLens external comparison: mean $\pm$ seed standard deviation over three seeds on the same 128 users. Bold denotes the highest mean among the displayed methods. Reward is simulated 20-step return. All ranking columns are scaled by $1{,}000$.

Method	Reward ↑	NDCG@10 ↑	Recall@10 ↑	MAP@10 ↑
SASRec	16.478 ± 0.361	$\mathbf{11.250 \pm 2.360}$	$\mathbf{26.040 \pm 4.510}$	6.690 ± 1.940
SQN	17.361 ± 0.427	10.950 ± 3.890	23.440 ± 7.810	$\mathbf{7.300 \pm 3.100}$
Self-Supervised AC	17.342 ± 0.062	8.720 ± 4.950	18.230 ± 11.930	5.870 ± 4.120
CoFARL–SAC	17.173 ± 0.123	5.550 ± 4.830	7.810 ± 7.810	4.770 ± 4.190
CoFARL–PPO	17.156 ± 0.157	5.620 ± 6.640	10.420 ± 11.930	4.200 ± 5.180
CoFARL–TD3	$\mathbf{17.487 \pm 0.177}$	6.580 ± 3.110	15.620 ± 0.000	4.100 ± 3.970
CoFARL–DQN	17.053 ± 0.556	6.790 ± 6.730	13.020 ± 9.020	4.880 ± 5.940

External reward–ranking trade-off. CoFARL–TD3 achieves the highest mean return among the displayed methods, 17.487, corresponding to gains of 6.12% over SASRec, 0.73% over SQN and 0.83% over Self-Supervised Actor-Critic. Its NDCG@10 ($\times 1{,}000$) is 6.580, retaining 58.4%, 60.1% and 75.4% of these methods' respective ranking scores. These results locate CoFARL–TD3 at a reward-oriented operating point. The observed mean frontier among the displayed methods comprises SASRec, SQN and CoFARL–TD3: each attains a different balance of ranking quality and return. This frontier describes measured configurations, rather than a continuous set of achievable policies.

Table 2: Internal MovieLens TD3 comparison (separate protocol). All models use ranking pretraining. Continued supervision adds ranking loss during RL. NDCG@10 is scaled by 1,000. Values are mean ± SD over three seeds; bold marks the highest mean.

Aggregation	Continued supervision	Reward ↑	NDCG@10 ↑
Attention	No	17.210 ± 0.466	13.704 ± 1.055
Attention	Yes	17.475 ± 0.190	17.675 ± 1.582
Copula	No	$\mathbf{17.482 \pm 0.011}$	13.133 ± 0.475
Copula	Yes	17.388 ± 0.166	$\mathbf{17.789 \pm 0.840}$
Identity	No	17.480 ± 0.136	13.006 ± 1.276
Identity	Yes	17.216 ± 0.170	14.878 ± 1.680

Table 3: MovieLens TD3 aggregation comparison over 20 seeds. Values are mean ± sample standard deviation. Bold marks the highest mean in each column. NDCG@10 is multiplied by 100.

Aggregation	Simulated reward ↑	NDCG@10 ×100 ↑
Aligned copula	$\mathbf{17.495 \pm 0.247}$	$\mathbf{1.451 \pm 0.124}$
Attention	17.295 ± 0.241	0.912 ± 0.145
Identity	17.388 ± 0.230	0.928 ± 0.202
Raw correlation	17.369 ± 0.261	0.883 ± 0.166
Permuted copula	17.304 ± 0.241	0.933 ± 0.200
Sequence-only	17.135 ± 0.400	0.592 ± 0.149

Internal aggregation controls. The internal MovieLens study contains 72 fits: four learners, three aggregation variants, two supervision settings and three seeds. Its training-data protocol differs from the external comparison, so its results are analyzed separately. Table 2 examines the reward–ranking balance under TD3 across aggregation and supervision settings. Appendix C reports all 24 configurations.

Continued ranking supervision. For copula–TD3, continued ranking supervision raises NDCG@10 (×1,000) from 13.133 to 17.789 (+35.45%), while mean return changes from 17.482 to 17.388 (−0.54%; Table 2). This comparison demonstrates how the training objective changes the ranking–reward balance within CoFARL. Attention–TD3 with continued supervision attains NDCG@10 (×1,000) of 17.675 and return of 17.475, illustrating the importance of the matched aggregation control when attributing this effect to dependence conditioning.

Uncertainty. Standard deviations describe variation across three training seeds. The existing paired seed-level tests retain their original experiment-wide Holm corrections; adjusted p-values are 1 for the reported comparison families. We therefore interpret the mean differences and frontier descriptively. These tests establish neither superiority nor equivalence.

5.2 Twenty-seed TD3 aggregation comparison

Table 3 compares aligned copula conditioning with attention, identity conditioning, raw correlation, permuted copula conditioning and a sequence-only model over 20 training seeds. These controls examine feature weighting, group-specific offsets, dependence estimation, coordinate alignment and heterogeneous information, respectively.

Sequential utility. Aligned copula achieves the highest mean return, 17.495 ± 0.247. Paired comparisons across 20 seeds show reward gains over attention of 0.200 (95% CI $[0.055, 0.345]$, Holm-adjusted $p = 0.032$), permuted copula of 0.191 ($[0.046, 0.336]$, $p = 0.041$), and sequence-only of 0.360 ($[0.154, 0.566]$, $p = 0.008$). Mean gains over identity and raw correlation are 0.107 and 0.126, with adjusted $p = 0.263$ and 0.205. Table 9 reports all paired intervals and tests, correcting across ten contrasts: five controls and two endpoints.

Joint reward–ranking performance. Under the matched TD3 protocol, aligned copula improves ranking over all five controls (Holm-adjusted $p < 0.001$ each), with concurrent reward improve-

ments over attention, permuted copula and sequence-only surviving correction. Aligned copula also achieves the highest mean NDCG@10, 1.451 ± 0.124 on the $\times 100$ scale. Using unrounded scores, its mean ranking quality exceeds attention by 59.0%, identity by 56.3%, raw correlation by 64.3%, permuted copula by 55.4% and sequence-only modelling by 145.2%. Aligned copula therefore improves both evaluated means over every displayed control, placing it on the empirical reward–ranking frontier as the sole nondominated configuration in this comparison. This frontier describes the observed means of the evaluated configurations.

Isolating dependence conditioning. The controls address distinct explanations for the observed operating point. Attention tests feature weighting without an explicit dependence descriptor; identity retains learned group-specific offsets; raw correlation supplies linear dependence summaries; and permutation tests the correspondence between descriptor coordinates and aggregation groups. The aligned-versus-permuted comparison supports the practical importance of that correspondence, while the sequence-only comparison examines the contribution of heterogeneous information. These questions are evaluated within the same TD3 protocol, with 20 training seeds characterizing optimization variability on the fixed evaluation cohort.

5.3 Sequential reward with retained ranking quality

We compare the 20-seed aligned-copula TD3 model from Table 3 with three fusion adaptations (Table 4). CoFARL achieves the highest mean reward, 17.495 ± 0.247, exceeding DIF-SR by 3.313% and DIFF by 4.804%. Both reward gains have conservative multiplicity-adjusted bounds below 0.001 across six comparisons. Its NDCG@10 retains 97.082% of DIFF's mean ranking score and exceeds DIF-SR and TASIF, demonstrating a reward-oriented operating point with substantial ranking retention.

Table 4: MovieLens side-information fusion comparison: mean $\pm$ sample SD across 20 seeds. Bold identifies the highest mean in each column. Ranking is scaled by 100; the CoFARL row is shared with Table 3.

Method	Simulated reward ↑	NDCG@10 $\times 100$ ↑
CoFARL–TD3	$\mathbf{17.495 \pm 0.247}$	1.451 ± 0.124
DIF-SR adaptation	16.934 ± 0.218	1.386 ± 0.187
DIFF adaptation	16.693 ± 0.379	$\mathbf{1.494 \pm 0.194}$
TASIF adaptation	17.183 ± 0.307	1.187 ± 0.158

The comparison uses common evaluation access: full-catalog ranking and 128 simulator starts with 20-step returns. CoFARL and the fusion adaptations train on logged transitions. The reported external results use no additional simulator-training interactions. The six contrasts cover three baselines and two endpoints. Conservative Bonferroni reward bounds are 0.000199 (DIF-SR), 0.000097 (DIFF) and 0.126 (TASIF); these also bound Holm-adjusted values. CoFARL's mean reward gain over TASIF is 0.312. Appendix C.2 gives the bound calculation. Ranking retention denotes a ratio of means.

5.4 Expanding-window cryptocurrency evaluation

Building on the primary MovieLens study, the financial extension evaluates whether CoFARL can translate heterogeneous evidence into sequential asset selection under changing market conditions. We compare its copula-conditioned variants with three financial representation-learning adaptations: SARL-price/Top2, DeepTrader-ASU/Top2, and EarnMore-MAE/Top2, based on Ye et al. (2020), Wang et al. (2021b), and Zhang et al. (2023), respectively. All methods select two assets with equal, long-only weights from BTC, ETH, XRP, DOGE, LTC, and BCH.

Protocol. Training expands from 2018, with annual out-of-sample tests from 2019 through 30 August 2025. Each annual fit uses 8,192 RL interactions and a fixed final checkpoint. We report five-seed means of annualized Sharpe computed from 2,434 concatenated daily net returns, with a zero risk-free rate, 365-day annualization and ten basis points per unit turnover. Transformations and descriptors use only each training window. Appendix D details costs and evaluation scope.

Table 5: Cryptocurrency expanding-window evaluation. Entries are mean ± sample SD across five seeds. Pre denotes ranking pretraining; Aux adds continued ranking supervision. Every copula learner/supervision configuration is shown. External rows retain their adaptation labels. Bold marks the highest observed mean Sharpe in this table; it does not denote statistical significance.

Method	Supervision	Net OOS Sharpe ↑
CoFARL–PPO	Pre	0.759 ± 0.071
CoFARL–PPO	Pre + Aux	0.941 ± 0.052
CoFARL–SAC	Pre	0.956 ± 0.118
CoFARL–SAC	Pre + Aux	0.900 ± 0.072
CoFARL–TD3	Pre	0.882 ± 0.163
CoFARL–TD3	Pre + Aux	$\mathbf{0.957} \pm 0.185$
CoFARL–DQN	Pre	0.858 ± 0.095
CoFARL–DQN	Pre + Aux	0.947 ± 0.010
SARL-price/Top2	Price prediction	0.909 ± 0.108
DeepTrader-ASU/Top2	RL	0.801 ± 0.261
EarnMore-MAE/Top2	Reconstruction	0.743 ± 0.032

External comparison. Each of the four RL learners has a CoFARL configuration whose mean Sharpe exceeds all three external adaptations (Table 5). TD3 with continued ranking supervision and SAC with ranking pretraining attain closely matched means of 0.957 and 0.956, compared with 0.909 for SARL-price/Top2. The favourable mean comparison therefore spans both on-policy and off-policy learners, rather than depending on a single optimization algorithm.

Figure 2 shows five-seed mean Sharpe and seed variability across the complete configuration grid. Figure 3 provides the annual breakdown from the three archived seeds.

Interpretation. The comparison evaluates complete systems: the external methods adapt their native architectures to Top2 selection and use close-derived histories, whereas CoFARL uses heterogeneous feature groups. All eight copula configurations are reported; five-seed means describe the tested grid, with statistical and market-history uncertainty detailed in Appendix D.

5.5 RL4RS NATIVE-SIMULATOR RECOMMENDATION

We evaluate CoFARL in the RL4RS recommendation simulator (Wang et al., 2023), retaining its native action masks and reward checkpoint. The experiment compares copula, attention, identity-descriptor and raw-correlation actor representations under a common sequence/context critic, testing heterogeneous aggregation in multi-page recommendation with price-weighted predicted-response rewards.

Protocol. Each variant uses 256 ranking-pretraining updates followed by 1,296 PPO interaction steps (36 complete episodes), auxiliary ranking weight 0.1, and a fixed final checkpoint. Twenty paired training seeds per variant yield 80 fits. Final reward evaluation uses 32 fixed test simulator starts, each spanning 36 decisions. Ranking uses eligible pages from 128 test sessions and evaluates nine logged exposures per page. Both endpoints use the final trained checkpoint. Appendix D specifies the evaluation scope.

Table 6: RL4RS final-test performance: mean ± sample SD across 20 training seeds. Reward is native simulated cumulative return over 32 test starts. Ranking metrics are multiplied by 100. Bold identifies the highest observed mean in each column.

Actor representation	Reward ↑	NDCG@5 ↑	Recall@5 ↑	MAP@5 ↑
Copula	$\mathbf{229.163 \pm 11.631}$	$\mathbf{94.715 \pm 0.907}$	69.293 ± 0.659	92.220 ± 1.162
Attention	228.600 ± 9.903	94.558 ± 0.933	$\mathbf{69.356 \pm 0.662}$	92.135 ± 1.220
Identity descriptor	224.545 ± 17.844	94.654 ± 0.873	69.290 ± 0.759	$\mathbf{92.328 \pm 1.046}$
Raw correlation	223.947 ± 13.945	94.509 ± 0.937	69.237 ± 0.811	92.115 ± 1.183

Table 7: RL4RS final-test copula-minus-control differences across 20 paired seeds. Intervals are ordinary, unadjusted 95% CIs; NDCG differences use the $\times 100$ scale. Full multiplicity-adjusted inference appears in Appendix E.

Control	Reward difference [95% CI]	NDCG@5 difference [95% CI]
Attention	$0.563\ [-3.20, 4.33]$	$0.157\ [-0.11, 0.43]$
Identity	$4.618\ [0.30, 8.94]$	$0.061\ [-0.29, 0.41]$
Raw correlation	$5.216\ [1.45, 8.98]$	$0.206\ [-0.08, 0.49]$

Joint reward–ranking performance. Copula-conditioned PPO achieves the highest mean final-test reward and NDCG@5 among the four variants (Table 6), occupying the sole nondominated point of the observed mean reward–NDCG frontier. Its largest mean reward gains are against identity and raw correlation: $+4.618$ and $+5.216$, with ordinary paired 95% intervals $[0.30, 8.94]$ and $[1.45, 8.98]$. These controls distinguish estimated copula dependence from learned group offsets and linear dependence summaries. Holm-adjusted p-values are 0.40 and 0.12, retaining uncertainty after multiplicity correction. Against attention, the reward difference is $+0.563\ [-3.20, 4.33]$. Attention and identity lead mean Recall@5 and MAP@5, respectively; all ranking-difference intervals include zero. Table 7 gives the core contrasts; Appendix E reports all endpoints and adjusted tests.

Evaluation scope. The experiment evaluates held-out policy sessions within the native simulator. Exposure-conditioned ranking measures relevance within logged pages, and the public reward model's training coverage is unknown. The reported rewards are simulator predictions. Full split and metric definitions appear in Appendix D.

Policy sensitivity. A separate audit of 20 archived copula policies changes only the descriptor at identical states. Identity replacement, permutation and attenuation change 0.820%, 1.393% and 0.360% of actions, respectively. This establishes sensitivity of those policies to the descriptor; Appendix E distinguishes this audit from final-test inference.

6 Discussion and Conclusion

Matched aggregation controls, external system comparisons and frozen-policy interventions provide complementary evidence. The controls distinguish dependence conditioning from attention, group-specific offsets and simpler dependence summaries; interventions examine whether the descriptor affects decisions. RL4RS places copula conditioning at the highest observed mean reward and NDCG among its controls, with paired intervals and multiplicity-adjusted tests defining the statistical scope. MovieLens full-catalog ranking, RL4RS exposure-conditioned ranking and financial Sharpe retain their distinct interpretations.

The joint evaluation makes the design objective explicit: improve sequential utility while preserving recommendation relevance. The empirical reward–ranking frontier identifies the operating points offered by the evaluated systems. A natural extension is to select checkpoints on validation data subject to a prespecified ranking-quality constraint, then assess reward on held-out trajectories. This connects the aggregation mechanism to an actionable selection criterion while preserving separate uncertainty estimates for reward and ranking.

CoFARL provides a common architecture for combining temporal preferences, heterogeneous candidate information and sequential decision learning. The matched MovieLens study establishes ranking gains across all tested aggregation controls and concurrent reward gains over attention, permuted copula and sequence-only models. External fusion comparisons demonstrate the practical value of evaluating cumulative reward alongside ranking quality, while RL4RS and expanding-window financial experiments extend the framework to distinct decision settings. These findings support explicit dependence conditioning as a useful architectural component and motivate learning descriptors that adapt to evolving relationships, together with validation procedures that select policies under explicit ranking-quality requirements.

AI USE STATEMENT

AI tools assisted manuscript drafting and editing, checks of mathematical proofs, and literature-review verification through comparison of bibliographic records with primary sources. They also assisted code debugging.

REPRODUCIBILITY STATEMENT

The appendix provides mathematical proofs, implementation details, data splits, training and evaluation protocols, metric definitions and statistical analyses for the recommendation and financial experiments. The experiment archive contains source versions, configurations, random seeds, checkpoints and per-user or per-day evaluation outputs.

REFERENCES

Andrea Baisero and Christopher Amato. Unbiased Asymmetric Reinforcement Learning under Partial Observability. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 44–52. IFAAMAS, 2022. URL https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p44.pdf.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 2048–2056, 2020.

Erica Coppolillo, Giuseppe Manco, and Aristides Gionis. Relevance meets diversity: A user-centric framework for knowledge exploration through recommendations. arXiv preprint arXiv:2408.03772, 2024. URL https://arxiv.org/abs/2408.03772.

Romain Deffayet, Thibaut Thonet, Jean-Michel Renders, and Maarten de Rijke. Generative Slate Recommendation with Reinforcement Learning. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 580–588. ACM, 2023. doi: 10.1145/3539597.3570412. URL http://dx.doi.org/10.1145/3539597.3570412.

Daniel Ebi, Damien Ernst, Klemens Böhm, and Gaspard Lambrechts. Informed asymmetric actor-critic: Leveraging privileged signals beyond full-state access. arXiv preprint arXiv:2509.26000, 2026. URL https://arxiv.org/abs/2509.26000.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1587–1596, 2018.

Samuel Garcin, Trevor McInroe, Pablo Samuel Castro, Christopher Lucas, David Abel, Prakash Panangaden, and Stefano V. Albrecht. Studying the interplay between the actor and critic representations in reinforcement learning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 35590–35616, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/57da66da25d0ce77e0129b246f358851-Paper-Conference.pdf.

Kevin Gmelin, Shikhar Bahl, Russell Mendonca, and Deepak Pathak. Efficient RL via disentangled environment and agent representations. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 11525–11545, 2023.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1861–1870, 2018.

Pouya Hamadanian, Arash Nasr-Esfahany, Malte Schwarzkopf, Siddhartha Sen, and Mohammad Alizadeh. Online reinforcement learning in non-stationary context-driven environments. In *International Conference on Learning Representations*, 2025.

Eugene Ie, Vihan Jain, Jing Wang, Sanmit Narvekar, Ritesh Agarwal, Rui Wu, Heng-Tze Cheng, Morgane Lustman, Vince Gatto, Paul Covington, Jim McFadden, Tushar Chandra, and Craig Boutilier. Reinforcement learning for slate-based recommender systems: A tractable decomposition and practical methodology. arXiv preprint arXiv:1905.12767, 2019. URL https://arxiv.org/abs/1905.12767.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, pp. 3543–3556, 2019.

Wang-Cheng Kang and Julian McAuley. Self-Attentive Sequential Recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 197–206. IEEE, 2018. doi: 10.1109/icdm.2018.00035. URL http://dx.doi.org/10.1109/ICDM.2018.00035.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3744–3753, 2019.

Jie Luo, Wenyu Zhang, Xinming Zhang, and Yuan Fang. Time-aware adaptive side information fusion for sequential recommendation. arXiv preprint arXiv:2512.24246, 2025. URL https://arxiv.org/abs/2512.24246.

Jayanta Mandi, Víctor Bucarey, Maxime Mulamba Ke Tchomba, and Tias Guns. Decision-focused learning: Through the lens of learning to rank. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14935–14947. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/mandi22a.html.

Hanyu Peng, Guanhua Fang, and Ping Li. Copula for instance-wise feature selection and rank. In Robin J. Evans and Ilya Shpitser (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1651–1661. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/peng23a.html.

Aleksandr Petrov and Craig Macdonald. A systematic review and replicability study of bert4rec for sequential recommendation. arXiv preprint arXiv:2207.07483, 2022. URL https://arxiv.org/abs/2207.07483.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric Actor Critic for Image-Based Robot Learning. In *Robotics: Science and Systems XIV*. Robotics: Science and Systems Foundation, 2018. doi: 10.15607/rss.2018.xiv.008. URL http://dx.doi.org/10.15607/RSS.2018.XIV.008.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *arXiv preprint arXiv:1707.06347*, 2017.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Álvaro Labarca Silva, Denis Parra, and Rodrigo Toro Icarte. On the unexpected effectiveness of reinforcement learning for sequential recommendation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45432–45450. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/silva24b.html.

Gautam Singh, Skand Peri, Junghyun Kim, Hyunseok Kim, and Sungjin Ahn. Structured world belief for reinforcement learning in POMDP. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9744–9755, 2021.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1441–1450. ACM, 2019. doi: 10.1145/3357384.3357895. URL http://dx.doi.org/10.1145/3357384.3357895.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Hongwei Wang, Lantao Yu, Zhangjie Cao, and Stefano Ermon. Multi-agent imitation learning with copulas. arXiv preprint arXiv:2107.04750, 2021a. URL https://arxiv.org/abs/2107.04750.

Kai Wang, Zhene Zou, Minghao Zhao, Qilin Deng, Yue Shang, Yile Liang, Runze Wu, Xudong Shen, Tangjie Lyu, and Changjie Fan. Rl4rs: A real-world dataset for reinforcement learning based recommender system. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pp. 2935–2944. ACM, July 2023. doi: 10.1145/3539618.3591899. URL http://dx.doi.org/10.1145/3539618.3591899.

Zhicheng Wang, Biwei Huang, Shikui Tu, Kun Zhang, and Lei Xu. Deeptrader: A deep reinforcement learning approach for risk-return balanced portfolio management with market conditions embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):643–650, May 2021b. ISSN 2159-5399. doi: 10.1609/aaai.v35i1.16144. URL http://dx.doi.org/10.1609/aaai.v35i1.16144.

Chen-Yu Wei and Haipeng Luo. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. In *Proceedings of the 34th Conference on Learning Theory*, pp. 4300–4354, 2021.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, pp. 11–20, 2019.

Annie Xie, James Harrison, and Chelsea Finn. Deep reinforcement learning amidst continual structured non-stationarity. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11393–11403, 2021.

Yueqi Xie, Peilin Zhou, and Sunghun Kim. Decoupled Side Information Fusion for Sequential Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1611–1621. ACM, 2022. doi: 10.1145/3477495.3531963. URL http://dx.doi.org/10.1145/3477495.3531963.

Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, and Joemon M. Jose. Self-Supervised Reinforcement Learning for Recommender Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 931–940. ACM, 2020. doi: 10.1145/3397271.3401147. URL http://dx.doi.org/10.1145/3397271.3401147.

Yunan Ye, Hengzhi Pei, Boxin Wang, Pin-Yu Chen, Yada Zhu, Ju Xiao, and Bo Li. Reinforcement-learning based portfolio management with augmented asset movement prediction states. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1112–1119, April 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i01.5462. URL http://dx.doi.org/10.1609/aaai.v34i01.5462.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021.

Wentao Zhang, Yilei Zhao, Shuo Sun, Jie Ying, Yonggang Xie, Zitao Song, Xinrun Wang, and Bo An. Reinforcement learning with maskable stock representation for portfolio management in customizable stock pools, 2023. URL https://arxiv.org/abs/2311.10801.

Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. S3-Rec: Self-Supervised Learning for Sequential Recommendation with Mutual Information Maximization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1893–1902. ACM, 2020. doi: 10.1145/3340531.3411954. URL http://dx.doi.org/10.1145/3340531.3411954.

A Proofs and Analytical Details

The following derivations establish the propositions in Section 4. Each result states its own information, regularity or control assumptions.

A.1 Fixed-descriptor equivalence

Proof of Proposition 1.

Proof. For group basis vector e_g^{basis}, the descriptor contribution is $W_D R e_g^{\text{basis}}$. Set $\widetilde{W}_D = W_D R$ in the identity model. This preserves every logit and score for every history and candidate. Conversely set $W_D = \widetilde{W}_D R^{-1}$. Shrinkage makes our fitted R invertible. Equal function classes do not imply equal initialization, optimization trajectories or regularized objectives. □

The equivalence also identifies an optimization effect. Writing $A = W_D R$, ordinary gradient descent on W_D induces

$$A^+ = A - \eta \nabla_A \mathcal{L}\, R^T R.$$

Thus R acts as a fixed preconditioner on the effective projection. Moreover, an ℓ_2 penalty on W_D becomes $\|AR^{-1}\|_F^2$. This interpretation assumes that the loss depends on W_D through A; it describes a gradient-descent step, with other parameters held fixed. It explains why equal function classes can have different training behaviour.

A.2 Dependence and state distinguishability

Proof of Proposition 2.

Proof. For a standard bivariate normal with correlation c, its positive-quadrant probability is $1/4 + \arcsin(c)/(2\pi)$: differentiate the Gaussian quadrant integral in c, obtaining $1/(2\pi\sqrt{1-c^2})$, and integrate from $c = 0$. Hence $\Pr(UV > 0 \mid b) = 1/2 + b \arcsin(\rho)/\pi$. Without regime information any randomized action averages to $1/2$ over b. Choosing $a = b$ yields the stated gain. □

With a symmetrically misclassified regime estimate of error probability η, conditional independence of the estimation error and future outcome given the regime gives

$$\mathbb{E}[r] = \tfrac{1}{2} + (1 - 2\eta) \arcsin(\rho)/\pi.$$

This follows by weighting correct and incorrect regime decisions by $1 - \eta$ and η. An informative estimate ($\eta < 1/2$) retains a positive benefit. The construction isolates dependence information available to the decision rule; its base representation is identical in the two regimes.

A.3 Descriptor error and ranking stability

Proof of Proposition 3.

Proof. Tanh is 1-Lipschitz, so each logit changes by at most $\|a\|_2 \|W_D\|_2 \delta$. The softmax Jacobian gives the bound $\|\widehat{\alpha} - \alpha\|_1 \le 2\|\widehat{\ell} - \ell\|_\infty$. Weighted aggregation therefore changes by at most $2B\|a\|_2\|W_D\|_2\delta$; apply the score Lipschitz assumption. A boundary gap greater than twice the maximum error prevents a selected/unselected pair crossing. For the additive regret, insert and subtract the two perturbed set sums and use optimality under perturbed scores; each size-K sum changes by at most $K\varepsilon_s$. □

For completeness, let $p = \mathrm{softmax}(\ell)$ and let h perturb its logits. The Jacobian satisfies $(Jh)_g = p_g(h_g - \sum_j p_j h_j)$, giving $\|Jh\|_1 \le 2\|h\|_\infty$. Integrating along the segment between the two logit vectors proves the softmax bound used above.

If S^* and $\widehat{S}$ maximize reference and perturbed additive scores over sets of size K, respectively, then

$$\sum_{i\in S^*} s_i - \sum_{i\in\widehat{S}} s_i \le K\varepsilon_s + \left(\sum_{i\in S^*} \widehat{s}_i - \sum_{i\in\widehat{S}} \widehat{s}_i\right) + K\varepsilon_s \le 2K\varepsilon_s.$$

Recall depends on set membership. NDCG and MAP also depend on ordering within the selected set; preserving them requires sufficient margins for the relevant internal comparisons as well.

A.4 Value error and greedy control

Proof of Proposition 4.

Proof. Insert $\widehat{Q}$ between the optimal and selected actions. Greediness bounds their Q^* difference by $2\varepsilon_Q$. The Bellman recursion yields $\Delta \le 2\varepsilon_Q + \gamma\Delta$. For the residual statement, contraction of $\mathcal{T}^*$ gives $\varepsilon_Q \le \beta/(1-\gamma)$. □

More explicitly, for $a^* \in \arg\max_a Q^*(s,a)$ and $\widehat{a} \in \arg\max_a \widehat{Q}(s,a)$,

$$Q^*(s,a^*) - Q^*(s,\widehat{a}) \le |Q^*(s,a^*) - \widehat{Q}(s,a^*)| + |\widehat{Q}(s,\widehat{a}) - Q^*(s,\widehat{a})| \le 2\varepsilon_Q.$$

The remaining loss propagates through the transition kernel under $\widehat{\pi}$, yielding the geometric factor $1/(1-\gamma)$. For the residual form, the optimal Bellman operator gives

$$\|\widehat{Q} - Q^*\|_\infty \le \beta + \gamma\|\widehat{Q} - Q^*\|_\infty.$$

The result concerns a greedy policy in a discounted Markov model. Applying it to a learned history representation requires that representation to satisfy the stated state and error assumptions.

A.5 Numerical checks

Archived numerical checks support the algebra and probability calculations. Reparameterizing the fixed-descriptor scorer produced a maximum score difference of zero. In 200,000 Gaussian draws per correlation, the observed probabilities of equal signs were 0.498750, 0.597685, 0.706160 and 0.854165 for correlations 0, 0.3, 0.6 and 0.9, respectively; the analytical values were 0.500000, 0.596987, 0.704833 and 0.856434. A further 1,000 perturbation trials checked Top-K stability under the stated margin condition. These checks supplement the derivations above.

B MovieLens Implementation and Evaluation

B.1 Temporal representation and dependence descriptor

The history encoder uses at most 20 preceding positive interactions, two causal Transformer layers, two heads and width 64. Padding is masked and empty histories use a beginning-of-sequence token. Candidate embeddings combine content, history-affinity and popularity features. The Gaussian descriptor uses training-fitted empirical marginals, clipped Gaussian quantiles and shrinkage $R = 0.9\,\mathrm{Corr}(z) + 0.1I$. It is frozen during RL.

The original three-seed factorial fits retain the implementation's positional mapping: dependence evidence is ordered as affinity, popularity and repeat, while aggregation groups are ordered as content, history and popularity. Thus a descriptor row is a fixed conditioning input to a group rather than a verified semantically aligned dependence profile for that group. Those original factorial results concern this implementation. The aligned 20-seed TD3 experiment is reported separately in Section 5.2.

B.2 External comparison

The main table uses three seeds (8801–8803) and the same 128 evaluation users. CoFARL, SASRec, SQN and Self-Supervised Actor-Critic train on filtered logged transitions. The latter two use the

recorded recurrent reference implementations, while SASRec uses causal attention. These reference implementations differ in architecture and loss from CoFARL.

CoFARL uses filtered-history ranking pretraining. The reported external results use no additional simulator-training interactions. During simulator evaluation, each decision presents 32 candidates and selects two items. The simulator ranks the eligible catalog by retriever score, breaking ties by descending training count and ascending catalog index. It retains the first 16 items and fills the remaining slots with nonduplicate popularity-ranked items. This deterministic retrieval order defines the candidate slots; the concatenated state and action heads are not permutation invariant. Ranking auxiliary batches instead contain the observed target and 31 uniformly sampled distinct non-target items, with randomized target position. PPO and DQN choose a discrete candidate pair; SAC and TD3 output continuous scores whose Top-2 set is deployed. The ten context entries are history-length fraction, distinct-item fraction, mean normalized popularity, empty-history indicator, recent response mean, smoothed cumulative response, log exposure count, last response, log clipped elapsed seconds, and normalized episode clock.

Validation native-scorer NDCG@10 selects the CoFARL checkpoint. Test scorer ranks and simulated return are evaluated from the same selected checkpoint.

B.3 Internal factorial comparison

The internal study crosses PPO, SAC, TD3 and DQN with attention, identity and copula aggregation, and with ranking pretraining alone or pretraining plus continued supervision. Three seeds give 72 MovieLens fits and 24 configuration means. The internal preparation supplies 30,000 positive training contexts. The external preparation retains 9,048 after excluding 20,952 examples with ambiguous simultaneous positive timestamps. Both use a 1,024-item output catalog. Internal ranking pretraining uses ten epochs, batch size 32 and 128 sampled candidates, with one seed-8101 parent per aggregation mode reused across the three RL seeds. External CoFARL uses seed-specific parents trained for 10,000 updates, batch size 32 and 32 candidates. Internal auxiliary batches have size 16; external auxiliary batches have size 32. Internal RL checkpoint selection maximizes validation simulated return at 4,096/8,192 interactions. External CoFARL selection maximizes validation native-scorer NDCG@10. The external displayed ranking metrics use the same 128 users as policy evaluation, selected from 1,486 saved test users. Shared retrieval, response and feature-statistic assets retain their broader training-data provenance. These differences affect both data access and model selection; numerical values from the two studies are not pooled or interpreted as an isolated architectural effect. Copula and identity have the same descriptor-projection parameterization; attention removes that term.

B.4 Endpoints and inference

Ranking uses the full eligible catalog and the held-out next-positive target. A target outside that catalog counts as a miss. We retain NDCG, Recall and MAP at $K = 5, 10, 20$. With one relevant target, MAP is truncated reciprocal rank. Simulated reward is cumulative return over 20 interaction steps. Actual deployed Top-2 relevance is saved separately from scorer ranking.

All tables report mean and sample standard deviation across training seeds. Paired seed-level sign-flip tests use their recorded comparison families and Holm corrections, which are preserved when presenting MovieLens. Three paired seeds permit a minimum two-sided exact sign-flip p-value of 0.25. The reported corrected comparisons therefore support descriptive analysis rather than statistical superiority or equivalence. Seed uncertainty describes variation under the recorded MovieLens preparation and simulator, rather than transfer to another dataset or real-world deployment.

B.5 Computational cost

Let H be history length, C the number of scored candidates, G the number of groups, F the group input width, and d the hidden width. For a fixed number of Transformer layers, temporal encoding costs $O(H^2d + Hd^2)$. Group projection costs $O(CGFd)$, descriptor projection $O(G^2d)$, and gate evaluation and aggregation $O(CGd + d^2)$. The feature-score MLP costs $O(Cd^2)$. These costs are per scorer forward pass; actor and critic copies and auxiliary minibatches add corresponding evaluations. Attention activations require $O(H^2 + Hd)$ space and candidate group embeddings $O(CGd)$,

excluding parameters and optimizer/replay storage. Pair-action PPO/DQN heads additionally scale with $C(C-1)/2$ outputs; their dense input-output projection cost must be added. Full-catalog ranking substitutes the eligible catalog size for C and can be evaluated in candidate batches. This analysis states asymptotic costs rather than a measured speedup over external models.

For the external CoFARL–TD3 Recall@10 entry, independent reconstruction from per-user ranks gives exactly two hits among 128 users for each seed. Checkpoint hashes, hit users and hit ranks differ across seeds. The reported zero sample standard deviation is therefore correct for the aggregate hit fraction; it does not imply identical predictions or zero uncertainty over users.

C Complete Internal MovieLens Results

Table 8 contains all 24 internal configuration means. Pretrain denotes ranking pretraining followed by RL; continued adds ranking supervision during RL. These results use the internal protocol and are separate from the external comparison in Table 1.

Table 8: Internal MovieLens configurations, mean ± seed standard deviation over three seeds. NDCG@10 is scaled by 1,000.

Learner	Aggregation	Supervision	Reward	NDCG@10
DQN	attention	pretrain	17.138 ± 0.194	1.891 ± 0.337
DQN	attention	continued	17.448 ± 0.109	4.362 ± 1.081
DQN	copula	pretrain	17.438 ± 0.637	5.595 ± 0.775
DQN	copula	continued	17.215 ± 0.145	4.993 ± 1.526
DQN	identity	pretrain	17.358 ± 0.571	5.995 ± 0.280
DQN	identity	continued	17.372 ± 0.442	5.685 ± 0.412
PPO	attention	pretrain	17.199 ± 0.188	14.111 ± 1.538
PPO	attention	continued	17.076 ± 0.212	16.934 ± 1.021
PPO	copula	pretrain	17.043 ± 0.119	15.409 ± 1.784
PPO	copula	continued	17.167 ± 0.022	16.320 ± 1.181
PPO	identity	pretrain	17.085 ± 0.091	13.819 ± 1.632
PPO	identity	continued	17.277 ± 0.072	16.917 ± 0.820
SAC	attention	pretrain	17.385 ± 0.069	10.924 ± 4.399
SAC	attention	continued	17.285 ± 0.175	9.510 ± 2.622
SAC	copula	pretrain	17.488 ± 0.149	12.739 ± 1.514
SAC	copula	continued	17.375 ± 0.077	12.412 ± 1.771
SAC	identity	pretrain	17.492 ± 0.223	12.504 ± 1.099
SAC	identity	continued	17.422 ± 0.102	13.271 ± 2.265
TD3	attention	pretrain	17.210 ± 0.466	13.704 ± 1.055
TD3	attention	continued	17.475 ± 0.190	17.675 ± 1.582
TD3	copula	pretrain	17.482 ± 0.011	13.133 ± 0.475
TD3	copula	continued	17.388 ± 0.166	17.789 ± 0.840
TD3	identity	pretrain	17.480 ± 0.136	13.006 ± 1.276
TD3	identity	continued	17.216 ± 0.170	14.878 ± 1.680

C.1 Paired inference for the twenty-seed TD3 comparison

Table 9 reports aligned-copula-minus-control contrasts for the matched experiment in Table 3. Differences are paired across 20 training seeds. The 95% confidence intervals are unadjusted paired intervals; Holm correction covers the ten reward and ranking comparisons jointly. NDCG differences and intervals use the $\times 100$ scale. Seed variation characterizes training randomness on the fixed evaluation cohort.

Table 9: MovieLens TD3 paired comparisons across 20 seeds. Bold adjusted p-values identify comparisons passing the 0.05 threshold after Holm correction across all ten tests.

Endpoint	Copula minus	Difference	Paired 95% CI	Holm p
Reward	Attention	$+0.200$	$[+0.055, +0.345]$	$\mathbf{0.032}$
	Identity	$+0.107$	$[-0.034, +0.248]$	0.263
	Raw correlation	$+0.126$	$[-0.025, +0.277]$	0.205
	Permuted copula	$+0.191$	$[+0.046, +0.336]$	$\mathbf{0.041}$
	Sequence-only	$+0.360$	$[+0.154, +0.566]$	$\mathbf{0.008}$
NDCG@10 $\times 100$	Attention	$+0.539$	$[+0.452, +0.626]$	$< \mathbf{0.001}$
	Identity	$+0.523$	$[+0.409, +0.637]$	$< \mathbf{0.001}$
	Raw correlation	$+0.568$	$[+0.470, +0.666]$	$< \mathbf{0.001}$
	Permuted copula	$+0.518$	$[+0.405, +0.631]$	$< \mathbf{0.001}$
	Sequence-only	$+0.859$	$[+0.765, +0.953]$	$< \mathbf{0.001}$

C.2 Multiplicity bounds for the external fusion comparison

Section 5.3 compares CoFARL with DIF-SR, DIFF and TASIF on reward and NDCG@10, giving six contrasts. For $n = 20$ paired observations, the sample standard deviation of the differences satisfies $s_\Delta \leq s_{\mathrm{CoFARL}} + s_{\mathrm{baseline}}$. Consequently, $|\bar{\Delta}|\sqrt{20}/(s_{\mathrm{CoFARL}} + s_{\mathrm{baseline}})$ lower-bounds the absolute paired-t statistic. The corresponding two-sided tail probability with 19 degrees of freedom, multiplied by six and capped at one, upper-bounds both Bonferroni- and Holm-adjusted p-values. Applied to the displayed reward summaries, this gives 0.000199, 0.000097 and 0.126 against DIF-SR, DIFF and TASIF, respectively. These are conservative bounds rather than exact paired tests. Ranking retention is descriptive.

D Extension protocols and interpretation

D.1 Cryptocurrency

Chronological evaluation. Training begins in 2018 and expands annually: training on 2018 is followed by testing on 2019; training on 2018–2019 is followed by testing on 2020; this continues through the final test period ending on 30 August 2025. Feature transformations and dependence descriptors are fitted on each training window, and policy parameters remain fixed during the following test year. Each annual fit uses 8,192 RL interactions and a fixed final checkpoint. We evaluate five training seeds, giving 280 copula fits across four learners and two supervision settings, alongside 105 external-adaptation fits. The complete internal architecture campaign additionally contains attention, identity, and sequence-only controls.

The primary endpoint is the annualized Sharpe ratio of the concatenated 2,434 daily out-of-sample net returns, with zero risk-free rate and a 365-day annualization factor. Returns include ten basis points per unit of turnover, entry costs, and terminal liquidation at each annual boundary. We compute Sharpe on the concatenated return series, rather than averaging annual Sharpe ratios.

Comparison scope. The baseline adaptations retain SARL's price-prediction augmentation, DeepTrader's asset-scoring unit, and EarnMore's masked representation learning within a shared Top2 action constraint. Their native allocation mechanisms and some auxiliary objectives are changed. Baselines receive standardized close-derived return histories, whereas CoFARL uses its heterogeneous feature groups; the comparison therefore measures complete systems with different representations and information access. The displayed configurations were evaluated as a grid, and the highest observed means are descriptive rather than validation-selected winners. With five paired seeds, the minimum exact two-sided sign-flip p-value is 0.0625; these results do not establish statistical superiority. Seed variation also does not capture uncertainty across market histories. The fixed retrospective asset universe and current data vintage delimit the financial generalization claim.

D.2 RL4RS

Final-test protocol. We report the final-checkpoint endpoints, which were secondary endpoints in the frozen study protocol. The prespecified learning-curve analysis is retained in the experiment archive. Each variant uses 256 exposure-ranking pretraining updates followed by 1,296 PPO simulator steps, corresponding to 36 complete 36-step training episodes, with continued ranking supervision of weight 0.1. Twenty paired seeds per variant yield 80 fits across copula, attention, identity and raw-correlation representations. All variants use a fixed final checkpoint. The prepared cohort contains 4,096 training sessions and 128 sessions for each validation and test split, selected in deterministic file order from an existing session-hash split. Dependence and exposure statistics are fitted on the training split. Reward evaluation uses 32 fixed simulator starts; ranking evaluation uses eligible pages from the 128 test sessions.

Evaluation scope. Ranking is exposure-conditioned: duplicate-item and zero-positive pages are excluded, and the metric assesses ordering among nine logged exposures. These values measure a different task from full-catalog recommendation. The session split is neither chronological nor guaranteed user-disjoint, and the public simulator checkpoint's training coverage is unknown. Consequently, the evidence concerns held-out policy-session evaluation within this simulator, rather than independently held-out real-user outcomes. Together with the financial extension, this evaluation broadens the settings in which CoFARL can be evaluated while preserving separate interpretations of ranking quality, simulated reward, and realized financial returns.

E RL4RS FINAL-TEST PAIRED COMPARISONS

Table 10 reports secondary final-checkpoint comparisons across 20 paired seeds, using the same final-test results as Table 6. Positive differences favour copula. Reward uses 32 test starts and ranking uses 128 test sessions. Ordinary 95% confidence intervals are unadjusted; Holm-adjusted values address the 12 endpoint–control comparisons. Seed uncertainty concerns training variability on these fixed evaluation cohorts.

Table 10: Final-test copula-minus-control differences. Ranking differences and intervals use the $\times 100$ scale.

Endpoint	Control	Difference	95% CI	Holm p
Reward	Attention	$+0.563$	$[-3.20, 4.33]$	1.00
Reward	Identity	$+4.618$	$[0.30, 8.94]$	0.40
Reward	Raw correlation	$+5.216$	$[1.45, 8.98]$	0.12
NDCG@5	Attention	$+0.157$	$[-0.11, 0.43]$	1.00
NDCG@5	Identity	$+0.061$	$[-0.29, 0.41]$	1.00
NDCG@5	Raw correlation	$+0.206$	$[-0.08, 0.49]$	1.00
Recall@5	Attention	-0.063	$[-0.32, 0.20]$	1.00
Recall@5	Identity	$+0.003$	$[-0.35, 0.36]$	1.00
Recall@5	Raw correlation	$+0.056$	$[-0.22, 0.33]$	1.00
MAP@5	Attention	$+0.085$	$[-0.29, 0.46]$	1.00
MAP@5	Identity	-0.108	$[-0.56, 0.34]$	1.00
MAP@5	Raw correlation	$+0.105$	$[-0.28, 0.49]$	1.00

Policy sensitivity to dependence. A separate frozen-checkpoint audit evaluates 20 archived copula policies with their original descriptor, identity replacement, a coordinate permutation and half-strength dependence. At identical states and native action masks, identity replacement changes 0.820% of actions, permutation changes 1.393%, and attenuation changes 0.360%. Thus the copula descriptor participates in the learned decision rule. The intervention audit characterizes its archived policies separately from the final-test summary in Table 6; its reward and ranking contrasts do not pass the six-comparison Holm correction. Action sensitivity and training comparisons provide complementary evidence about policy use and evaluated utility.

F EXTENDED RELATED-WORK CONTEXT

Sequential recommendation and feature fusion. SASRec learns causal attention over interaction histories (Kang & McAuley, 2018), while BERT4Rec uses bidirectional masked-item prediction (Sun et al., 2019). S^3-Rec combines attribute and sequence self-supervision (Zhou et al., 2020). For side-information fusion, DIF-SR separates item and attribute attention to address restrictions on attention rank and gradient flow (Xie et al., 2022); TASIF combines temporal information with adaptive fusion (Luo et al., 2025). CoFARL uses a causal sequence encoder but conditions its feature-group weights on an explicit dependence descriptor. The relevant comparison is therefore with both sequence-only models and models that adaptively fuse side information.

Aggregation and conditional computation. Deep Sets characterizes permutation-invariant functions of sets (Zaheer et al., 2017), while Set Transformer models interactions between their elements through attention (Lee et al., 2019). Transformer attention (Vaswani et al., 2017) and sparsely gated mixtures of experts (Shazeer et al., 2017) provide complementary ways to condition computation on an input. CoFARL applies conditional weighting to semantically defined feature groups. Its copula descriptor supplies a statistical relationship between those groups alongside the history and candidate embeddings. This distinction motivates the MovieLens controls: identity preserves the descriptor projection while changing its input, and attention removes the dependence-conditioning term.

Supervised objectives and reinforcement learning. Xin et al. (2020) combine supervised recommendation with Q-learning and actor–critic heads. Their Self-Supervised Actor-Critic framework is distinct from the Soft Actor-Critic algorithm evaluated in our MovieLens study (Haarnoja et al., 2018). Silva et al. (2024) show that next-item accuracy and long-term reward can favour different policies. They also find that a simple interaction-count objective can reproduce some of the ranking gains obtained with auxiliary RL. These results motivate separate evaluations of ranking and sequential utility, together with a non-RL auxiliary-supervision control.

Actor and critic representations. Asymmetric actor–critic learning supplies training information to the critic while retaining the actor's deployment observations (Pinto et al., 2018). More recent work examines the interplay between actor and critic representations, including their effects on data collection (Garcin et al., 2025), and the informativeness of privileged signals (Ebi et al., 2026). Under partial observability, the conditioning of an asymmetric value function also affects the validity of policy-gradient estimates (Baisero & Amato, 2022). These results place representation choice within the learning dynamics. For CoFARL, this literature motivates examining the representations consumed by native policy and value updates. The MovieLens comparisons vary the learner and aggregation mechanism; branch-specific placement is a separate experimental question.

Task-relevant state structure. Bisimulation-based representations preserve behavioural distinctions while reducing irrelevant observation variation (Zhang et al., 2021). Structured world beliefs organize information for partially observed control (Singh et al., 2021), and disentangled agent and environment representations provide another structural prior (Gmelin et al., 2023). The shared issue is how a representation retains information relevant to decisions. Our dependence-aliasing construction makes this issue explicit: regimes with equal marginals admit different decisions when their joint distributions differ. It motivates dependence summaries that complement a compressed history, and the MovieLens analysis measures ranking and return as distinct outcomes.

Copula models. Peng et al. (2023) use Gaussian copulas to sample dependent feature selections and rankings. CoFARL instead uses a copula descriptor to condition deterministic feature-group weights. Copula-based imitation learning models dependence between agents' actions (Wang et al., 2021a), and copula-based recommendation has been used to combine relevance and diversity (Coppolillo et al., 2024). The random variables differ across these settings. Our descriptor represents dependence between specified feature-evidence summaries and conditions their aggregation before scoring.

Slate decisions and task-aligned estimation. SlateQ decomposes slate values under assumptions on user choice (Ie et al., 2019). Generative slate recommendation learns a continuous representation

of complete slates for RL (Deffayet et al., 2023). These methods contextualize slate selection beyond additive item scores. Decision-focused learning connects predictive models to downstream ranking and optimization objectives (Mandi et al., 2022). Decision-focused estimation offers an extension in which downstream utility shapes a statistically valid dependence model. Its evaluation combines dependence calibration with decision quality.

Changing dependence. Continual structured non-stationarity (Xie et al., 2021), non-stationary RL with regret guarantees (Wei & Luo, 2021), and policy optimization under time-varying exogenous context (Hamadanian et al., 2025) motivate adaptation to changing contexts. The MovieLens architecture studied here uses a descriptor fitted once on training observations. History-conditioned gates adapt their weights while that descriptor remains fixed. Estimating an evolving descriptor is a natural extension of this design.

Interpreting attention and evaluating generalization. Work on attention as explanation examines the relationship between model weights and faithful accounts of predictions (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). For dependence-conditioned aggregation, intervention on a descriptor or representation provides a direct test of its effect on the score or policy. Routing weights serve as accompanying diagnostics. Evaluation across procedurally generated environments, as in Procgen (Cobbe et al., 2020), also distinguishes training replication from generalization across environments. We accordingly separate training-seed uncertainty, catalog coverage and the domain covered by each experiment.

Evaluation. Training duration and implementation choices can substantially affect recommender comparisons (Petrov & Macdonald, 2022). We distinguish logged-data training from additional simulator access and report seed-level uncertainty. Our empirical question concerns the reward–ranking operating points reached on MovieLens. The identity and attention controls examine the role of dependence conditioning within the internal scoring architecture.

G Cryptocurrency and RL4RS visual analyses

The figures report the completed cryptocurrency grid and RL4RS final-test comparison separately. Cryptocurrency panels contain all eight copula learner/supervision configurations and all three external adaptations. RL4RS panels contain all four actor representations. Individual seed points, standard deviations, and paired confidence intervals are identified explicitly.

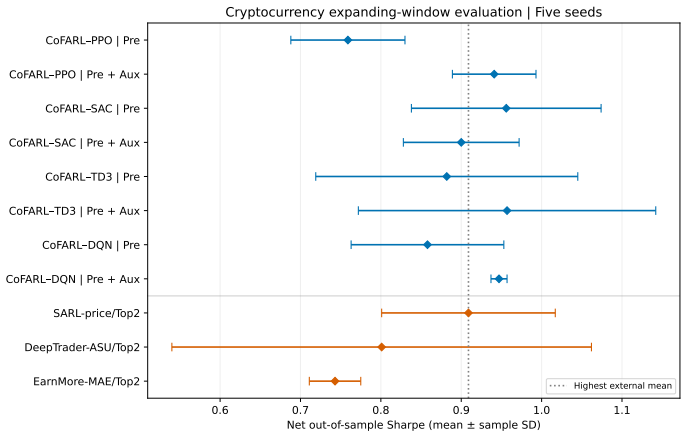

Figure 2: Stitched out-of-sample net Sharpe over 2,434 days. Diamonds and bars show mean and sample SD across five training seeds. The dotted line is the SARL-price/Top2 mean, the highest external-adaptation mean. Pre denotes ranking pretraining and Aux denotes continued ranking supervision. Four configurations, spanning all four learners, exceed this reference on mean Sharpe. This is a descriptive comparison across the complete tested grid.

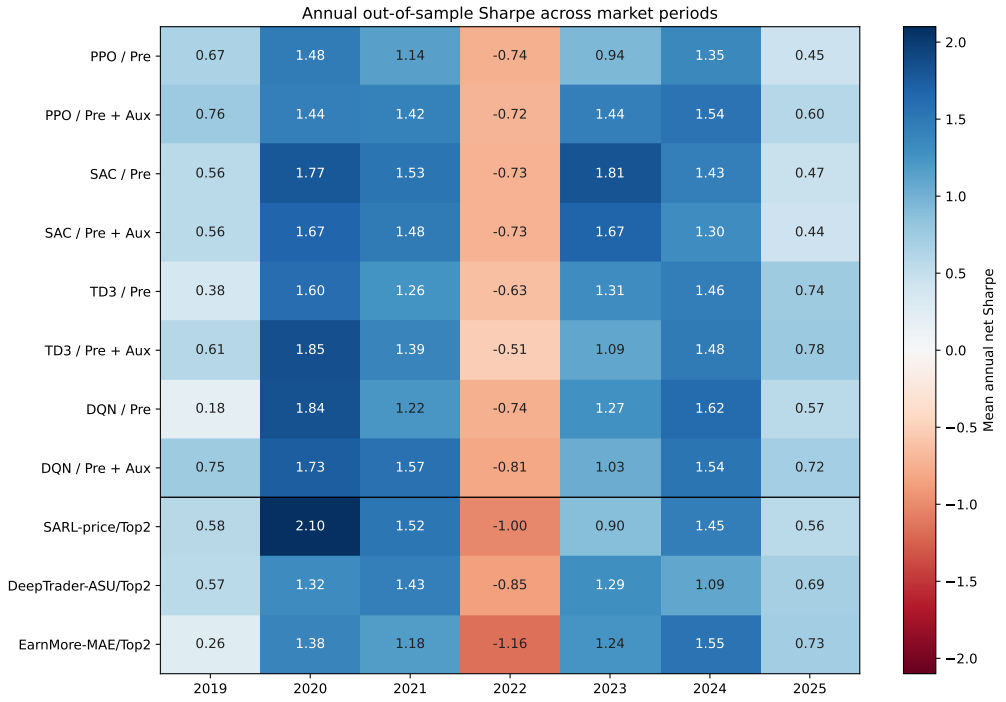

Figure 3: Annual test-period net Sharpe, averaged across the three archived training seeds. Every model is retrained on the expanding history preceding its test year. The final column covers the available 2025 test period specified in Appendix D. The annual breakdown shows variation across market periods; the primary stitched Sharpe is computed from daily returns rather than averaging these cells.

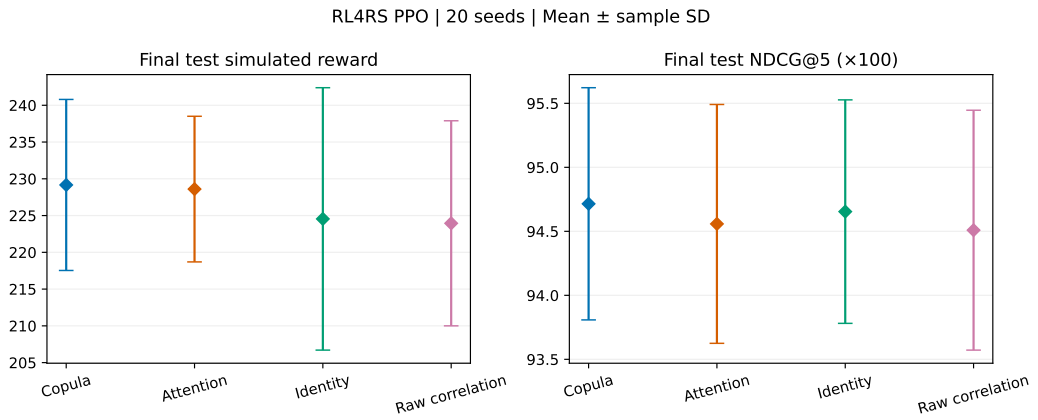

Figure 4: RL4RS final-test reward and NDCG@5 across four PPO aggregation variants. Points show means and bars show sample SD across 20 training seeds. Reward uses 32 test simulator starts; ranking uses nine logged exposures per page across 128 test sessions. Copula has the highest mean reward and NDCG@5. These are evaluated operating points; bars represent seed variability rather than paired confidence intervals.