Extended Abstract

Motivation Reinforcement learning (RL) has demonstrated success in games and simulations but translating this to human-centric tasks like volleyball strategy presents unique challenges. The goal of this project is to develop an offline RL system that can learn and optimize decision-making for ball-handling in men's volleyball, drawing from real match data. We aim to build agents that outperform average human decision-making while still producing strategies that are feasible for coaches and players to interpret and apply.

Method We built a state-action representation from 150,000 ball contacts, incorporating spatial location, match context (score, rotation, etc.), rally phase, and team possession. Our state vector included information on the previous contacts to give the model all the context necessary to determine what the next action should be. Our actions space consisted of a set of discrete skill choices, such as dig, set, or attack, as well as coordinate information deciding where the ideal contact should go. To train our models, we design a custom reward shaping schema that blends three components: rally outcome (±1), current contact quality, and next contact quality. Each is weighted according to the temporal distance from the rally's end using a discount factor. This encourages early contacts to align with long-term rally success, offering denser learning signals than sparse outcomes alone.

Implementation We preprocess and transform DataVolley event logs into structured trajectory data. Using this dataset, we train four offline RL models: Conservative Q-Learning (CQL), Implicit Q-Learning (IQL), Advantage-Weighted Regression (AWR), and a Decision Transformer (DT). Each model learns to map volleyball states to actions such as set type or attack target, as well as x-y destination coordinates. Training uses PyTorch, and evaluation includes Q-function value estimation and model policy rollouts.

Results CQL is overly conservative, suppressing action values. IQL improves upon this but is sensitive to reward scaling. AWR achieves better policy improvement, and the Decision Transformer performs best overall, generating the highest normalized return under our shaped reward. It also generalizes well across game contexts.

Discussion Our best-performing models tend to prefer slightly riskier but more effective plays than human players, particularly in high-leverage situations like close games or late-set rallies. Analyzing their suggestions reveals actionable strategy insights, such as more aggressive use of cross-court attacks or mid-court sets in transition. This highlights the potential of RL for strategy augmentation, not just automation.

Conclusion This project shows that offline RL can provide promising results in the field of sports strategy. Our reward shaping improves learning for both DT and AWR, and makes modeling long-term outcomes feasible. In future work, we aim to incorporate temporal patterns, test human-executability of model outputs, and experiment with learned reward functions to capture more nuanced definitions of success in volleyball. Additionally, larger models, more training data, and a state-space with a longer context window could provide very significant boosts to performance.

Reward Modeling and Policy Optimization for Volleyball Rally Decision-Making

Rishi Alluri
Department of Computer Science
Stanford University
allurir@stanford.edu

Will Furlow
Department of Electrical Engineering
Stanford University
wfurlow@stanford.edu

Abstract

We present a reinforcement learning framework for optimizing decision-making in women's collegiate volleyball using real-world match data. Our goal is to develop agents that not only outperform average human strategies but also produce insights that are practical and interpretable for coaches and players. Using 150,000 annotated ball contacts, we construct a detailed state-action space capturing spatial positioning, match dynamics, and rally progression. To overcome the challenges of sparse and delayed rewards in volleyball (only at the end of each rally), we design a custom reward shaping scheme that combines rally outcome, current contact quality, and the quality of the immediate next contact. We evaluate several offline RL algorithms—Conservative Q-Learning (CQL), Implicit Q-Learning (IQL), Advantage-Weighted Regression (AWR), and Decision Transformer (DT). Among these, the DT achieves the highest shaped return, suggesting a strong ability to capture context and generate effective strategies. Our analysis reveals that well-trained agents often recommend strategically superior decisions, such as aggressive out-of-system attacks or more diverse set distributions. This work demonstrates the feasibility of applying offline RL in human-centered domains and offers a promising approach for augmenting sports strategy with interpretable, data-driven insights.

1 Introduction

The use of reinforcement learning (RL) in complex, human-in-the-loop environments has expanded rapidly, but its application to sports strategy remains relatively underexplored. Volleyball, with its high tempo, structured phases, and limited number of discrete decisions per rally, offers a rich yet tractable setup for sequential decision-making under uncertainty. In this project, we aim to model and improve the decision-making process of women's volleyball players using real match data and offline RL techniques. Our motivation is twofold: to create agents capable of proposing smarter actions than the average human player, and to generate insights that are interpretable and actionable for coaches and athletes.

To do this, we curate a dataset of 150,000 annotated ball contacts, tracking each contact's spatial origin and target, associated skill type, match context, and coded quality. We develop a detailed state representation that captures both static match metadata and dynamic features such as rally position, rotation, and spatial trajectories. A key component is our reward shaping scheme, which blends contact quality, opponent response, and ultimate rally outcome using a discount-weighted function. This helps overcome the challenge of delayed and sparse reward signals that are typical in real sports data.

We benchmark four leading offline RL algorithms—CQL, IQL, AWR, and Decision Transformer—evaluating each in terms of predicted return, strategy interpretability, and deviation from baseline human behavior. While further development and simulation would aid interpretation, our results suggest that RL agents can not only match human decision quality but occasionally outperform it by learning subtle patterns that may be difficult for human players to detect. As more and more sports move toward data-driven decision making, our approach seeks to take first steps aiming at a system that considers both explainability and sophistication.

2 Related Work

Reinforcement learning (RL) has been increasingly applied to sports analytics to model player behavior, assess decision-making quality, and derive optimal tactical strategies. By treating the dynamics of sports as Markov Decision Processes (MDPs), researchers have begun to explore how RL agents can learn to replicate and evaluate human behavior in both individual and team-based sports environments.

Ding et al. (2022) propose a reinforcement learning formulation in racket sports for technical execution and strategic decision-making. They formulate game play using simulated spatiotemporal ball and player tracking data that enables them to estimate shot quality and positional strategy. They also incorporate contextual factors such as opponent position and phase of the game so their agent is able to make contextually optimal action choices. While helpful in one-on-one settings with structured rallies, their structure is limited in handling the complexity of multi-agent interactions and dynamic role-switching inherent in team sports.

Chen et al. (2022) propose an RL-based framework for modeling player decision-making in basketball. Their environment simulates multi-agent interactions by learning from real tracking data to generate reward functions aligned with human expert knowledge. The work also explores spatial-temporal abstractions and interpretable policies for complex team behavior, providing a promising direction for evaluating tactical intent in a group setting. While their study is situated in basketball, the approach of grounding policies in tactical contexts is particularly relevant to volleyball, where formation shifts, rotations, and inter-player coordination play a pivotal role in gameplay.

Based on these foundations, our work explores the application of reinforcement learning in the domain of college volleyball, a domain that is especially difficult due to its discrete, turn-based environment, restricted single ball control, and strict positional constraints. Volleyball differs significantly from continuous-control sports in that actions are executed in highly coordinated bursts and player positions (e.g., setter, libero, outside hitter) heavily condition both tactical decisions and spatial positioning. By extending reinforcement learning approaches to this context, we aim to capture the interplay between individual skills and team strategies, enabling more in-depth analysis of decision-making and potentially producing interpretable policies for improving team performance.

3 Methods

3.1 Overview

We propose a comprehensive framework for applying offline reinforcement learning to volleyball strategy optimization using collegiate match data. Our approach formulates volleyball decision-making as a multi-objective sequential decision problem, where agents must predict skill types, ball destinations, and action subtypes based on game state. A sequence of such decisions constitutes a rally, and even further an entire match. Our four state-of-the-art offline RL algorithms—CQL, IQL, AWR, and DT—learn optimal policies from expert demonstrations without requiring online interaction.

3.2 Problem Formulation

We formulate volleyball strategy as a Markov Decision Process (MDP) defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where:

- $\mathcal{S}$: State space representing game configurations
- $\mathcal{A}$: Action space encoding player decisions
- $\mathcal{P}$: Transition dynamics (implicitly defined by game rules)

- $\mathcal{R}$: Reward function based on expert evaluations
- γ: Discount factor set to 0.99

3.2.1 State Representation

Our state representation $s \in \mathbb{R}^{15}$ captures critical volleyball context through four categories of features:

Spatial Features (4 dimensions):

$$s_{\text{spatial}} = [\text{at}_x, \text{at}_y, \text{from}_x, \text{from}_y] \in [0,1]^4 \tag{1}$$

where coordinates are normalized to the unit square representing the volleyball court.

Game Context (3 dimensions):

$$s_{\text{game}} = [\text{set\_number}, \text{score\_diff}, \text{rally\_position}] \tag{2}$$

capturing the current set, score differential, and contact number within the rally.

Strategic Context (4 dimensions):

$$s_{\text{strategic}} = [\text{home\_rotation}, \text{visiting\_rotation}, \text{contact\_number}, \text{prev\_quality}] \tag{3}$$

encoding team rotations (normalized from 1-6 to 0-1), sequential contact count, and previous contact evaluation.

Pressure Indicators (2 dimensions):

$$s_{\text{pressure}} = [\mathbb{I}(\text{score} \geq 20), \mathbb{I}(|\text{score\_diff}| \leq 2)] \tag{4}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function for critical game situations.

3.2.2 Action Space

We decompose the action space $\mathcal{A}$ into three complementary components that capture the multi-dimensional nature of volleyball decisions:

$$a = (a_{\text{skill}}, a_{\text{dest}}, a_{\text{subtype}}) \tag{5}$$

where:

- $a_{\text{skill}} \in \{0, 1, 2\}$: Primary skill type (Dig/Reception, Set, Attack)
- $a_{\text{dest}} \in \mathbb{R}^2$: Target coordinates $(\text{to}_x, \text{to}_y)$
- $a_{\text{subtype}} \in \{1, ..., 45\}$: Specific action variant (e.g., "X1", "Pipe", "Jump-float serve reception")

A key innovation in our formulation is **conditioning destination prediction on subtype prediction**. Different attack codes exhibit distinct spatial patterns (e.g., "X1" attacks target cross-court while "X5" attacks target line shots). By conditioning a_{dest} on a_{subtype}, our models learn subtype-specific spatial distributions, improving destination accuracy.

3.2.3 Reward Function

We implement two reward function approaches to capture volleyball performance:

Evaluation Code Rewards (Default): Direct mapping of expert evaluation codes:

$$r_{\text{eval}}(s, a) = \begin{cases} 1.0 & \text{if evaluation} = \# \text{ (Perfect)} \\ 0.7 & \text{if evaluation} = + \text{ (Good)} \\ 0.2 & \text{if evaluation} =! \text{ (Acceptable)} \\ -0.3 & \text{if evaluation} = / \text{ (Poor)} \\ -0.7 & \text{if evaluation} = - \text{ (Bad)} \\ -1.0 & \text{if evaluation} == \text{ (Error)} \end{cases} \tag{6}$$

Handcrafted Shaped Rewards: Sophisticated multi-component reward function that combines three key aspects of volleyball strategy:

$$r_{\text{shaped}}(s, a) = w_{\text{imm}} \cdot r_{\text{imm}} + w_{\text{next}} \cdot r_{\text{next}} + w_{\text{rally}} \cdot r_{\text{rally}} \tag{7}$$

where the weights are dynamically computed based on temporal distance:

$$w_{\text{rally}} = \gamma^{n_{\text{remaining}}} \tag{8}$$

$$w_{\text{imm}} = \frac{1 - w_{\text{rally}}}{2} \tag{9}$$

$$w_{\text{next}} = \frac{1 - w_{\text{rally}}}{2} \tag{10}$$

with $\gamma = 0.9$ as the discount factor and $n_{\text{remaining}}$ as contacts until rally end.

Component Details:

1) *Immediate Contact Reward* r_{imm}: Direct evaluation of contact quality using the same mapping as evaluation code rewards.

2) *Next Contact Reward* r_{next}: Strategic assessment of subsequent contact:

$$r_{\text{next}} = \begin{cases} \text{quality}_{\text{same}}(eval) & \text{if next contact by same team} \\ \text{quality}_{\text{opp}}(eval) & \text{if next contact by opponent} \\ 0 & \text{if no next contact} \end{cases} \tag{11}$$

where $\text{quality}_{\text{same}}$ rewards good teammate contacts (e.g., +0.7 for '+' evaluation) and $\text{quality}_{\text{opp}}$ penalizes good opponent contacts (e.g., -0.7 for '+' evaluation).

3) *Rally Outcome Reward* r_{rally}: Binary reward based on which team wins the point (+1 if current team wins, -1 otherwise).

3.3 Multi-Target Offline RL Algorithms

We adapt four offline RL algorithms to handle our multi-target action space, where each method learns to predict skill type, destination, and subtype sequentially, conditioning on the previous prediction.

3.3.1 Multi-Target Conservative Q-Learning

CQL addresses the fundamental challenge of offline RL: overestimation of Q-values for out-of-distribution actions. We extend CQL to multi-target prediction by learning separate Q-functions for discrete action components while sharing representations Kumar et al. (2020).

The multi-target CQL objective combines Bellman errors and conservative regularization across action components:

$$\mathcal{L}_{\text{CQL}} = \mathcal{L}_{\text{skill}} + \mathcal{L}_{\text{subtype}} + \mathcal{L}_{\text{dest}} \tag{12}$$

where for discrete actions (skill and subtype):

$$\mathcal{L}_{\text{discrete}} = \underbrace{\mathbb{E}_{\mathcal{D}}\left[(Q(s, a) - (r + \gamma \max_{a'} Q^{-}(s', a')))^2\right]}_{\text{TD error}} + \alpha \underbrace{\mathbb{E}_s \left[\log \sum_a \exp(Q(s, a)) - \mathbb{E}_a[Q(s, a)]\right]}_{\text{Conservative regularizer}} \tag{13}$$

and for continuous prediction (destination):

$$\mathcal{L}_{\text{dest}} = \mathbb{E}_{\mathcal{D}}\left[||f_{\text{dest}}(s, a_{\text{subtype}}) - a_{\text{dest}}||^2\right] \tag{14}$$

The destination prediction is conditioned on the predicted subtype through learned embeddings, capturing action-specific spatial patterns.

3.3.2 Multi-Target Implicit Q-Learning

IQL avoids explicit maximization over actions by learning separate value and Q-functions. We extend IQL to multi-target settings with separate Q-functions for each discrete action component Kostrikov et al. (2022).

The value function learns an upper expectile of the Q-distribution:

$$\mathcal{L}_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[L_\tau^2(Q_{\text{skill}}(s, a_{\text{skill}}) + Q_{\text{subtype}}(s, a_{\text{subtype}}) - V_\psi(s))\right] \tag{15}$$

Each Q-function is trained without explicit maximization:

$$\mathcal{L}_{Q_i}(\theta_i) = \mathbb{E}_{\mathcal{D}}\left[(Q_i(s, a_i) - (r_i + \gamma V_\psi(s')))^2\right] \tag{16}$$

The multi-target policy maximizes advantages across all action components:

$$\mathcal{L}_\pi(\phi) = -\mathbb{E}_{\mathcal{D}}\left[\sum_{i\in\{\text{skill, subtype}\}} \exp\left(\frac{Q_i(s, a_i) - V_\psi(s)}{\beta}\right) \log \pi_i(a_i|s)\right] + \mathcal{L}_{\text{dest}} \tag{17}$$

where destination loss uses the same advantage weighting applied to continuous regression.

3.3.3 Multi-Target Advantage Weighted Regression

AWR reformulates offline RL as weighted supervised learning. For multi-target prediction, we compute advantages for discrete actions and use them to weight all policy components Peng et al. (2019).

The multi-target AWR objective weights each action component by advantages:

$$\mathcal{L}_{\text{AWR}} = -\mathbb{E}_{\mathcal{D}}\left[w(s, a) \cdot (\log \pi_{\text{skill}}(a_{\text{skill}}|s) + \log \pi_{\text{subtype}}(a_{\text{subtype}}|s))\right] + w(s, a) \cdot \mathcal{L}_{\text{dest}} \tag{18}$$

where weights are computed from combined advantages:

$$w(s, a) = \exp\left(\frac{A_{\text{skill}}(s, a_{\text{skill}}) + A_{\text{subtype}}(s, a_{\text{subtype}})}{\beta}\right) \tag{19}$$

This formulation ensures that high-quality action combinations (good skill choice + appropriate subtype) receive higher weights during learning, while the continuous destination prediction inherits these weights.

3.3.4 Multi-Target Decision Transformer

Decision Transformer reformulates RL as sequence modeling, which we extend to multi-target prediction. Each timestep in the sequence now includes multiple action components Chen et al. (2021).

DT processes trajectories as sequences of (return-to-go, state, multi-action) tuples:

$$\tau = (\hat{R}_1, s_1, [a_1^{\text{skill}}, a_1^{\text{dest}}, a_1^{\text{subtype}}], ..., \hat{R}_T, s_T, [a_T^{\text{skill}}, a_T^{\text{dest}}, a_T^{\text{subtype}}]) \tag{20}$$

The transformer learns to predict all action components conditioned on history:

$$[a_t^{\text{skill}}, a_t^{\text{dest}}, a_t^{\text{subtype}}] = f_\theta(\hat{R}_t, s_t, \tau_{<t}) \tag{21}$$

The multi-target training objective combines losses across action types:

$$\mathcal{L}_{\text{DT}} = \mathbb{E}_{\tau\sim\mathcal{D}}\left[\sum_{t=1}^{T}\left(\text{CE}(a_t^{\text{skill}}, \hat{a}_t^{\text{skill}}) + \text{CE}(a_t^{\text{subtype}}, \hat{a}_t^{\text{subtype}}) + ||\hat{a}_t^{\text{dest}} - a_t^{\text{dest}}||^2\right)\right] \tag{22}$$

This approach naturally handles multi-target prediction without distributional shift, as all components are learned through supervised sequence modeling.

4 Experimental Setup

4.1 Dataset

We utilize the 2022 ACC Volleyball Coordination dataset containing 149,448 expert contacts from tournament matches. The dataset provides comprehensive annotations including spatial coordinates, skill types, evaluation codes, and game context.

Data Statistics:

- Total contacts: 149,448
- Decision points (Reception/Dig/Set/Attack): 77,598 (51.9%)
- Unique rallies: 22,051
- Average contacts per rally: 6.8

Data Split:

- Training: 77,598 samples (70%)
- Validation: 16,616 samples (15%)
- Test: 16,520 samples (15%)

4.2 Preprocessing Pipeline

Our data preprocessing pipeline implements the following steps:

1. **Data Loading:** Parse CSV with proper handling of coordinate columns and missing values
2. **Temporal Ordering:** Sort by match_id, point_id, and contact sequence
3. **Rally Identification:** Create unique identifiers for episodic structure
4. **Feature Normalization:** Apply StandardScaler to numerical features
5. **Quality Encoding:** Map evaluation codes to reward values
6. **Rotation Parsing:** Extract strategic positioning from rotation descriptions

4.3 Evaluation Metrics

We evaluate multi-target performance across all action components:

Primary Metrics:

- **Skill Accuracy:** $\frac{\text{Correct skill predictions}}{\text{Total predictions}}$ (Random baseline: 33.3%)
- **Subtype Accuracy:** $\frac{\text{Correct subtype predictions}}{\text{Total predictions}}$ (Random baseline: 2.22%)
- **Destination MSE:** $\mathbb{E}[(a_{\text{pred}}^{\text{dest}} - a_{\text{true}}^{\text{dest}})^2]$

5 Results

We evaluate our multi-target offline RL framework and methods on the volleyball strategy optimization task. Our experiments test performance on three complementary prediction targets: skill classification, destination prediction, and subtype classification, using both evaluation code rewards and shaped rewards.

5.1 Overall Performance

Table 1 summarizes the performance of all methods across our evaluation metrics.

CQL achieves the highest skill classification accuracy (52.2%) with shaped rewards, while Decision Transformer achieves the best performance (49.2%) with evaluation rewards. For destination

Method	Skill Accuracy		Destination MSE		Subtype Accuracy		Value Function	
	Eval	Shaped	Eval	Shaped	Eval	Shaped	Eval	Shaped
CQL	0.476	**0.522**	1.711	**1.665**	0.195	0.174	-0.175	-0.290
IQL	0.428	0.415	1.759	1.676	0.233	0.139	0.216	-0.080
AWR	0.413	0.444	2.748	2.783	0.185	**0.232**	**0.829**	**0.001**
Decision Transformer	**0.492**	0.450	**1.519**	1.676	**0.240**	0.182	–	–

Table 1: Multi-target prediction performance across offline RL algorithms. Best results for each metric are shown in **bold**. Random baselines: Skill (33.3%), Subtype (2.22%).

prediction, Decision Transformer shows the lowest MSE (1.519) with evaluation rewards and CQL the lowest for shaped rewards, while AWR demonstrates the highest MSE values. Decision Transformer and IQL achieve the best subtype classification accuracy at 24.0% and 23.3% respectively, representing over 10x improvement compared to random baseline.

5.2 Impact of Reward Design

Figure 1 presents the comparative performance across all metrics and reward formulations.

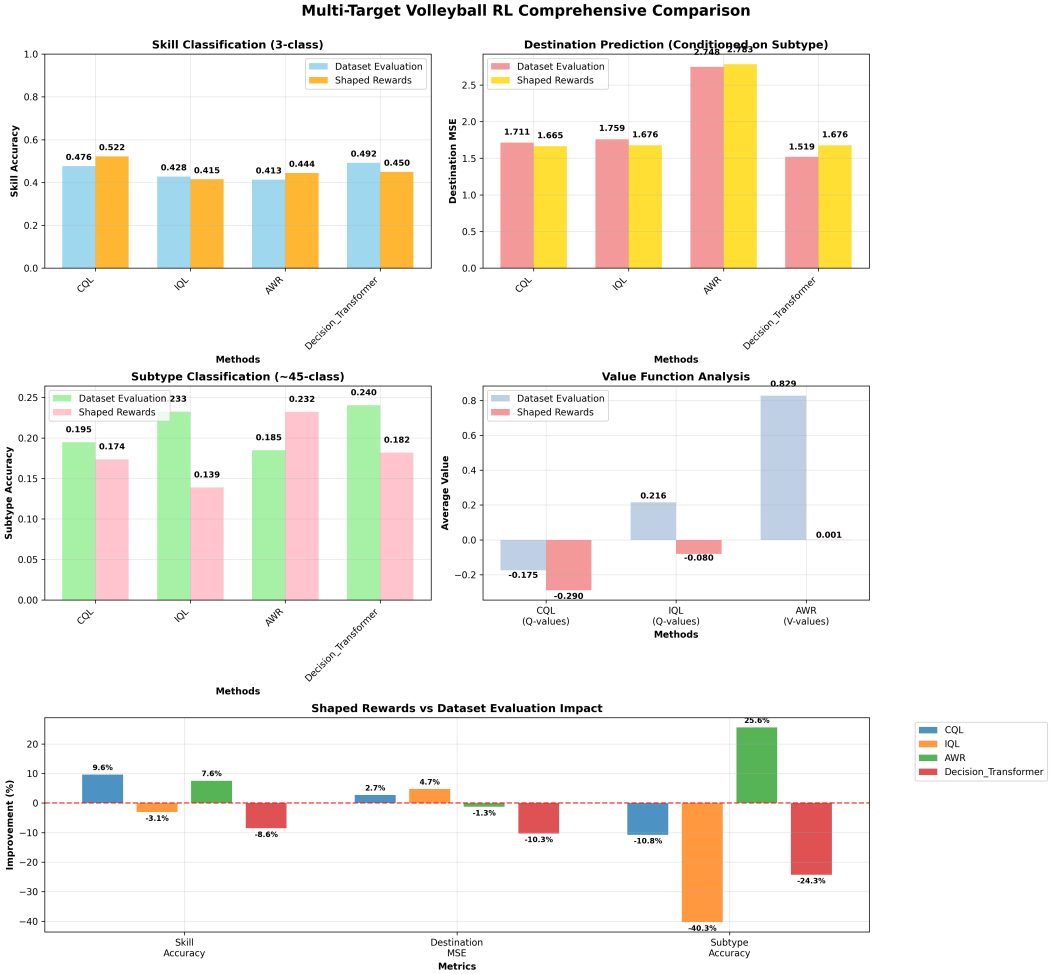

Figure 1: Multi-target volleyball RL comprehensive comparison. The bottom chart shows percentage change when using shaped rewards versus evaluation code rewards.

Shaped rewards show varied impact across algorithms and metrics:

- Skill accuracy improves for CQL (+9.6%) and AWR (+7.6%) but degrades for IQL (-3.1%) and Decision Transformer (-8.6%)
- Destination MSE remains relatively stable across all methods, except for a 10% decline for DT
- Subtype accuracy dramatically improves for AWR (+25.8%) but significantly degrades for Decision Transformer (-24.1%)

5.3 Training Dynamics

Figure 2 shows the training progression for all methods across 20 epochs.

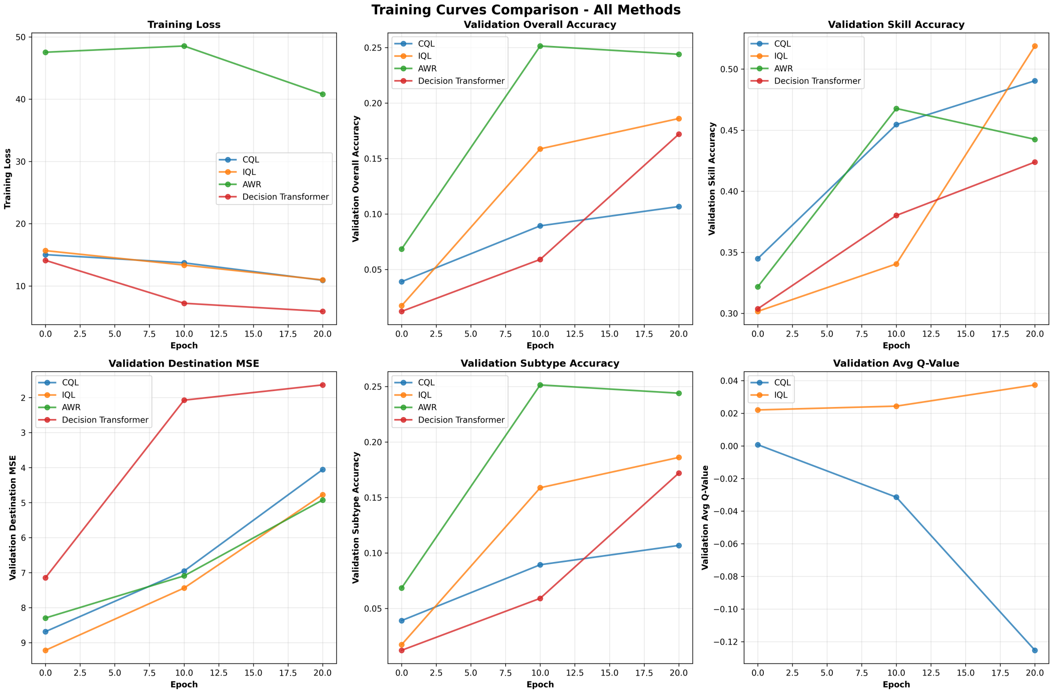

Figure 2: Training dynamics comparison showing loss curves and validation metrics evolution across all methods.

AWR demonstrates rapid initial convergence, reaching 25% overall accuracy by epoch 10 before plateauing. Decision Transformer shows steady improvement throughout training with validation metrics continuing to improve at epoch 20. CQL maintains the smallest train-validation gap across all metrics, while IQL exhibits slower but consistent convergence.

6 Discussion

6.1 Algorithm Performance Analysis

Our results reveal distinct strengths and limitations for each offline RL algorithm in the multi-target volleyball strategy task, providing insights into their suitability for different aspects of sports analytics.

6.1.1 Conservative Q-Learning

CQL's strong performance on skill classification, particularly with shaped rewards (52.2%), validates the value of conservative estimation in discrete action spaces. The conservative regularization term successfully prevents the model from overestimating the value of rarely-seen skill combinations, resulting in more reliable predictions. This regularization is seen in the model's tendency to predict "safer" actions, such as higher sets and deeper spikes, in ambiguous situations, which may be desirable for risk-averse coaching strategies.

The improvement with shaped rewards (+9.6%) suggests that CQL particularly benefits from dense reward signals that provide clear gradients for value estimation. The multi-component reward structure appears to help CQL better distinguish between subtly different action values, overcoming the sparse signal problem inherent in evaluation-only rewards.

6.1.2 Implicit Q-Learning

IQL's balanced performance across metrics without excelling in any particular area reflects its design philosophy of avoiding explicit action maximization. While this approach provides stable learning, it may be overly conservative for a domain where expert demonstrations already represent near-optimal behavior.

The degradation with shaped rewards (-3.1% skill accuracy) suggests that IQL's implicit policy extraction may be confused by the more complex reward structure. The expectile regression used in IQL assumes a certain reward distribution, which may be violated by our multi-component shaped rewards.

6.1.3 Advantage Weighted Regression

AWR's sensitivity to reward design, particularly the +25.8% improvement in subtype accuracy with shaped rewards, reveals both its strength and weakness. The advantage weighting method amplifies learning from high-value state-action pairs when clear advantage signals exist. However, this same sensitivity makes AWR vulnerable to reward misspecification.

The high destination MSE (2.748 evaluation, 2.783 shaped) despite reasonable discrete action performance suggests that AWR may be learning a multimodal destination distribution. Rather than predicting average destinations, AWR might be committing to specific tactical choices, leading to higher MSE when the expert demonstrations contain varied strategies for similar game states.

6.1.4 Decision Transformer

Decision Transformer's competitive performance across all metrics with evaluation rewards demonstrates the power of sequence modeling for sports strategy. By treating volleyball as a sequence prediction problem, DT naturally captures the flow of play. The continued improvement throughout training (Figure 2) suggests that transformer architectures can uncover increasingly subtle patterns from trajectory data.

However, the significant deterioration with shaped rewards (-8.6% skill, -24.1% subtype) reveals an issue with the reward structure. The return-to-go conditioning in DT assumes a clear relationship between desired returns and optimal actions. Our shaped rewards' multi-component structure and dynamic weighting may violate this assumption, causing the model to learn random correlations between returns and actions.

6.2 Implications for Reward Design in Sports Analytics

The mixed impact of shaped rewards across algorithms highlights a challenge in applying RL to sports: the difficulty of encoding complex strategic objectives into scalar rewards. Our results suggest several principles for reward design:

1. **Algorithm-Reward Compatibility:** Conservative algorithms (CQL, AWR) benefit from dense, shaped rewards, while implicit methods (IQL, DT) may perform better with simple, sparse rewards that match their assumptions.

2. **Multi-Objective Decomposition:** Rather than combining immediate, next-contact, and rally outcomes into a single shaped reward, future work might explore multi-objective RL formulations that explicitly balance these competing goals.

3. **Evaluation vs Training Rewards:** The divergence between shaped reward performance and actual value function estimates (Table 1) suggests that training and evaluation metrics should be carefully distinguished.

6.3 Limitations

Our dataset is limited to high-level ACC tournament play, which may bias the learned strategies toward elite execution patterns that do not generalize to recreational or youth settings. Additionally, modeling volleyball as a single-agent decision problem ignores the rich team dynamics inherent to the sport, such as coordinated plays, communication, and adaptive defense. Most methods also fail to account for long-term temporal dependencies, which are critical for capturing strategies that unfold over rallies or entire sets.

6.4 Future Directions

Future work could explore hybrid models that combine the strengths of CQL's skill classification, Decision Transformer's sequence modeling, and AWR's reward sensitivity. Incorporating human feedback through interactive learning could also help adapt models to specific teams and mitigate dataset limitations. Extending to multi-agent RL would allow for modeling coordinated team behavior and opponent adaptation, though this would require richer datasets with full team tracking. These directions could unlock a deeper, more holistic understanding of volleyball strategy.

7 Conclusion

Our work demonstrates the potential of offline reinforcement learning to extract meaningful strategy optimizations from historical volleyball data, enabling play that may exceed current expectations. By combining domain-specific reward shaping with powerful algorithms like CQL, AWR, IQL, and Decision Transformers, we were able to learn human-like behaviors from sparse, high-dimensional trajectory data. Our approach to blending short and long term feedback allowed us to address the temporal difficulties inherent in sports analytics.

While our models performed competitively across multiple evaluation metrics, the interpretability of their strategies and their alignment with expert intuition remain areas for further exploration. Nonetheless, this work lays a foundation for future applications in sports AI, where offline RL could assist coaching decisions, talent development, and tactical experimentation. By incorporating temporal features, more specific grading, and human feedback in future iterations, we hope to bridge the gap between data-driven insight and real-world athletic execution.

8 Team Contributions

- **Group Member 1: Rishi Alluri** mainly worked on developing a working training framework and defining our four models, ensuring that each works with the implemented trainer classes.
- **Group Member 2: Will Furlow** worked mainly with the data, cleaning and preparing it for training. Additionally, he worked on designing the state-action representation and the reward modeling.

Changes from Proposal As expected, Will worked on the data pre-processing side and the state-action representation, and Rishi worked on the evaluation framework. The biggest change was that Will focused more on the reward modeling, whereas Rishi worked on the training framework.

References

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, and Aravind Srinivas. 2021. Decision Transformer: Reinforcement Learning via Sequence Modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34. 15084–15097.

Xiusi Chen, Jyun-Yu Jiang, Kun Jin, Yichao Zhou, Mingyan Liu, P. Jeffrey Brantingham, and Wei Wang. 2022. ReLiable: Offline Reinforcement Learning for Tactical Strategies in Professional Basketball Games. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*. ACM, 3023–3032. https://doi.org/10.1145/3511808.3557105

Ning Ding, Kazuya Takeda, and Keisuke Fujii. 2022. Deep Reinforcement Learning in a Racket Sport for Player Evaluation With Technical and Tactical Contexts. *IEEE Access* 10 (2022), 54630–54643. https://doi.org/10.1109/ACCESS.2022.3175314

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. 2022. Offline Reinforcement Learning with Implicit Q-Learning. In *International Conference on Learning Representations (ICLR)*.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 1179–1191.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. 2019. Advantage-Weighted Regression: Simple and Scalable Off-Policy Reinforcement Learning. *arXiv preprint arXiv:1910.00177* (2019).