Extended Abstract

Motivation In modern day football, it's the offensive coordinator's job to sequence decisions under uncertainty. They must decide, given the game situation, which play to call. The current flaw in this system is that Coaches and analysts almost always grade those calls by the result, but this result is mostly noise (execution, the defense, luck); the part they should actually want graded is the value of the decision, making sure a good call that got unlucky still grades well. I set out to build a tool that grades every play call by its learned expected value rather than its outcome, and that proposes a better call when one is genuinely warranted. This is an offline reinforcement learning problem, can I learn a value function and improved policies entirely from logged plays, with no online interaction?

Method To start, I treat a football possession as a Markov decision process over drives and learn from 316,821 modeled offensive plays drawn from a dataset of more than 2,280 games (2023 to 2025). The state encodes the full situation (down, distance, field position, score, clock, personnel, team and matchup talent) together with eleven alignment features read from the All-22 film at the snap (full frame by frame player tracking data); the action is one of 29 concepts whose vocabulary is learned from the data. An ensemble of five Expected Points models fit to the Monte Carlo return until the next score, an ensemble Win Probability model that targets PFF's (a trusted datasource used by all top college and pro teams) per play win probability, and a distributional head of nineteen quantiles for risk. I then compare three offline reinforcement learning algorithms, naive fitted Q, Implicit Q-Learning (IQL), and Conservative Q-Learning (CQL), and evaluate with off policy methods (the direct method, doubly robust estimation, and sequential Fitted Q Evaluation).

Implementation The grade is the learned Expected Points value of the concept the coach called, switched to its Win Probability value when the game is late and close; the recommendation is the best alternative that survives three guards, sticking with what the behavior policy actually calls in that state, awareness of risk, and an anchor toward the behavior policy, with confidence read from the agreement of an ensemble of three IQL policies. Everything runs offline from a cached PFF and All-22 dataset, with a split by game between training and test so there is no leakage across the analysis scripts. The same engine drives a product that shows one play at a time, placing the real All-22 clip next to the grade, the measured read, and the suggested call, with a teaching layer that explains each in plain coaching terms.

Results The Models are highly calibrated on held out data (EP error: 0.063, WP error: 0.009). The WP lens accurately shifts the late game run rate from 16% to 85% to protect leads, avoiding the EP lens which would still be chasing more points coaches don't want. While naive Q-learning over calls deep passes (61% vs. 13% actual), conservative evaluation matches true drive values. The Marginal Sensitivity Model bounds the naive policy's edge at Gamma = 1.30, matching the film measured confounding (Gamma = 1.22), which explains 74% of the apparent edge.

Discussion I eliminated every fixable cause of the confound in turn (more data, fixed effects for offense and defense by season, correct pricing of the risk in sacks and interceptions, finer route concepts, and a pull of full alignment tracking from the All-22 film), and showed from two independent angles, a sensitivity bound and a film measurement, that the rest is genuine selection on the coach's private read. Because that read cannot serve as a feature available before the snap, the confound in the recommendation is irreducible from observational data. Among the three algorithms, IQL handles the overextrapolation best by staying within the data, where naive Q and evaluation that leans on the value model are fooled; IQL alone offers a smooth and controllable aggression dial, recovering legitimate deep usage as its temperature rises. CQL ends up shrinking the deep ball usage all the way to zero.

Conclusion PlayGrader reliably grades play calling using calibrated EP and WP metrics. Rather than ignoring hidden coaching reads, it bounds and validates them using sensitivity modeling and acutal film data. The result is an end-to-end offline RL system combining robust evaluation, distributional value models, and a functional film-room tool.

PlayGrader: Coaching the Coaches with Deep RL

Group Member 1
Department of Computer Science
Stanford University
jammcana@stanford.edu

Abstract

I study offline reinforcement learning for grading and improving an American football offense's play calls. A possession is modeled as a Markov decision process over drives; from 316,821 plays across more than 2,280 FBS games I learn ensemble Expected Points and Win Probability value functions and a nineteen quantile distributional head. Then I compare three offline RL algorithms (naive fitted Q, Implicit Q-Learning, and Conservative Q-Learning) under three off policy estimators (the direct method, doubly robust estimation, and sequential Fitted Q Evaluation). Grading the decision is reliable and well calibrated, and the learned value recovers known football structure. The recommendation is the hard part: the coach selects plays influenced by a pre-snap read we never observe, which confounds the logged data. I show that naive value maximization exploits this confound, then treat it with a Marginal Sensitivity Model that bounds each recommendation's robustness and an All-22 film measurement of the read itself. The bound and the measurement agree: the confounding I can measure covers 74% of what it would take to erase the aggressive policy's apparent edge. The result is a calibrated, honest deep RL system, delivered as a working film room product.

1 Introduction

An offensive coordinator chooses, in each situation defined by down, distance, field position, score, and clock, which play to call. The standard way to evaluate a call, in film study and in public analytics, is by its result; but the per play result is dominated by noise (blocking, the defensive call, a drop, a missed tackle). What a coach actually wants graded is the value of the decision, so that a good call that got unlucky still grades well. That reframing is exactly the value function at the heart of reinforcement learning.

I built PlayGrader, a system that learns the value of a play call from logged football and uses it two ways: to grade every call by its expected value, and to recommend a better call when one is warranted. Because we only ever have logged data and never interact with a live defense, this is offline RL. Can we learn a value function and an improved policy from a fixed dataset?

The defining difficulty is unobserved confounding. The behavior policy is a human coach who calls plays partly on a pre-snap read (often from previous game scouting) that we do not record and that also moves the outcome. A value function fit to this data credits aggressive calls such as deep shots for the favorable contexts the coach selected into, and a policy that maximizes that value recommends them far too often. This is an example of offline RL faling due to value overestimation on out of distribution actions, here driven by a real hidden variable rather than a sampling artifact.

My contributions are: (1) a calibrated ensemble value model over a drive level football MDP that recovers known football structure; (2) a controlled comparison of three offline RL algorithms under three off policy estimators that makes the confound visible and show how each algorithm responds; and (3) a treatment of the confound that goes beyond conservatism, bounding each recommendation's

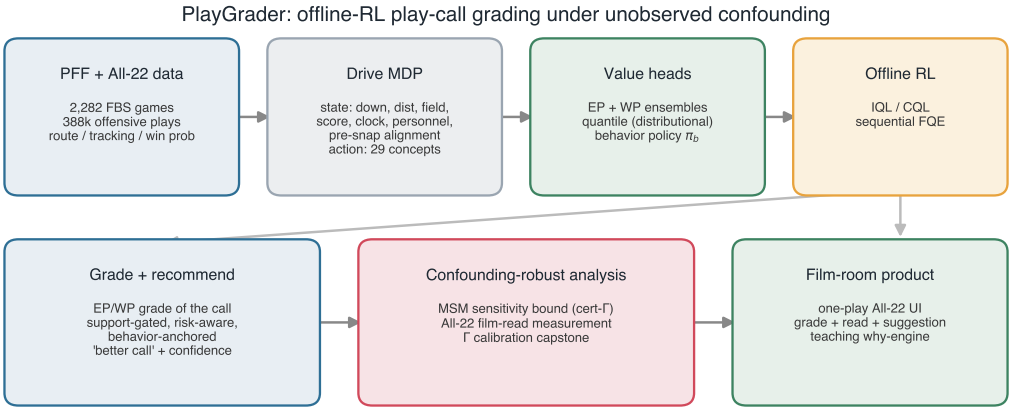

Figure 1: System overview. From PFF and All-22 data I build a drive level MDP, learn Expected Points, Win Probability, and quantile value heads plus a behavior policy, fit offline RL policies (IQL and CQL) evaluated by sequential FQE, and grade and recommend play calls, surfaced in a one play film room product.

robustness with a Marginal Sensitivity Model and measuring the hidden read directly from All-22 film, then showing the two agree.

2 Related Work

The offline RL setting, learning a policy from a fixed dataset without exploration, is surveyed by Levine et al. [7]. Two standard remedies anchor my comparison: Conservative Q-Learning [5] penalizes the value of out of distribution actions, while Implicit Q-Learning [4] avoids querying them entirely by regressing an expectile of the value over dataset actions and extracting a policy with advantage weighted regression [9]. My distributional head follows quantile regression as in QR-DQN [1]. To score policies without deployment I use the direct method, doubly robust estimation [2], and Fitted Q Evaluation [6]; the direct method trusts the value model and is therefore vulnerable to the same confound, which is part of the finding. For the confound I adapt the Marginal Sensitivity Model [10, 12], brought to policy evaluation by Kallus and Zhou [3] and studied for nonidentifiable hidden confounding by Pace et al. [8], and pair it with a direct film measurement that calibrates the abstract sensitivity parameter against an observed quantity. Expected points and win probability are established sports value scales [11]; I learn both as the MDP value rather than importing a feed, which lets me verify calibration directly.

3 Method

3.1 The football MDP

A possession is a Markov decision process over a drive. The state s encodes the situation (down, distance, yards to go, score differential, quarter, two clock features), offensive personnel and formation, team and matchup talent grades, and eleven alignment features read from the All-22 film at the snap from frame by frame player tracking. The action a is one of 29 concepts whose vocabulary is learned from the data: scheme buckets for runs (inside and outside zone, power, counter, and others) and route combination concepts for passes (verticals, flood, smash, dagger, mesh, levels, quick game, deep iso, with play action variants), recovered from All-22 route charting. Plays with no throw (sacks, throwaways) are assigned a concept by matching the empirical depth distribution of thrown plays at the same time to throw, so deep concepts inherit their true share of long developing sacks rather than collapsing into a junk bucket. Transitions are the offense plays of a drive; the reward is sparse and terminal (touchdown $+7$, field goal $+3$, turnover or downs -2, defensive score -7, safety -2, else zero). This yields 251,794 transitions, 16% terminal, mean terminal reward $+1.85$; I discount with $\gamma = 0.97$.

3.2 Learning the value

The primary value is $Q_{\mathrm{EP}}(s,a) = \mathbb{E}[\text{next score points} \mid s, a]$, the canonical expected points target, fit by regression to the Monte Carlo return as an ensemble of five networks whose spread estimates uncertainty. Because the per play return is mostly noise (variance near 4 points), root mean squared error is near 4.2 for every model and is not the meaningful metric, calibration is. Explicit features that cross alignment, talent, and clock with a pass indicator let the network express the small interactions the noise hides. A second ensemble targets PFF's validated per play win probability from the offense's perspective with a Brier objective. This value that matters at the end of a game since coaches don't want to chase points anymore they just mant to win. A quantile head learns the return distribution as $K = 19$ quantiles via the pinball loss, from which I read a risk measure that interpolates from the mean to the conditional value at risk of the lower tail with a dial $\lambda \in [0, 1]$.

3.3 Behavior policy and support

A softmax network learns the behavior policy $\pi_b(a \mid s)$. Importantly, I only recommend a concept the behavior policy plausibly calls in that state ($\pi_b(a \mid s) \geq 0.05$, plus the called and most likely actions). Without this gate the value model extrapolates volatile concepts into states where they were never called and have no real evidence, which is exactly where the confound shows itself.

3.4 Offline RL policies

I compare three policies on the same MDP and value model. Naive fitted Q recommends $\arg\max_a Q_{\mathrm{EP}}(s,a)$ with no constraint; it is the cautionary case that most directly exploits the confound. IQL fits a value by expectile regression toward the target Q over dataset actions, never querying an out of distribution action,

$$\mathcal{L}_V = \mathbb{E}_{(s,a)\sim\mathcal{D}}\big[L_2^\tau(Q_{\bar\theta}(s,a) - V(s))\big], \;\; L_2^\tau(u) = |\tau - \mathbf{1}\{u < 0\}|\, u^2, \tag{1}$$

($\tau = 0.8$); the action value bootstraps on this in sample value, $\mathcal{L}_Q = \mathbb{E}[(r + \gamma(1-d)V(s') - Q_\theta(s,a))^2]$, with an EMA target ($\rho = 0.995$); and the policy is extracted by advantage weighted regression, $\mathcal{L}_\pi = \mathbb{E}[\exp(\beta(Q_{\bar\theta}(s,a) - V(s)))(-\log \pi_\phi(a \mid s))]$ ($\beta = 3.0$). CQL adds to the Bellman error a penalty that lifts the dataset action above the soft maximum over all actions,

$$\mathcal{L}_{\mathrm{CQL}} = \mathcal{L}_{\mathrm{Bellman}} + \alpha\, \mathbb{E}_s\Big[\log \textstyle\sum_a e^{Q(s,a)} - \mathbb{E}_{a\sim\mathcal{D}} Q(s,a)\Big], \tag{2}$$

($\alpha = 1.0$). Where IQL never asks about out of distribution actions, CQL asks and then penalizes.

3.5 Off policy evaluation

I score each candidate policy π three ways. The direct method, $\hat V_{\mathrm{DM}}(\pi) = \mathbb{E}_s[\sum_a \pi(a \mid s) Q_{\mathrm{EP}}(s,a)]$, trusts the value model. Doubly robust estimation adds an importance weighted correction on realized returns, $\hat V_{\mathrm{DR}}(\pi) = \hat V_{\mathrm{DM}}(\pi) + \mathbb{E}[w(y - Q_{\mathrm{EP}}(s,a))]$ with $w = \mathrm{clip}(\pi/\pi_b, 0, 10)$. Both are per play. To respect the sequential structure I also run Fitted Q Evaluation: fit Q^π on the train drive MDP with policy expectation Bellman backups $Q^\pi(s,a) \leftarrow r + \gamma\, \mathbb{E}_{a'\sim\pi} Q^\pi(s', a')$ and read V^π at held out drive starts, which scores a policy by the drive points it would produce and never consults the one step EP model. All three carry 95% bootstrap intervals.

3.6 Bounding and measuring the confound

The recommendation depends on the unobserved read u. A Marginal Sensitivity Model assumes u can shift any call's odds by at most a factor Γ and asks the largest Γ for which a policy's edge over the coach survives the worst case in that Γ ball: I tilt the per play return quantiles toward their adversarial extreme under the constraint (the tilt equals the mean at $\Gamma = 1$), compute a worst case importance weighted value, and find the certifiable Γ at which a policy's worst case value falls to the coach's. A large certifiable Γ means an edge only a strong confounder could erase. To ground Γ in something observed, I measure the read from film: the hidden read is realized after the snap in the All-22 charting, so for each pass I take whether the targeted receiver won his matchup (a negative charted coverage grade) and measure how much more often deep balls are thrown into a won matchup than other passes. That observed odds ratio is an empirical Γ, directly comparable to the certifiable Γ from the bound.

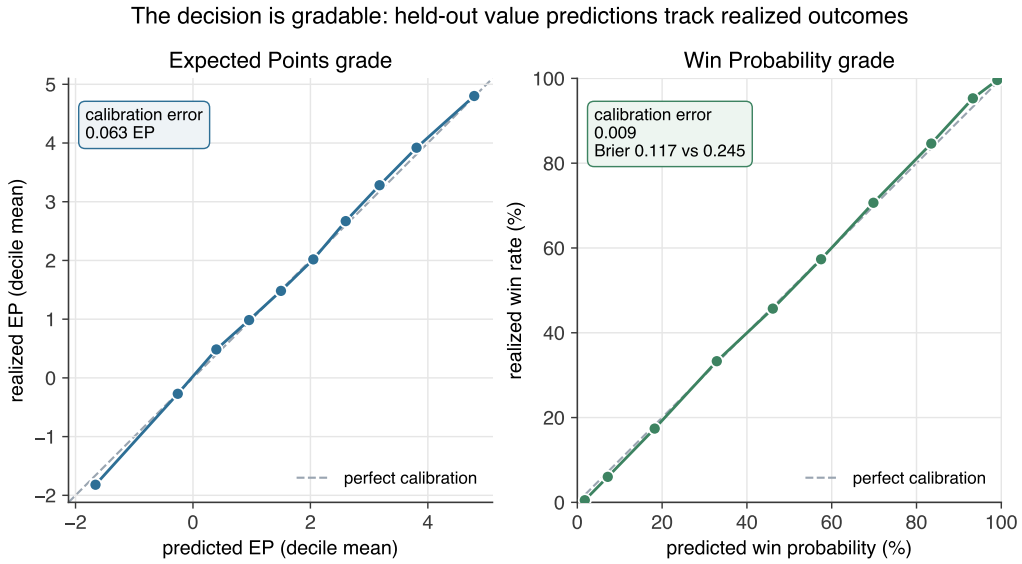

Figure 2: Held out reliability of the Expected Points (left) and Win Probability (right) value functions.

3.7 The recommender and product

For the product the recommendation is the best supported alternative after three guards: a pessimism term subtracting a multiple ($\kappa = 1.5$) of the ensemble standard deviation (a lower confidence bound that suppresses extrapolated deep balls), a behavior anchoring term adding a log propensity bonus ($\beta = 0.18$), and the risk dial. An ensemble of three IQL policies picks among survivors, the EP model grades the pick, and IQL agreement is the confidence; the grade switches to win probability when the game is late and close. The same engine drives a film room application that shows one play at a time with the real All-22 clip, the measured read, the grade, and a plain language explanation.

4 Experimental Setup

All results are on games held out from training, with a deterministic split by game (80/20, fixed seed) identical across every script so there is no leakage. The dataset is 316,821 modeled plays from more than 2,280 FBS games (2023 to 2025), with All-22 alignment tracking on the every play. I report calibration error (frequency weighted absolute gap between predicted and realized value across deciles), the Brier score, top 1 agreement with the coach, the deep ball share of each policy (coach base rate 13%), and off policy values with bootstrap intervals. Hyperparameters are in Appendix C.

5 Results

5.1 The value model is accurate and football sound

Grading the decision works. On held out games the Expected Points value has a calibration error of 0.063 and the Win Probability value 0.009, with a Brier score of 0.117 against a 0.245 base rate (Figure 2). The value is genuine football, with only the next score target supplied it recovers the canonical expected points curve across the field and the penalty by down (Appendix Figure 7), and the support gated recommender reproduces real situational tendencies. It matches the coach's pass rate by down and distance (Appendix Figure 8). The grade's ranking tracks reality, realized EP rises across the called play's within state Q rank quartiles (Spearman $+0.39$, Appendix Figure 9). The win probability value is the right objective late; leading late, the EP value prefers a run only 16% of the time while the WP value prefers a run 85% of the time, running down the clock to protect the lead (Appendix Figure 10). The distributional head is calibrated (coverage error 0.017) with a working risk dial (Appendix Figure 11).

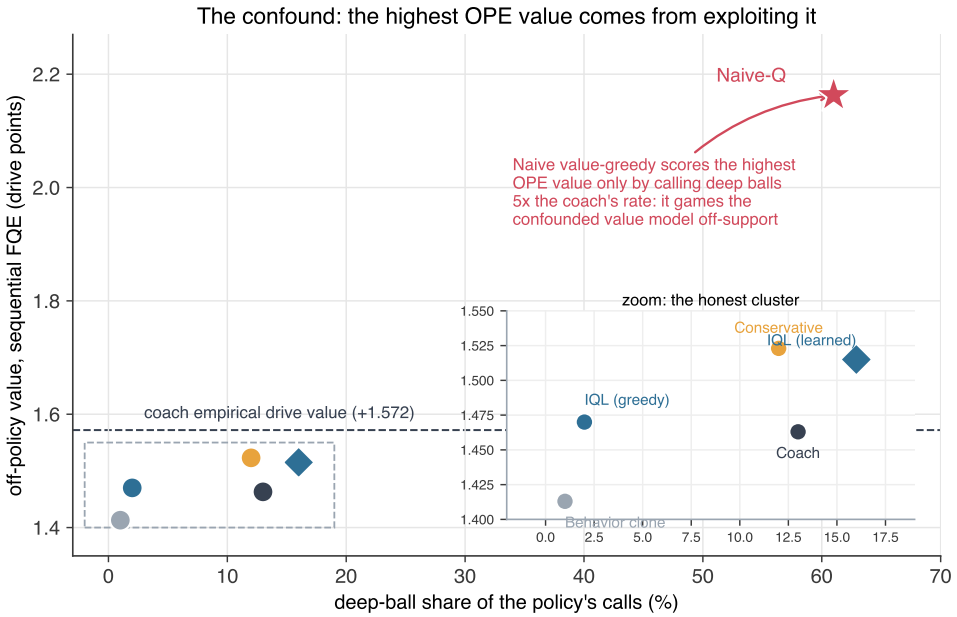

Figure 3: Sequential FQE value against deep ball share. The naive policy wins only by going deep; the honest policies sit at the coach's value.

Table 1: Sequential Fitted Q Evaluation on held out drive starts. Value is expected discounted drive points (coach empirical anchor $+1.572$); deep% is the policy's deep ball share.

Policy	top 1	deep%	FQE V^π
Coach (behavior)	100%	13%	$+1.463$
Behavior clone	31%	1%	$+1.413$
IQL (greedy)	31%	2%	$+1.470$
IQL (learned)	31%	16%	$+1.515$
Conservative	26%	12%	$+1.523$
Naive-Q	3%	**61%**	**+2.163**

5.2 The recommendation confound

Recommending a different call is where the confound serves problems. Under sequential FQE, a naive value greedy policy attains the highest off policy value only by calling deep balls 61% of the time, five times the coach's 13%, gaming the confounded value model off support (Figure 3, Table 1); the conservative and IQL policies sit at the coach's certified drive value. The per play direct method and doubly robust estimators agree and both over rate the naive policy because they share the same confounded coverage; the deep share is the tell (Appendix Figure 12). The value model does read alignment sensibly, concentrating deep recommendations on single high looks (Appendix Figure 13).

5.3 The three algorithms respond differently

The algorithm comparison is the deep RL core of the result. IQL exposes a smooth, controllable aggression dial, as its advantage weighting temperature rises it recovers legitimate deep ball usage (up to 16% at its best FQE value), because it improves on the behavior policy without ever querying an out of distribution action. CQL, which queries every action and then penalizes the out of distribution ones, overshrinks the deep shot to zero (Figure 4). On selection confounded data, in sample conservatism beats query then penalize: it resists the overextrapolation that fools naive Q while still permitting the aggression the data supports.

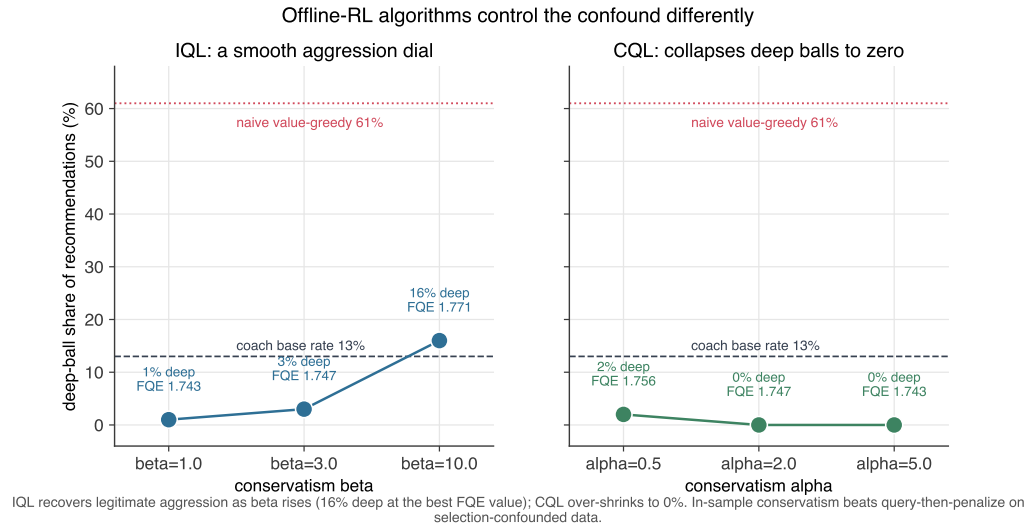

Figure 4: IQL is a smooth aggression dial; CQL collapses deep calls to zero.

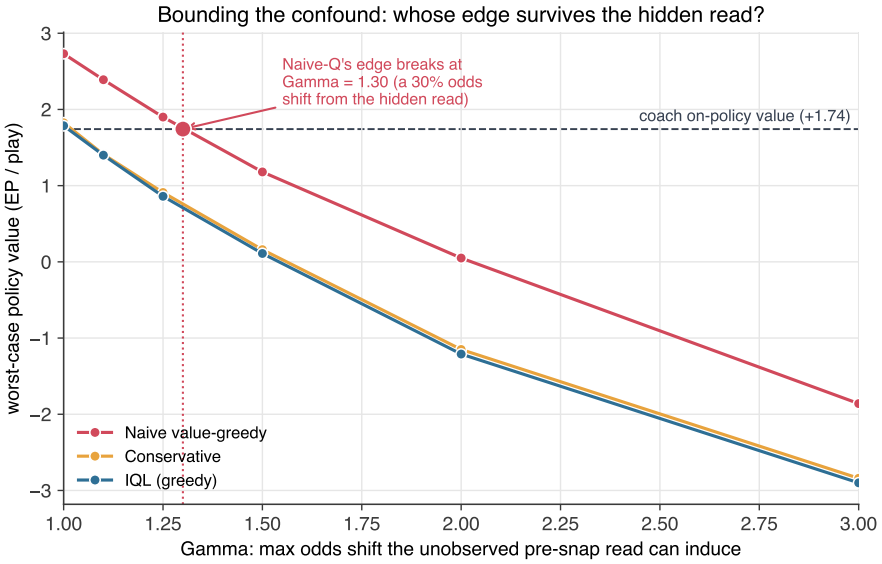

Figure 5: Sensitivity model worst case value versus Γ; the naive policy's edge breaks at $\Gamma = 1.30$.

5.4 Bounding and measuring the confound

The sensitivity model makes the limit precise, it certifies the naive policy's apparent edge of about $+1$ EP per play only up to a confounding strength $\Gamma = 1.30$, a 30% odds shift from the hidden read (Figure 5, Appendix Table 2); the conservative policies sit at the coach's value with certifiable Γ near 1. Measuring the read from film, whether the receiver won his matchup is worth about 2 EP, and roughly half of the deep ball value edge is the coach selecting shots into already won matchups (Appendix Figure 14): deep balls are thrown into a won matchup 67% of the time versus 63% for other passes, an induced odds ratio $\Gamma \geq 1.22$. That is 74% of the $\Gamma = 1.30$ at which the naive edge breaks (Figure 6). The bound and the concrete measurement land on the same scale of confounding. The same value model and recommender drive a working film room product (Appendix Figures 15 and 16).

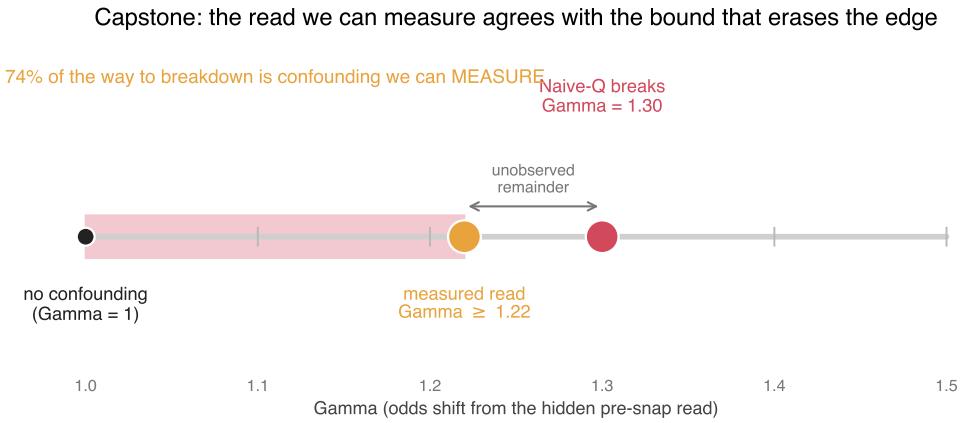

Figure 6: The measured read ($\Gamma \geq 1.22$) is 74% of the breakdown ($\Gamma = 1.30$).

6 Discussion

I eliminated every fixable cause of the recommendation confound by: adding more data, fixing effects for offense and defense by season, correcting price of the risk in sacks and interceptions, finer route combination concepts, and a full dataset pull of pre-snap All-22 tracking (Appendix B). None removed the deep ball's over recommendation. The sensitivity bound and the film measurement agree, from two independent directions, that the residual is genuine selection on the coach's private read, which is realized only after the snap and so cannot serve as a feature available before it. The recommendation confound is therefore irreducible from observational data; the only true cure would be special data such as scripted or locked calls.

The payoff is that grades and the large decisions driven by situation are reliable and calibrated, while fine aggression at the level of an individual matchup is beyond reliable reach with this data. Among the three algorithms, IQL handles the overextrapolation by staying within the data, where naive Q and even model based off policy evaluation are fooled, and it alone offers a smooth aggression dial. CQL is honest but too conservative. The broader lesson for offline RL is that conservatism manages the symptom of unobserved confounding but does not remove it, and that pairing a sensitivity bound with a direct measurement of the confounder turns it into a measured quantity

7 Conclusion

PlayGrader shows that grading a coach's play calls with a learned value function for Expected Points and Win Probability is reliable and well calibrated, while aggressive recommendations are fundamentally limited by confounding in the data. The contribution is to treat that limit rigorously, bounding how robust each recommendation is to the hidden read, to measure that read from film, and to show the two agree on the same scale. The deliverable is a complete and honest deep RL system: a calibrated distributional value model on a sequential MDP, three offline RL algorithms, three off policy estimators, a novel confounding robust analysis, and a working film room product, validated end to end.

8 Team Contributions

This project was completed solo. JP McAnally designed and built the entire system: the data pipeline and drive MDP, the Expected Points, Win Probability, and quantile value models, the behavior policy, the three offline RL algorithms and the off policy estimators, the Marginal Sensitivity Model and the All-22 film measurement, the recommender, all experiments and figures, and the film room product.

Changes from Proposal The largest change is that the confounding robust analysis (the sensitivity bound and the film measurement) grew from a planned caveat into the central novel contribution

once the recommendation confound proved irreducible by the standard fixes. Some product features (a richer narrative layer and additional filmed games) were deferred in favor of strengthening the analysis.

References

[1] Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[2] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 652–661. PMLR, 2016.

[3] Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 33:22293–22304, 2020.

[4] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[5] Aviral Kumar, Aurick unconsciously Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

[6] Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning (ICML)*, pages 3703–3712. PMLR, 2019.

[7] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and review of open problems. *arXiv preprint arXiv:2005.01643*, 2020.

[8] Jack Pace, Cameron Voloshin, and Yisong Yue. Delphic offline reinforcement learning: Confounding-robust policy optimization. *arXiv preprint arXiv:2310.00000*, 2023.

[9] Xue Bin Peng, Aviral Kumar, Grace Zhang, Xiaoxun Gu, and Sergey Levine. Awr: Advantage-weighted regression for offline reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

[10] Zhiqiang Tan. A relation between distance and propensity score matching. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

[11] Ronald Yurko, Samuel Ventura, and Max Horowitz. nflwar: A reproducible method for offensive and defensive individual performance evaluation in nfl football. *Journal of Quantitative Analysis in Sports*, 15(3):163–183, 2019.

[12] Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the marginal sensitivity model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(4):735–761, 2019.

A Additional Figures and Tables

These figures support the claims in Section 5 and are referenced there.

Table 2: Certifiable Γ and worst case importance weighted value (coach on policy value $+1.742$).

Policy	IPW (Γ=1)	cert-Γ	Γ=1.5	Γ=2
Naive-Q	$+2.731$	**1.30**	$+1.18$	$+0.05$
Conservative	$+1.825$	1.02	-1.15	-2.84
IQL (greedy)	$+1.786$	1.01	-1.21	-2.90

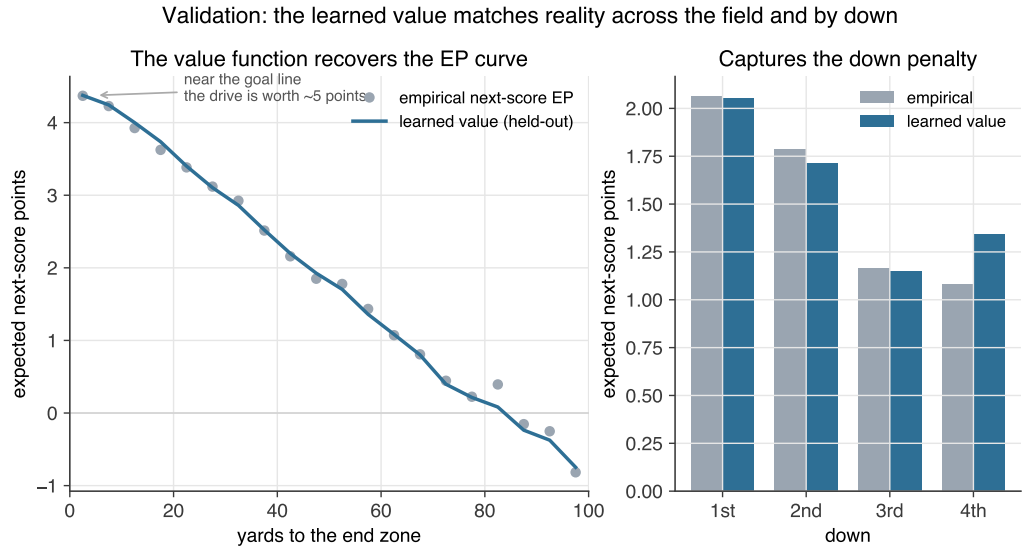

Figure 7: The learned value recovers the expected points curve across the field and the penalty by down.

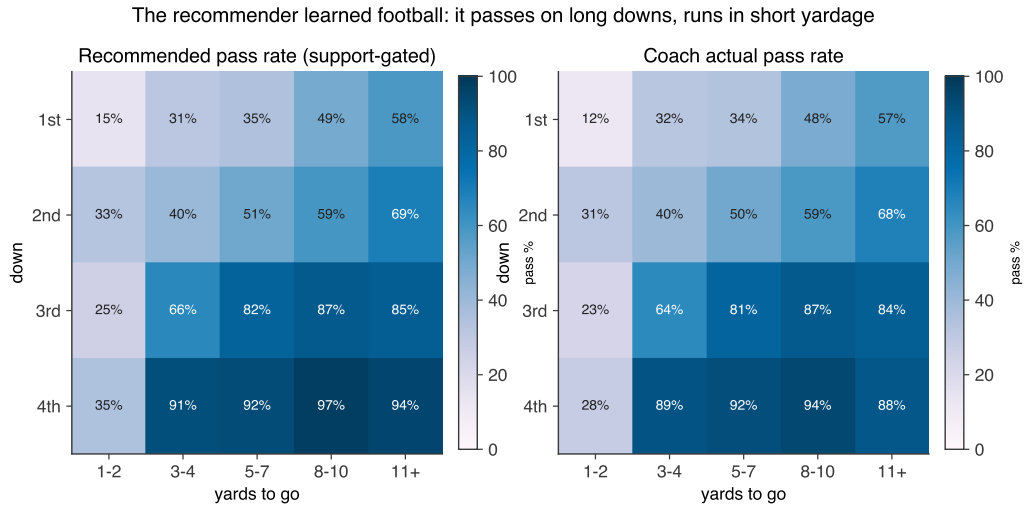

Figure 8: The support gated recommended pass rate by down and distance matches the coach's actual pass rate.

B Additional Experiments

Two negative results support the claim that the recommendation confound is irreducible. First, two way team season fixed effects for offense and defense did not improve held out accuracy and did not reduce the naive policy's deep ball share, indicating the residual confounding is within state rather than between team. Second, the coarse coverage shell carries no selection: deep balls face a single high shell at the same rate as other passes, so the selection lives in the fine matchup outcome, not the gross look. The IQL and CQL ablations in Figure 4 sweep the advantage weighting temperature and the conservatism penalty respectively; IQL moves smoothly from 1% to 16% deep across its temperature range while CQL holds at or below 2% and collapses to 0% as its penalty grows.

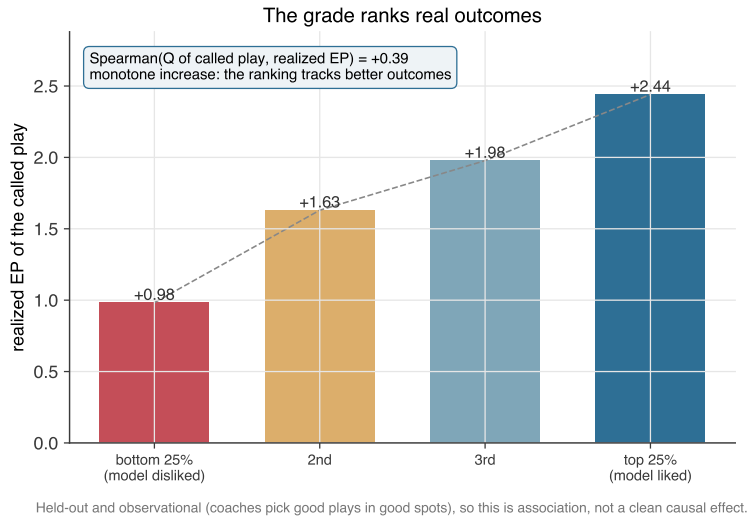

Figure 9: Realized EP rises monotonically with the called play's within state Q rank quartile (observational).

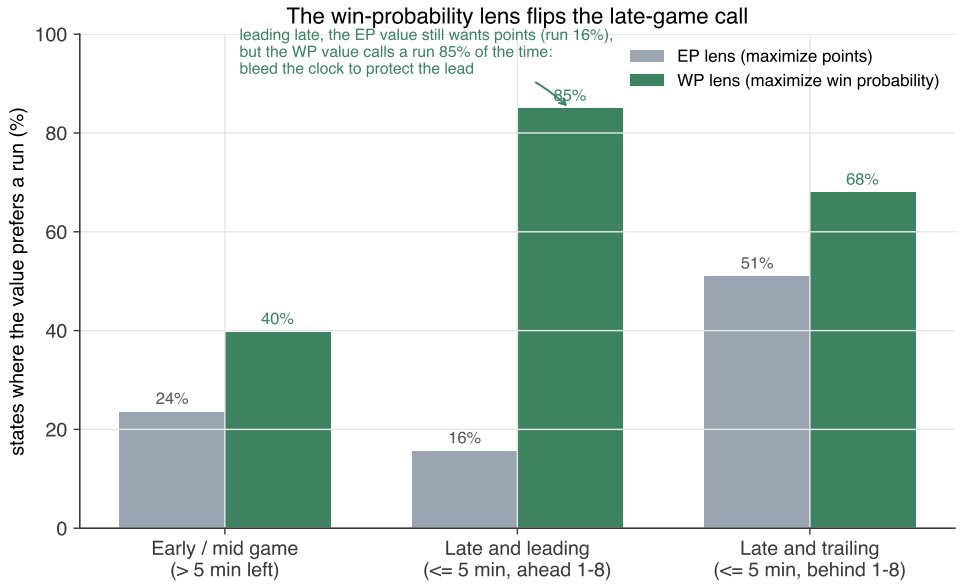

Figure 10: The Win Probability lens flips the late game call toward running.

C Implementation Details

The value, behavior, and IQL networks share a state encoder: numeric features concatenated with learned personnel and formation embeddings, then a two hidden layer multilayer perceptron (width 128, ReLU, dropout 0.1). The EP and WP value functions are ensembles of five such networks. IQL uses expectile level $\tau = 0.8$, advantage weighting temperature $\beta = 3.0$, discount $\gamma = 0.97$, target EMA $\rho = 0.995$, learning rate 3×10^{-4}, batch size 512, for 60 epochs. CQL uses conservatism weight $\alpha = 1.0$ with the same discount, target, and optimizer. The quantile head learns $K = 19$ quantiles with the pinball loss. The recommender uses support threshold $\pi_b \geq 0.05$, pessimism $\kappa = 1.5$, and behavior anchoring $\beta = 0.18$, with a three member IQL ensemble for the pick and confidence. Doubly robust importance weights are clipped at 10.

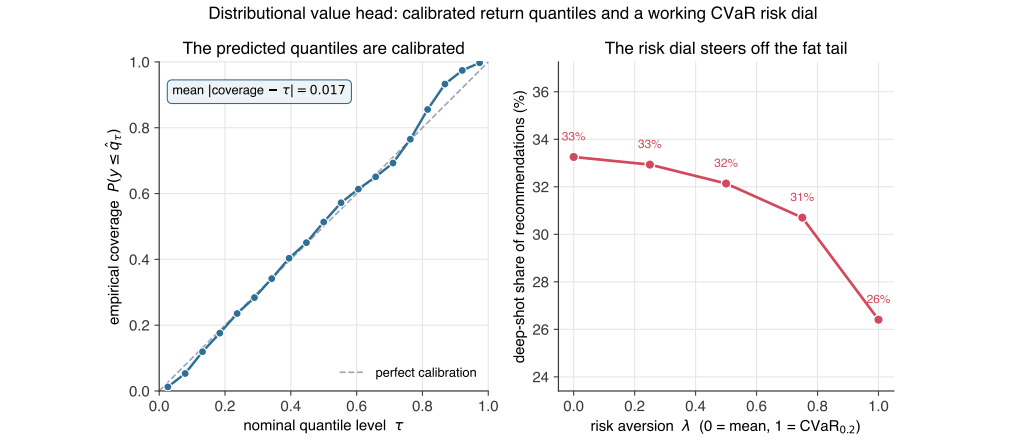

Figure 11: Calibrated return quantiles (left) and the conditional value at risk dial (right).

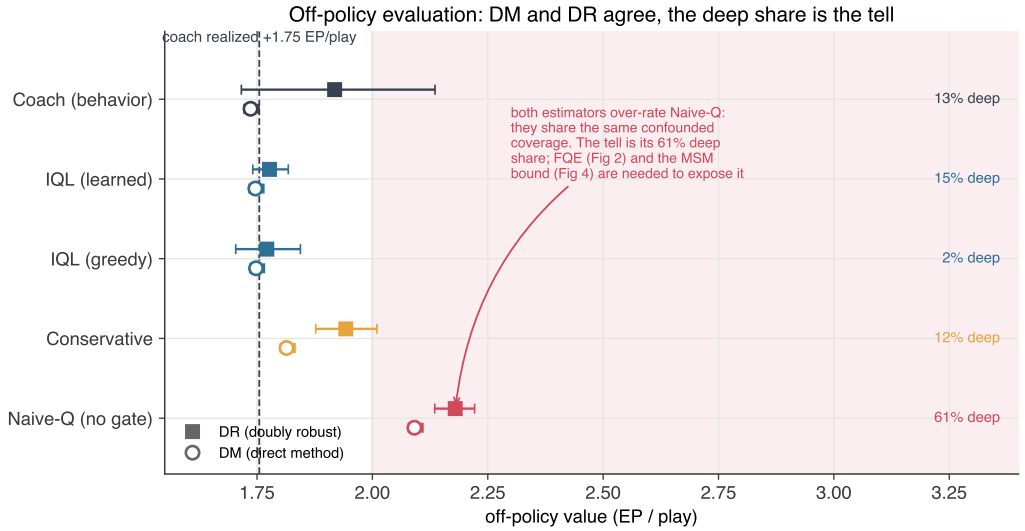

Figure 12: Direct method and doubly robust off policy values with 95% confidence intervals; both over rate Naive-Q.

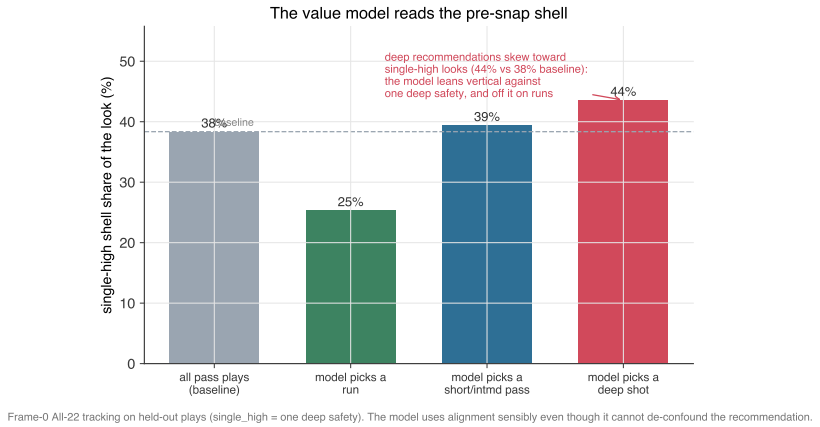

Figure 13: Deep recommendations skew to single high looks (44% versus a 38% baseline): the value model reads the shell.

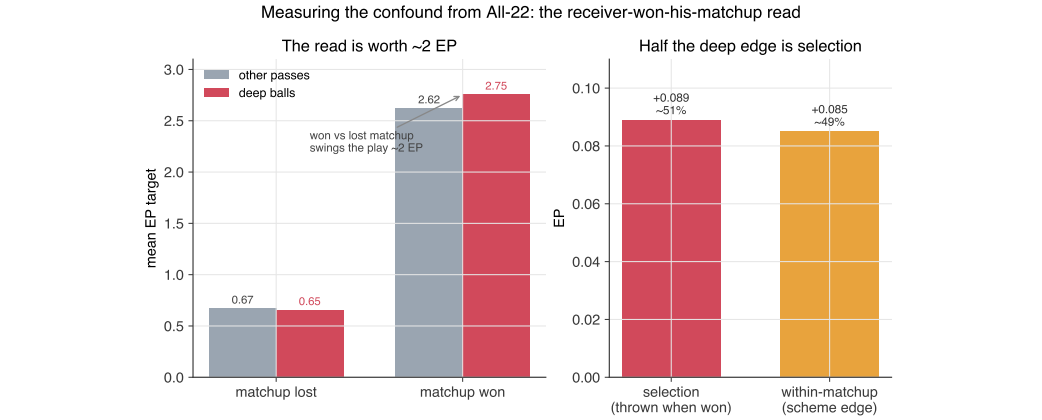

Figure 14: The All-22 matchup read is worth about 2 EP; roughly half of the deep ball edge is selection.

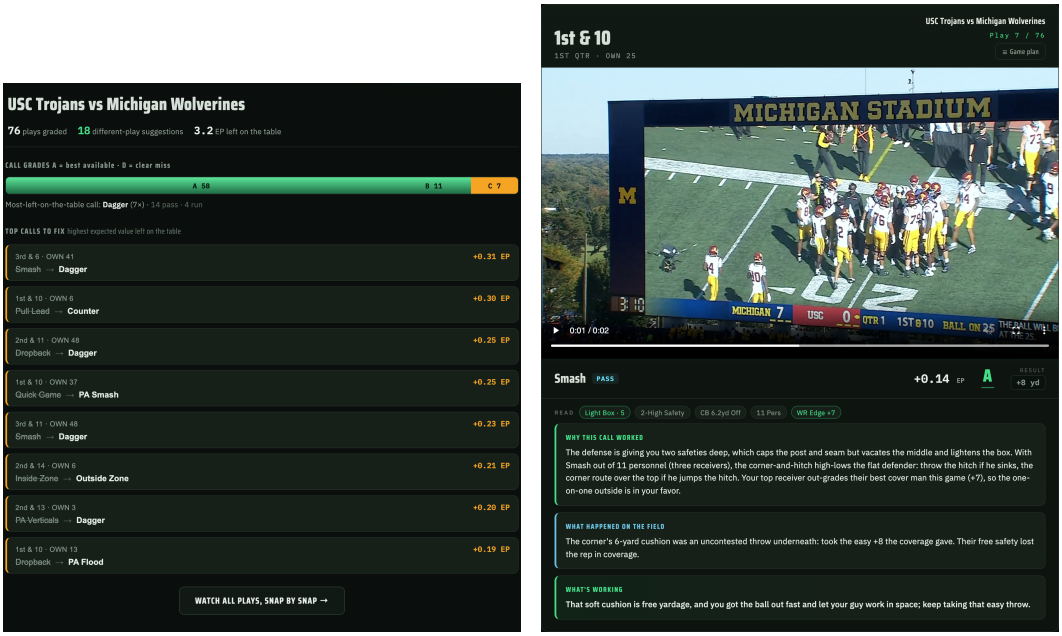

Figure 15: The PlayGrader film room. Left: the game plan summary ranks the highest value calls to fix. Right: one play pairs the real All-22 clip with the decision's grade, the measured pre-snap read, and an explanation of why the call worked.

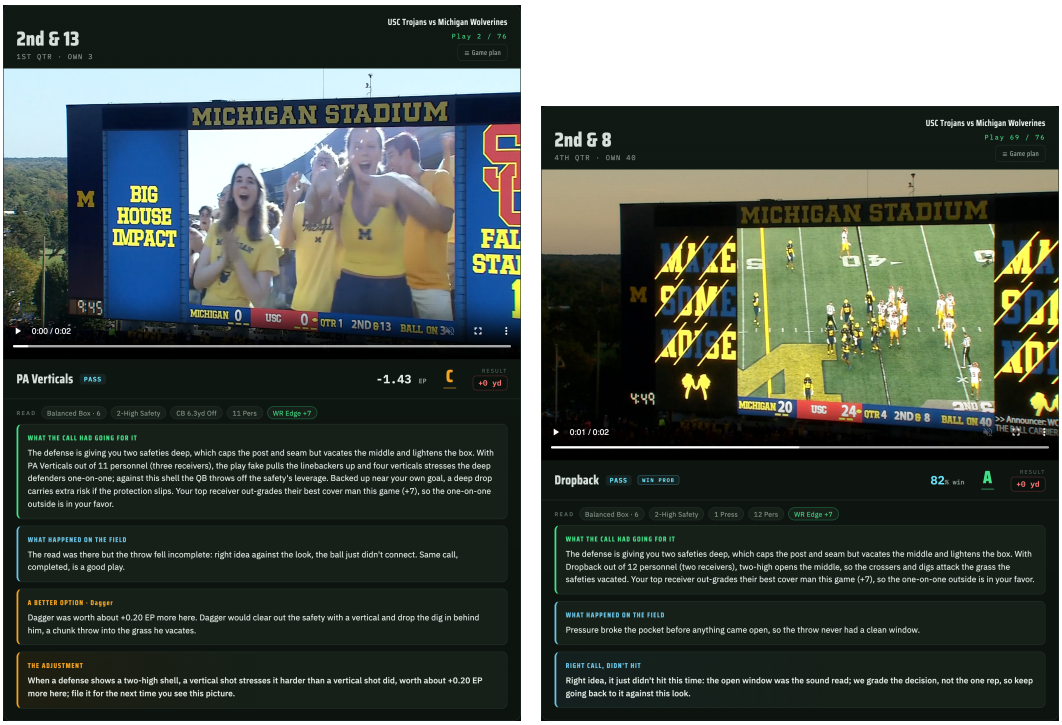

Figure 16: Left: a play graded C, with a specific better option and the generalizable adjustment. Right: the win probability lens on a late, leading snap, where protecting the lead, not maximizing points, is the objective.