Tackling Causality: Estimating Frame-Level Defensive Impact with Multi-Agent Transformers

Football Track
Paper ID 95

1. Introduction

Evaluating defensive performance in football remains one of the hardest open problems in sports analytics. While offensive value is captured through measurable outcomes such as yards gained, completion probabilities, or expected points added, defensive value is far more elusive. Most of a defender's impact occurs in the invisible domain of prevention, such as closing space, influencing angles, disrupting reads, and forcing suboptimal decisions that never appear in the box score. When a tackle finally occurs, traditional metrics assign credit solely to the tackler while ignoring the coordinated movements that created the opportunity in the first place.

This creates a fundamental attribution problem. Defensive plays unfold as tightly coordinated, multi-agent systems where individual contributions are interdependent. A linebacker's tackle may depend critically on a defensive end compressing the edge, a safety's pursuit angle may be determined by cornerback leverage, and a missed tackle may force substitutions in responsibility as multiple defenders converge. These interactions violate the independence assumptions underlying nearly all current defensive metrics. As a result, tackle counts and correlational models systematically misrepresent player value, often rewarding high-volume tacklers while overlooking defenders whose contributions are expressed indirectly through coordination.

Causal inference provides a natural framework for answering the central question facing NFL front offices: What is a defender's true impact on the outcome of a play, holding everything else equal? However, classical causal methods assume the Stable Unit Treatment Value Assumption (SUTVA), which requires no interference between units, and this assumption is fundamentally incompatible with how football defense operates. A defender's action directly alters the opportunities, responsibilities, and outcomes of teammates. Ignoring these interference effects introduces measurable bias into value estimates and obscures the contributions that teams seek to quantify.

This work introduces a causal inference framework explicitly designed for the realities of football defense. Using player tracking data from the 2022 NFL season, we build a multi-agent Transformer architecture that models all 22 players jointly, captures time-varying coordination patterns, and estimates counterfactual outcomes at the play level. The framework incorporates adversarial balancing to ensure fair comparison between tacklers and non-tacklers, and it uses doubly robust estimation and Bayesian uncertainty quantification to produce reliable and interpretable treatment effects.

Our approach produces Expected Points Saved (EPS), which is a causal estimate of how many expected points a defender prevents by making a tackle, accounting for coordination, interference, and the defensive context surrounding each play. We show that interference effects are widespread, affecting 68 percent of plays, and that ignoring these effects biases player evaluation by an average of 0.084 expected points per tackle. Through real-game analysis, synthetic experiments with ground-truth causal effects, and extensive robustness testing, we demonstrate that this framework recovers individual defensive value in ways that traditional metrics cannot.

By directly modeling the collective and interdependent nature of football defense, this work provides a principled path toward causal player evaluation. The framework reveals hidden contributions, corrects long-standing biases in tackle-based metrics, and offers teams a more accurate foundation for roster construction and strategic decision-making.

1.1 Background and Motivation

American football presents unique challenges for performance evaluation, especially when assessing the contributions of defensive players. Offensive value is often visible in direct outcomes such as yards gained or touchdowns. Defensive value, however, is distributed across multiple players, depends on coordinated movement, and is often expressed through actions that prevent events from occurring. This makes defensive impact difficult to isolate and quantify.

Traditional defensive metrics suffer from four limitations that obscure individual value creation:

Attribution Problem: A tackle that appears to be the result of one player's effort may depend critically on the positioning and actions of several teammates.

Counterfactual Nature: Many of the most valuable defensive contributions involve preventing events, such as closing passing lanes or forcing unfavorable ballcarrier paths, which do not appear in the play-by-play record.

Temporal Complexity: Defensive impact unfolds throughout the play, not only at the moment of contact. Angles, leverage, and pursuit dynamics evolve continuously.

Interference Effects (SUTVA Violations): A defender's actions directly influence the opportunities and responsibilities of teammates. This interdependence violates the independence assumptions behind most traditional and statistical evaluation methods.

Together, these challenges motivate the need for methods that move beyond simple counts or correlational models. A more rigorous framework must capture hidden contributions, account for interdependence, and estimate the true causal value of individual defenders.

1.2 Causal Inference in Sports

Causal inference has expanded in sports analytics, driven by high-resolution tracking data and advances in machine learning. These methods have provided new insight into offensive decision-making, play selection, and strategy. Defensive performance, however, is still commonly evaluated through aggregated outcomes or individual tackle counts.

Such approaches miss a central fact about defense: contributions are both individual and interdependent. One player's movement often shapes the opportunities and effectiveness of others. Defensive pressure can funnel ballcarriers into specific gaps, alter route progressions, or change blocking responsibilities. Ignoring these interactions leads to biased or incomplete assessments.

Our work addresses this gap by developing a causal inference framework specifically tailored to football defense. The framework models all 22 players jointly and explicitly accounts for the coordination and interference patterns that define defensive play.

1.3 Technical Contributions

This paper introduces six technical contributions to sports analytics:

Multi-Agent Transformer Architecture.
A novel attention-based model that captures interactions between defenders, teammates, and opponents across time, and adjusts for time-varying confounding.

Interference-Aware Modeling.
A three-part representation of coordination, interference, and substitution that directly addresses the violation of independence in defensive play.

Representation Balancing.
A combination of gradient reversal layers, multi-scale adversarial training, and distributional matching that produces balanced representations for estimating treatment effects.

Uncertainty Quantification.
A Bayesian framework that separates epistemic uncertainty from aleatoric variability, enabling confidence intervals and probabilistic statements about defensive value.

Robustness Testing.
A suite of sensitivity analyses, including Rosenbaum bounds, E-values, and multiple placebo tests to validate causal credibility.

Temporal Coordination Modeling.
A characterization of how defensive coordination evolves throughout a play, identifying when interdependence is strongest and how it influences treatment effect estimation.

Together, these contributions create the first interference-aware, transformer-based causal inference framework for defensive evaluation in football.

2. Methodology

2.1 Problem Formulation

For each play, we observe:

- **Units:** individual defensive players i
- **Time:** sequences of T frames per play
- **Features:** vectors $X_{it} \in R^F$ describing player position, movement, relational structure, and game context at frame t

The outcome of interest is **Expected Points Saved (EPS)**, defined as the counterfactual difference in expected points if the defender does not make the tackle compared to if the defender does make the tackle:

$$\tau_{EPS}(X) = E[Y_0|X] - E[Y_1|X],$$

where:

- Y_0: Expected Points Added (EPA) if the defender does not make the tackle
- Y_1: Expected Points Added (EPA) if the defender does make the tackle

Positive values indicate that the defender prevents expected points relative to the counterfactual scenario.

Causal Structure.
Figure 1 illustrates the directed acyclic graph that guides identification. Player features, ballcarrier features, relational features, teammate features, opponent features, and game context influence both:

1. The probability that a defender becomes the tackler
2. The eventual EPA of the play

These variables form the confounding pathways that must be blocked to estimate the causal effect of interest, represented by the red arrow between **IsTackler** and **EPA**.

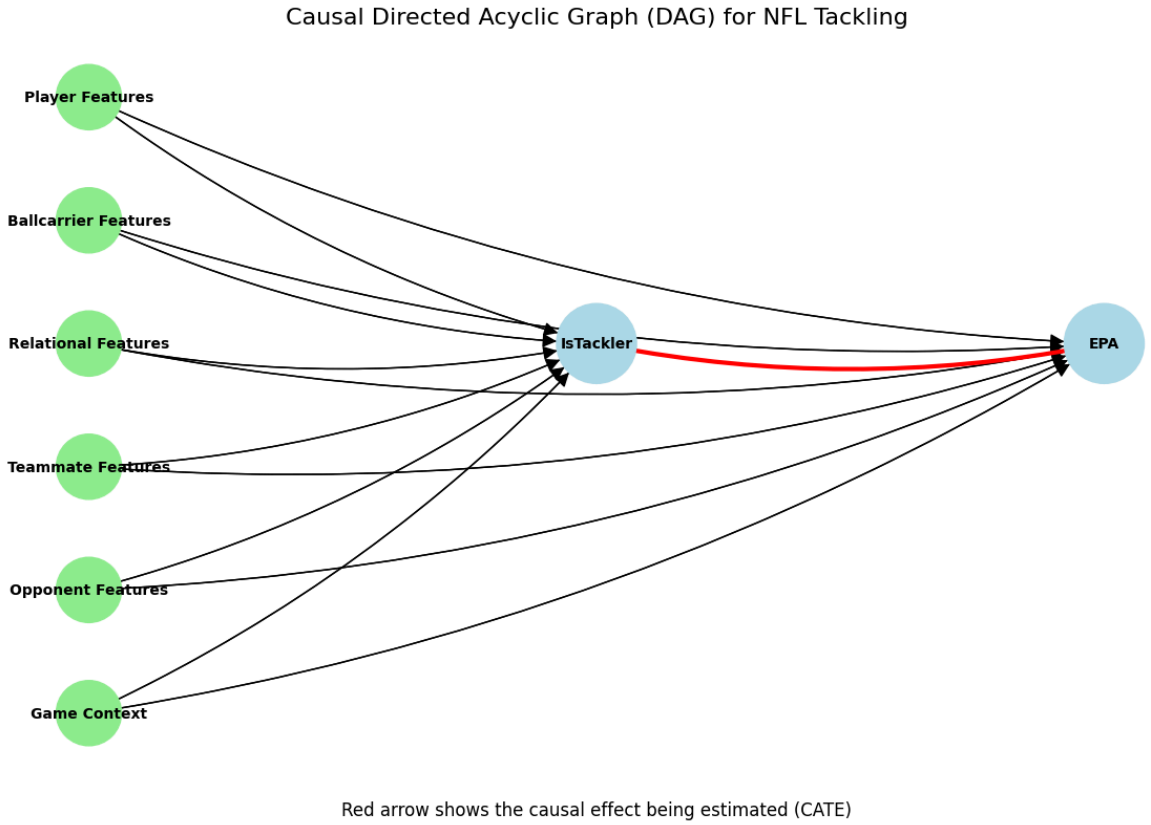


Figure 1. *Causal DAG for frame-level defensive impact estimation. The red arrow represents the causal effect of interest: the impact of a defender making a tackle on Expected Points Added (EPA).*

2.2 Multi-Agent Architecture and Interference Modeling

SUTVA Violations in Defensive Football

Classical causal inference requires the Stable Unit Treatment Value Assumption (SUTVA), which includes two principles. The first is that one unit's treatment does not affect another unit's outcome. The second is that treatment has a single consistent version for each unit. These assumptions often hold in medicine or economics. For example, one patient's drug intake does not alter another patient's results, and the formulation of a drug is stable.

Football defense violates both principles. Defenders operate as a coordinated unit, and one player's actions frequently alter the opportunities and effectiveness of teammates. A defensive end who forces a running back inside directly shapes the linebacker's chance to make the tackle. The nature of the treatment itself, the act of making a tackle, also varies by timing, location, leverage, and pre-snap or in-play context. A cornerback tackling after a large gain is fundamentally different from a linebacker tackling in the backfield.

Ignoring these violations produces biased causal estimates. A high tackle count might reflect favorable funneling created by teammates rather than individual excellence. Conversely, defenders who prevent plays from occurring at all, such as edge rushers who generate pressure that forces early throws, create value not captured by tackle outcomes.

Explicit Modeling of Interference

Our framework models how defenders influence one another through coordination, direct interference, and substitution. The multi-agent Transformer processes all 22 players simultaneously and uses hierarchical attention to capture spatial and temporal dependencies. For each defender i at time t, we estimate:

$$F_{i,t} = \sum_{j\neq i} C_{i,j,t} + I_{i,t} + S_{i,t} C_{i,j,t} + I_{i,t} + S_{i,t}$$

1. **Coordination Effects** ($C_{i,j,t}$):

   Defenders working together in real time, such as linebackers and safeties covering the same zone. Coordination strength declines exponentially with distance, reflecting how nearby defenders coordinate more intensively:

   $$C_{i,j,t} = exp\,(-\frac{\|x_{i,t}-x_{j,t}\|_2}{\lambda_c})\cdot\sigma\,(W_c^\top[E_{i,t} \oplus E_{j,t} \oplus f_{rel}(i,j,t)])$$

2. **Direct Interference** ($I_{i,t}$):

   One defender's action amplifying or diminishing another's impact. For example, when a defensive end forces a running back inside, the linebacker who makes the tackle benefits directly—the end's action interfered with (and improved) the linebacker’s effectiveness:

   $$I_{i,t} = \sum_{j\neq i} w_{ij,t} T_{j,t} (1 + \frac{\|x_{i,t}-x_{j,t}\|_2}{\lambda_I})^{-\alpha}$$

3. **Substitution Effects** ($S_{i,t}$):

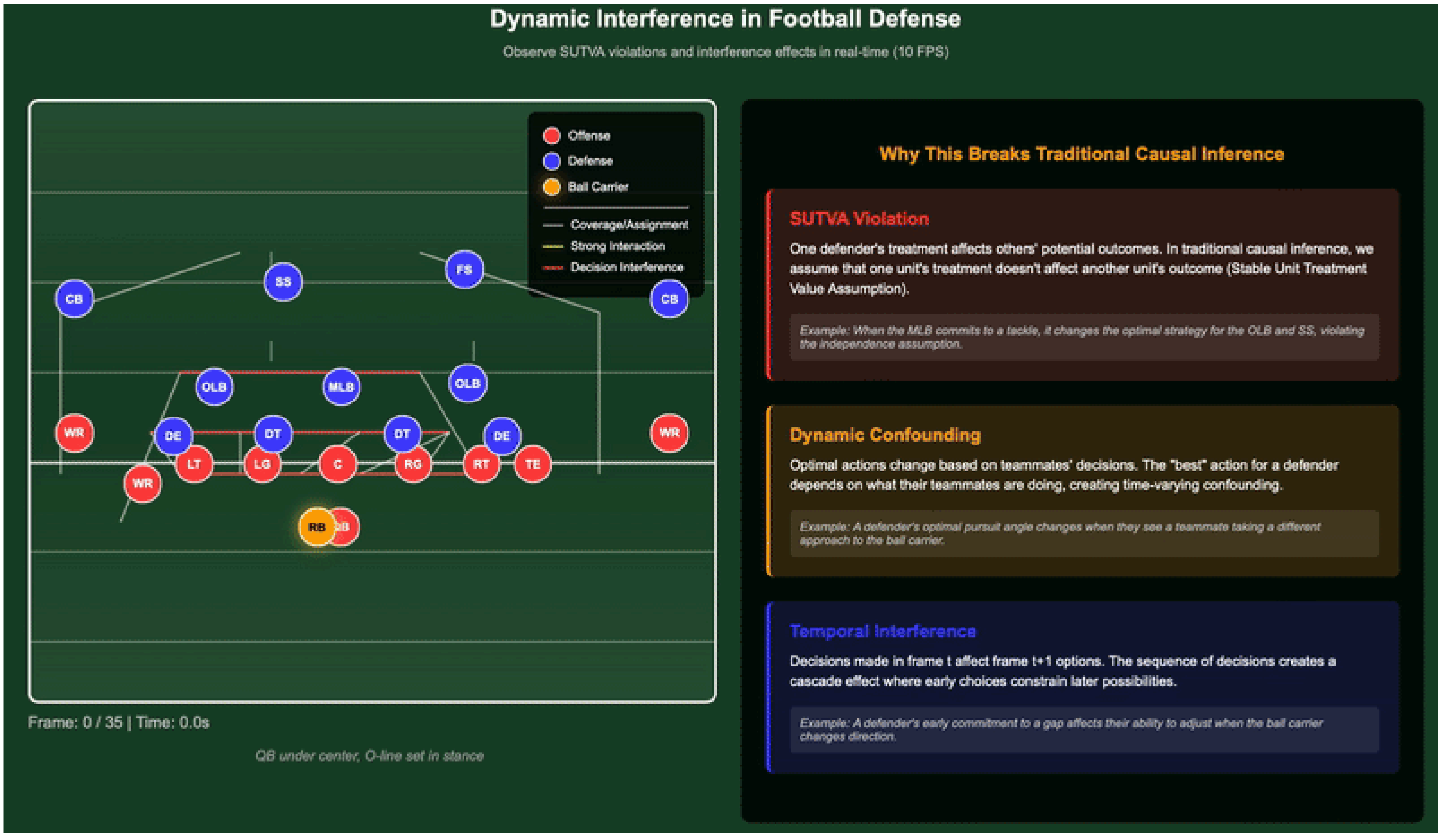
   Compensatory dynamics where defenders adjust responsibilities when teammates miss or abandon assignments. If one cornerback misses a tackle, a safety stepping in to make the stop exhibits substitution:

   $$S_{i,t} = \sigma\,(W_s^\top[E_{i,t} \oplus A_{i,t}])\cdot max_{j\neq i}(w_{ij,t} T_{j,t})$$

These functional forms reflect football domain knowledge while maintaining computational tractability.

Importance of Modeling Interference

By explicitly addressing SUTVA violations, our framework distinguishes between defenders who generate value independently and those who depend heavily on team-created opportunities. This distinction is essential for personnel evaluation. **Figure 2** shows how interference patterns evolve over the course of a play, illustrating why independence-based methods cannot capture the structure of defensive value.


Dynamic Interference in Football Defense
Observe SUTVA violations and interference effects in real-time (10 FPS)
Offense
Defense
Ball Carrier
Coverage/Assignment
Strong Interaction
Decision Interference
CB
SS
FS
CB
OLB
MLB
OLB
WR
DE
DT
DT
DE
WR
LT
LG
C
RG
RT
TE
WR
RB
QB
Frame: 0 / 35 | Time: 0.0s
QB under center, O-line set in stance

Why This Breaks Traditional Causal Inference

SUTVA Violation

One defender's treatment affects others' potential outcomes. In traditional causal inference, we assume that one unit's treatment doesn't affect another unit's outcome (Stable Unit Treatment Value Assumption).

Example: When the MLB commits to a tackle, it changes the optimal strategy for the OLB and SS, violating the independence assumption.

Dynamic Confounding

Optimal actions change based on teammates' decisions. The "best" action for a defender depends on what their teammates are doing, creating time-varying confounding.

Example: A defender's optimal pursuit angle changes when they see a teammate taking a different approach to the ball carrier.

Temporal Interference

Decisions made in frame t affect frame t+1 options. The sequence of decisions creates a cascade effect where early choices constrain later possibilities.

Example: A defender's early commitment to a gap affects their ability to adjust when the ball carrier changes direction.

Figure 2. *Dynamic interference in football defense. Coordination, interference, and substitution effects evolve across the play, violating standard independence assumptions required by traditional causal inference methods.*

2.3 Representation Balancing

Tacklers are not randomly assigned. To address this, we use multi-scale adversarial balancing. A discriminator attempts to infer treatment from learned representations, while the representation network seeks to make features informative for EPA prediction but uninformative for treatment assignment. This balancing occurs at the raw feature, interaction, and sequence levels, combined with distributional matching to create comparable treated and untreated groups.

2.4 Doubly Robust Estimation

We estimate Expected Points Saved (EPS) using Augmented Inverse Probability Weighting:

$$\hat{\tau}_{EPS} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{T_i(Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)} - \frac{(1-T_i)(Y_i - \hat{\mu}_0(X_i))}{1-\hat{e}(X_i)} + \hat{\mu}_0(X_i) - \hat{\mu}_1(X_i)\right]$$

where $\hat{e}(X_i)$ is the estimated propensity score, and $\hat{\mu}_1(X_i)$ $\hat{\mu}_0(X_i)$ are predicted EPAs under treatment and control. The doubly robust property ensures consistency if either the propensity model or the outcome model is correct.

2.5 Uncertainty Quantification and Temporal Modeling

We implement variational Bayesian neural networks by placing distributions over weights. This approach decomposes uncertainty into epistemic and aleatoric components. It enables statements such as: the defender prevents 0.15 ± 0.05 expected points with 95 percent confidence.

The model captures time-varying coordination patterns through the Transformer attention mechanism. Although treatment effects are computed at the play level, temporal variation in coordination is critical because early interactions influence downstream outcomes.

3. Experimental Design

3.1 Data and Preprocessing

We use the 2024 NFL Big Data Bowl dataset: 12,486 plays and 17,426 tackle attempts across Weeks 1 to 9 of the 2022 NFL Regular Season. Features include spatial trajectories, velocities, accelerations, orientations, and game context. We construct an extensive interaction feature set including distances between all pairs, pursuit angles, closing speeds, and contextual variables such as down, distance, and score.

Data splits prevent leakage: Weeks 1 to 6 for training, Weeks 7 to 8 for validation, and Week 9 for testing.

The evaluation throughout this paper emphasizes causal validity through propensity diagnostics, sensitivity tests, and placebo analyses.

4. Play Examples

Tackle Behind the Line of Scrimmage

In the play shown in **Figure 3**, Saquon Barkley (highlighted in yellow) takes a handoff and is tackled by Jeffery Simmons for a six-yard loss. Using our framework, we estimate each defender’s contribution on a frame-by-frame basis, evaluating what would have happened if they had or had not made the tackle.

The model correctly identifies Simmons as the primary contributor, with a peak **Expected Points Saved (EPS) of 0.66**. Importantly, it also highlights the hidden contributions of teammates that traditional statistics ignore. Interior defenders such as Tart and Weaver maintain consistently positive EPS values (0.25 to 0.35), reflecting their role in closing running lanes and forcing Barkley into Simmons’ path.

This example illustrates how our approach goes beyond simple tackle counts. By quantifying both the direct and indirect effects of defenders, it captures the coordinated dynamics of successful defensive plays, providing a richer and more accurate evaluation of individual value creation.

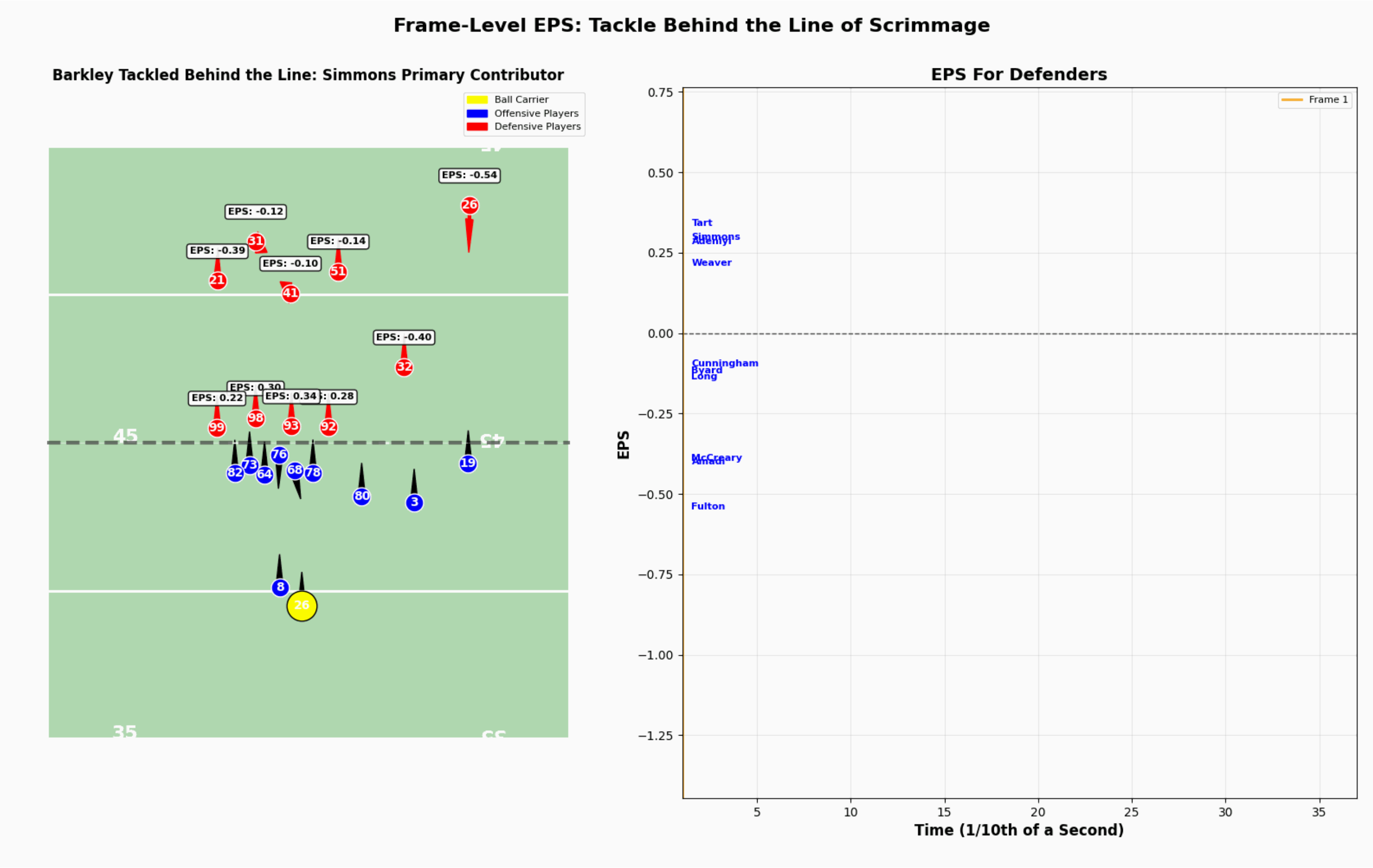


Figure 3. *Frame-by-frame EPS analysis of a running play, showing both the primary impact of the tackler and the supporting contributions of surrounding defenders.*

Missed Tackle

In the play shown in **Figure 4**, Russell Wilson completes a short pass to Javonte Williams, who breaks tackle attempts by Michael Jackson and Jordyn Brooks and gains nine yards before being stopped. Our framework captures defensive value even when tackles are missed, highlighting the cascading effects of failed attempts and the cumulative response of multiple defenders.

Jackson (**EPS** −0.15) and Brooks (**EPS** −0.35) register negative values, reflecting how their missed tackles directly increased the offense's advantage. The framework then tracks how the defense adapts, with additional defenders converging to contain the gain.

Barton emerges as the top positive contributor (**EPS** 0.40), demonstrating strong pursuit angles and effective positioning to prevent a larger gain. Other defenders vary in their impact: Nwosu remains near neutral (**EPS** 0.05), while Woolen (**EPS** −0.50), Coleman (**EPS** −0.60), and Diggs (**EPS** −0.90) accumulate increasingly negative values as the play extends and space opens for the offense.

This example shows how our approach uncovers both the immediate costs of missed tackles and the subsequent compensatory actions of teammates, offering a more complete view of defensive performance.

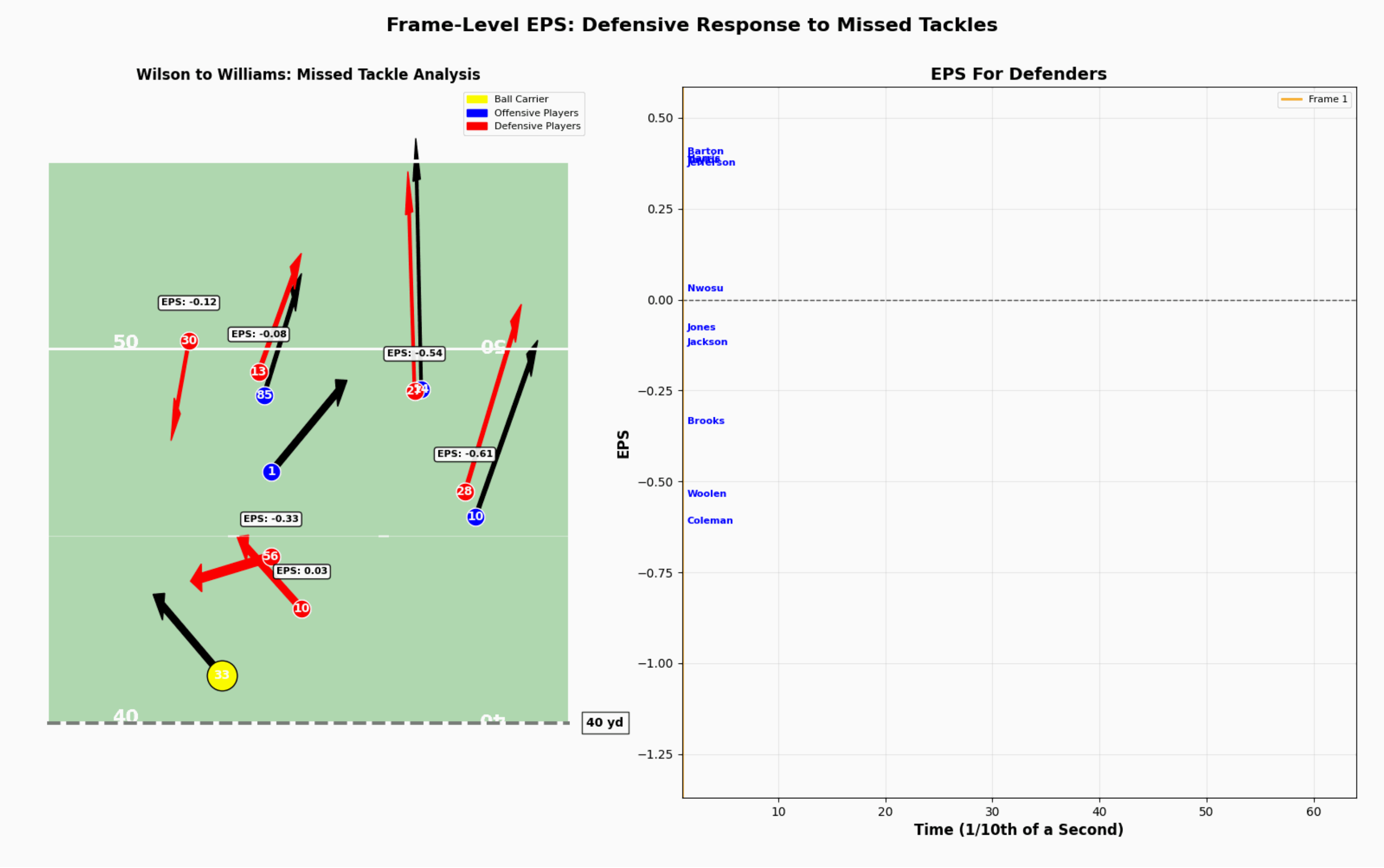


Figure 4. *Frame-level EPS analysis of a pass play, highlighting both the negative impact of missed tackles and the positive contributions of defenders who help limit the eventual gain.*

5. Results

All reported p-values are unadjusted; significance remains after Bonferroni correction (≈15 tests, threshold $p < 0.003$), with all major findings—interference prevalence, spatial decay, positional differences, temporal peaks, play-type dependencies, and treatment-effect heterogeneity—remaining highly significant ($p < 0.001$).

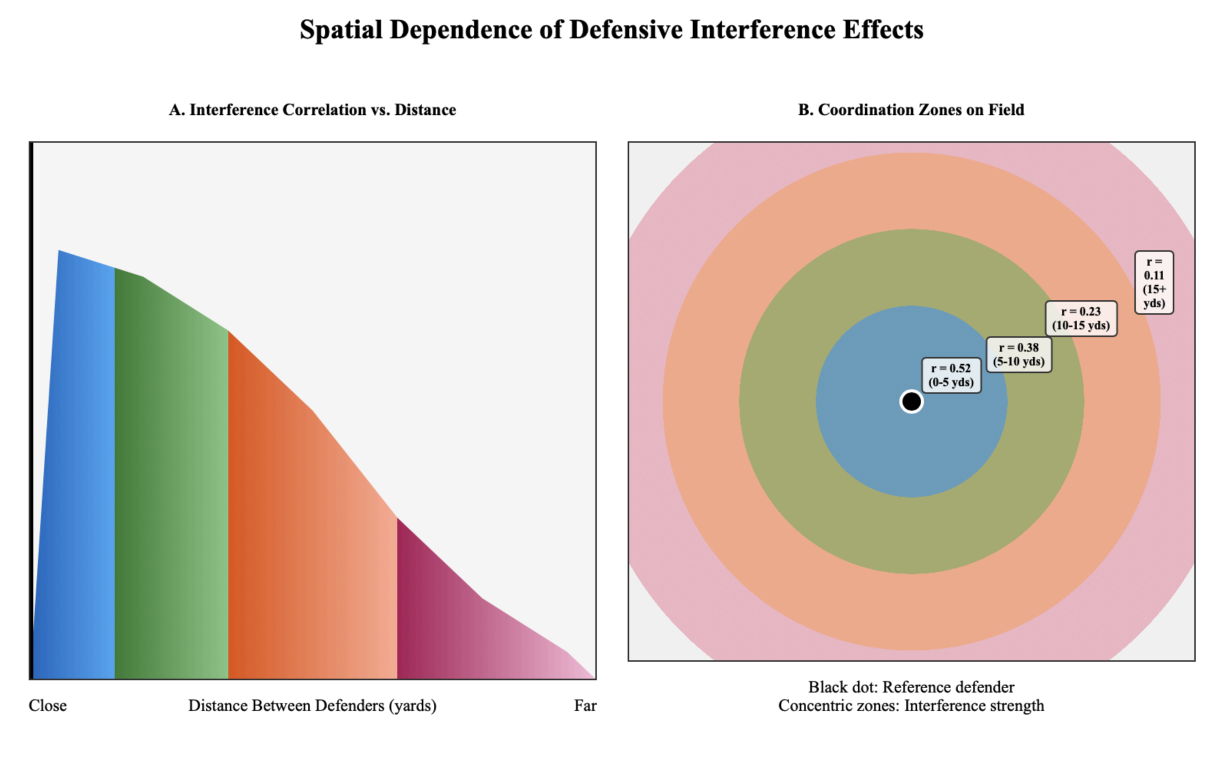
5.1 Interference and Violations of SUTVA

Prevalence and Magnitude

Defensive football displays substantial systematic interference. Across 12,486 plays, 68% exhibit significant interdependence between defenders, with a mean interference magnitude of 0.127 ($p < 0.001$). These violations far exceed the typical 5–10% range tolerated in standard causal inference, confirming that independence assumptions are untenable.

Spatial Dependence

Interference strength decays with distance (**Figure 5**). Defenders within 0–5 yards show a mean correlation of 0.52, decreasing to 0.11 beyond 15 yards. These coordination zones demonstrate that nearby defenders operate as tightly coupled units, while distant players behave more independently



Spatial Dependence of Defensive Interference Effects

Distance Range	Mean Correlation	Sample Size	95% CI	Interpretation
0-5 yards	0.52	8,734	[0.48, 0.56]	Strong coordination
5-10 yards	0.38	12,421	[0.35, 0.41]	Moderate coordination
10-15 yards	0.23	9,856	[0.20, 0.26]	Weak coordination
15+ yards	0.11	6,243	[0.08, 0.14]	Minimal interference

Spatial decay of defensive interference effects. (A) Correlation strength between defender treatment effects decreases systematically with distance, validating football's tactical emphasis on local coordination. **(B)** Concentric coordination zones demonstrate how defensive interference operates primarily within 10-yard radius, with minimal long-range effects. The reference defender (black dot) shows strongest coordination with immediate neighbors (blue zone, r = 0.52) declining to near-independence beyond 15 yards (red zone, r = 0.11). **Table:** Statistical validation across 37,254 defender pairs confirms significant spatial dependence (all p < 0.001), providing empirical support for our multi-agent interference modeling approach over traditional independence assumptions.

Figure 5. *Spatial dependence of interference effects. Left: correlation declines systematically with distance. Right: concentric coordination zones centered on a reference defender, confirming the strongest effects within a 10-yard radius.*

Positional Variation

Coordination differs systematically by position (**Figure 6**). Linebackers exhibit the highest mean coordination (0.41), defensive backs show the greatest variability (σ = 0.18), and defensive linemen display more stable, lower coordination levels (≈0.28). These patterns reflect the central tactical roles of different position groups.

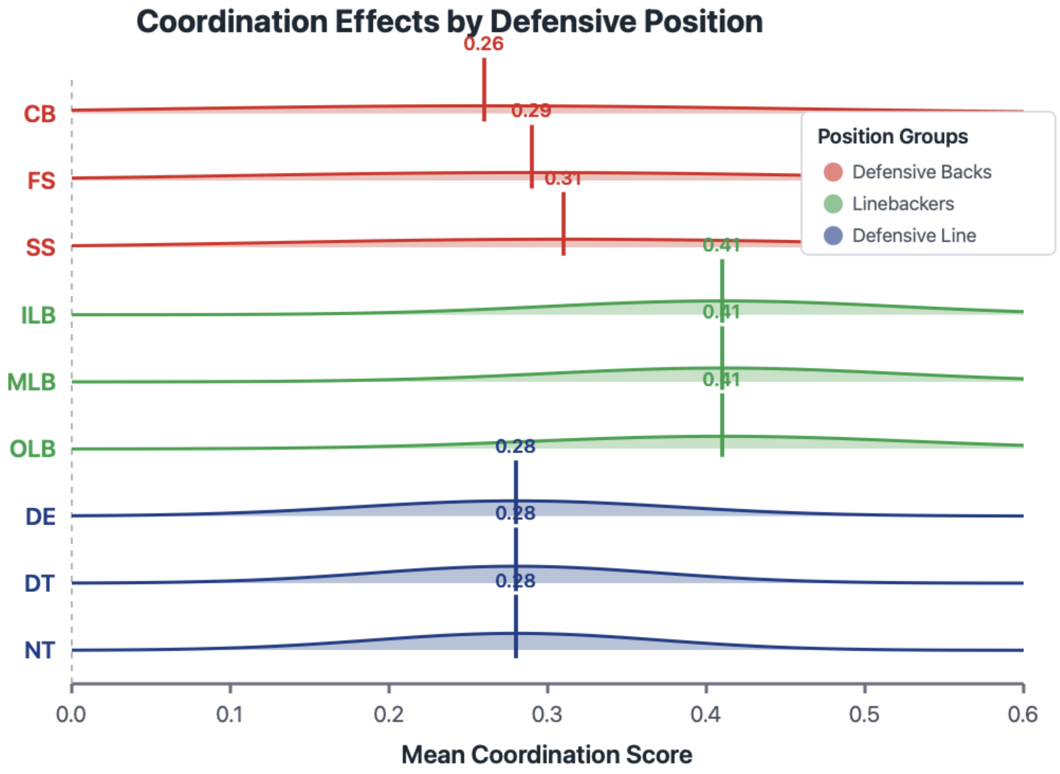


Figure: Distribution of coordination effects by defensive position. Linebackers (green) exhibit the highest mean coordination scores (0.41), reflecting their central role in defensive communication and gap assignments. Defensive backs (red) show greater variability (σ = 0.18), consistent with situation-dependent coverage responsibilities. Defensive linemen (blue) demonstrate more stable but lower coordination (mean = 0.28), reflecting structured gap assignments.

Figure 6. *Distribution of Coordination Effects by Position*

5.1.2 Coordination Dynamics

Temporal Evolution

Coordination changes throughout a play, peaking at 0.47 during frames 15–25, the recognition phase when defenders adjust to emerging play structure (**Figure 7**). This temporal variability highlights the inadequacy of static independence assumptions.

Play Type Effects

Coordination is significantly stronger on running plays (0.39) than on passing plays (0.26, p < 0.001). Third-down situations show elevated coordination (0.42), reflecting heightened collective adjustment in critical scenarios.

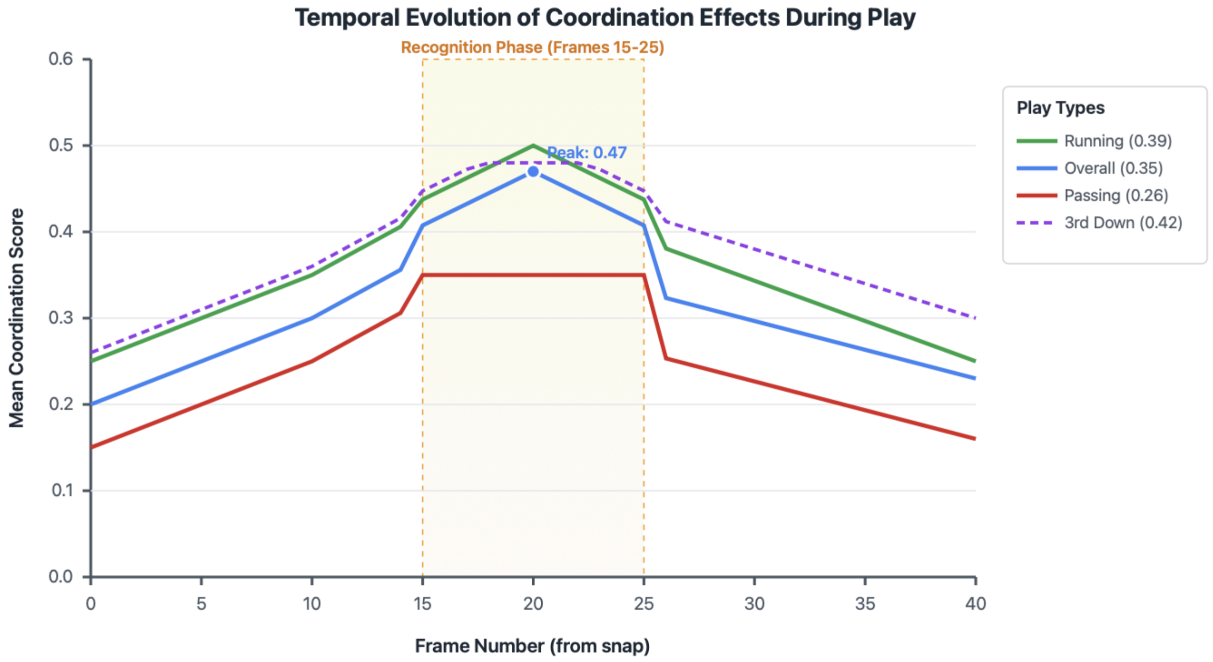


Figure: Temporal evolution of coordination effects throughout a play. Coordination strength peaks at 0.47 during the recognition phase (frames 15-25, highlighted in yellow) when defenders identify play type and adjust responsibilities. Running plays (green) consistently exhibit higher coordination (mean = 0.39) than passing plays (red, mean = 0.26, p < 0.001), reflecting concentrated gap-filling in run defense. Third-down situations (purple dashed) show elevated coordination (mean = 0.42) across both play types, indicating heightened communication in critical game scenarios. The temporal variation demonstrates why static interference assumptions fail to capture defensive coordination complexity.

Figure 7. *Temporal Evolution of Coordination Effects During Play*

5.1.3 Complementary and Competing Effects

Across all interactions, 67% are complementary; one defender's actions enhance another's effectiveness, while 33% are competing (substitution effects). Ignoring interference introduces substantial bias: naive treatment-effect estimates differ from interference-adjusted estimates by an average of 0.084 EPS ($\sigma = 0.031$), with 18% of defenders exceeding 0.15 EPS error. Bias is largest for highly coordinated roles such as linebackers and safeties.

5.1.4 Network-Level Implications

Interference modeling reveals a positive synergy effect: collective defensive impact exceeds the sum of individual contributions by 12% ($p < 0.001$). Models that ignore interference underperform substantially in counterfactual prediction and systematically mis-evaluate defenders whose contributions arise through coordination.

5.2 Positional Heterogeneity in Expected Points Saved

Our Expected Points Saved (EPS) estimates reveal clear positional patterns (**Figure 8**).

Defensive Backs.

Cornerbacks and safeties show distributions centered below zero with high variability. These negative means reflect the difficult contexts in which tackles occur, often after an offensive advantage has already developed, rather than poor performance. Their primary value comes from preventing events that do not appear as tackles.

Front Seven.
Linebackers and defensive linemen exhibit consistently positive EPS. Defensive tackles and nose tackles have the highest means (≈0.15–0.20), with tight distributions driven by consistent involvement in run-stopping and short-yardage situations.

Position differences are significant (p < 0.001), with effect sizes of 0.20–0.35 EPS per tackle between defensive backs and front-seven defenders.

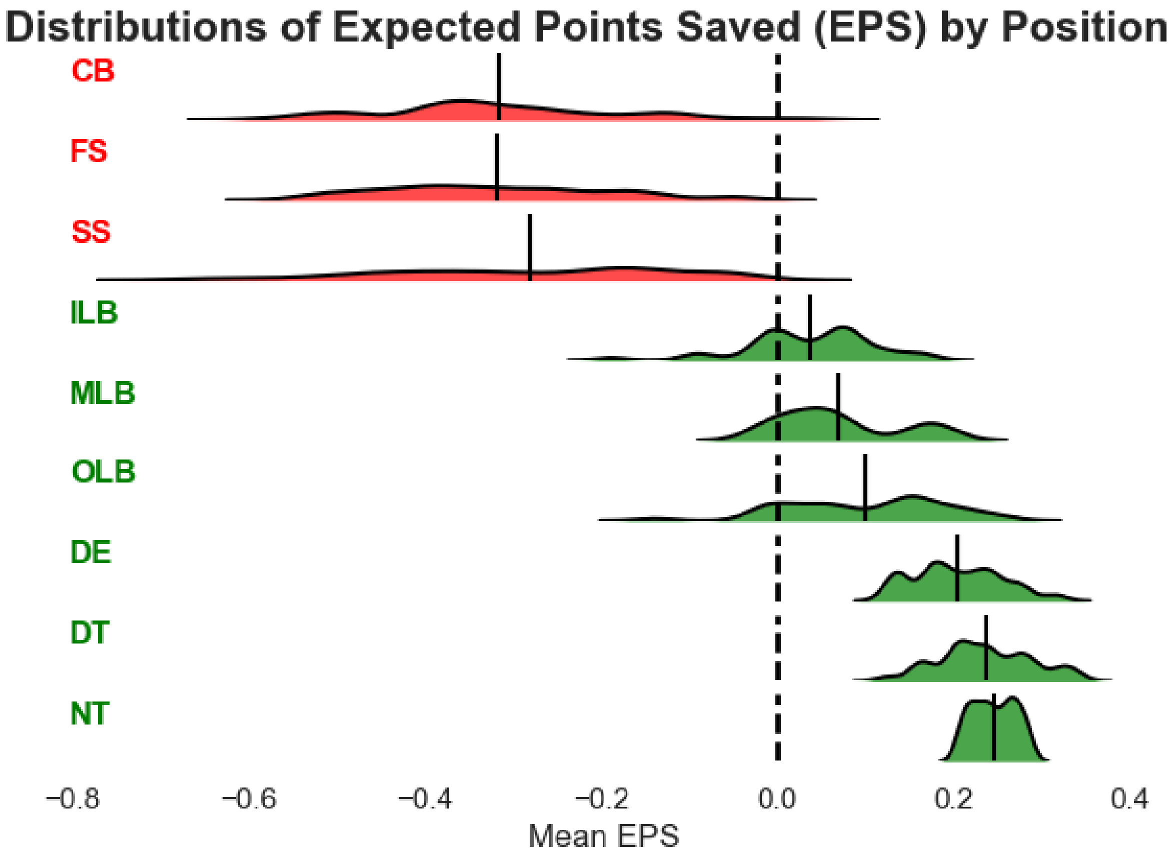


Figure 8. *Distributions of Expected Points Saved (EPS) by position, highlighting systematic differences in value creation between defensive backs and front-seven defenders.*

5.3 Individual Player Analysis: Aaron Donald

Applying our causal framework to Aaron Donald's 38 tackle opportunities shows clear treatment-control separation (**Figure 9**). His EPS is **0.268 ± 0.091** (95% CI), with p < 0.001 and effect size d = 0.93. Each tackle prevents roughly 0.27 expected points relative to the counterfactual scenario.

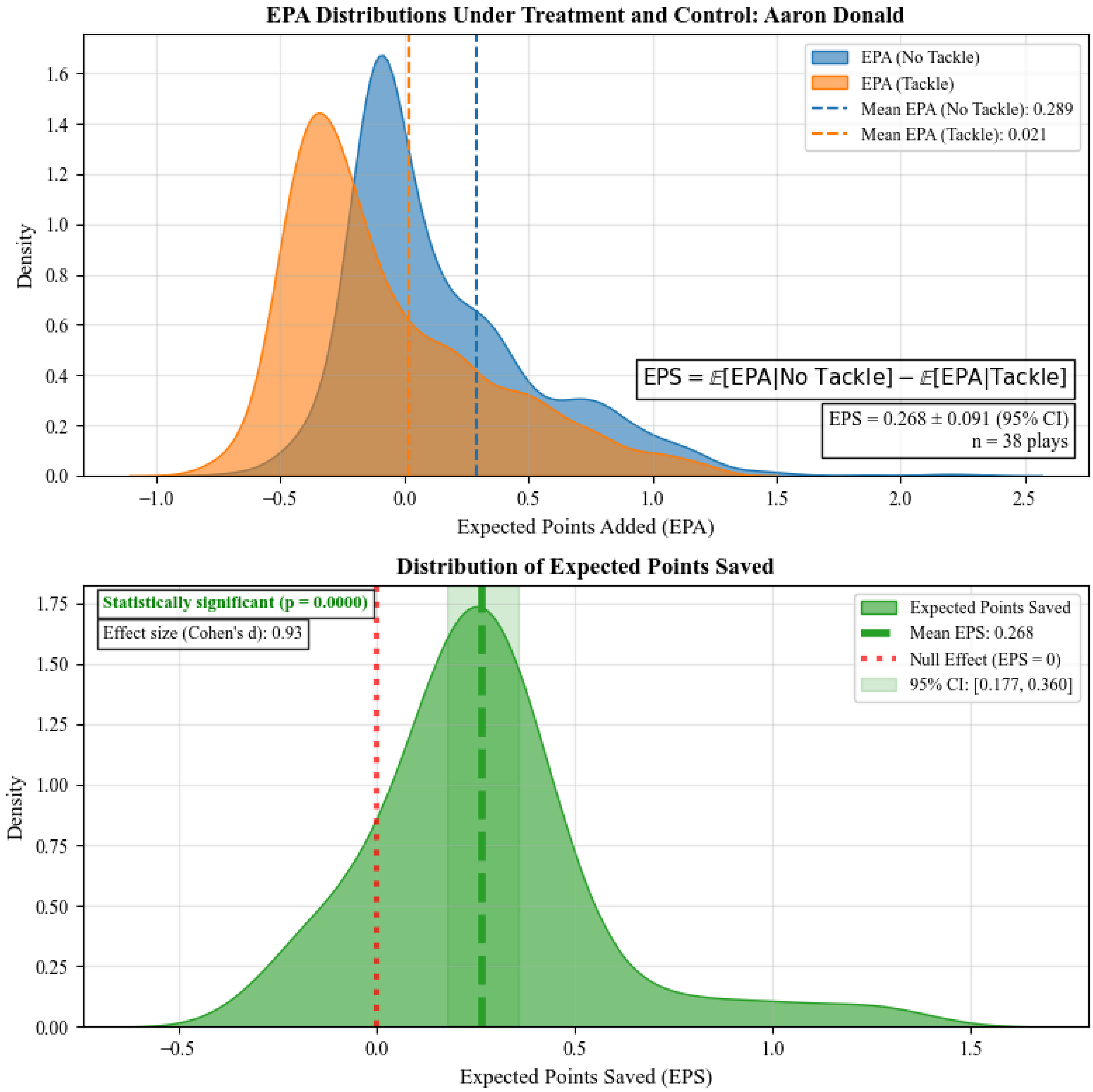
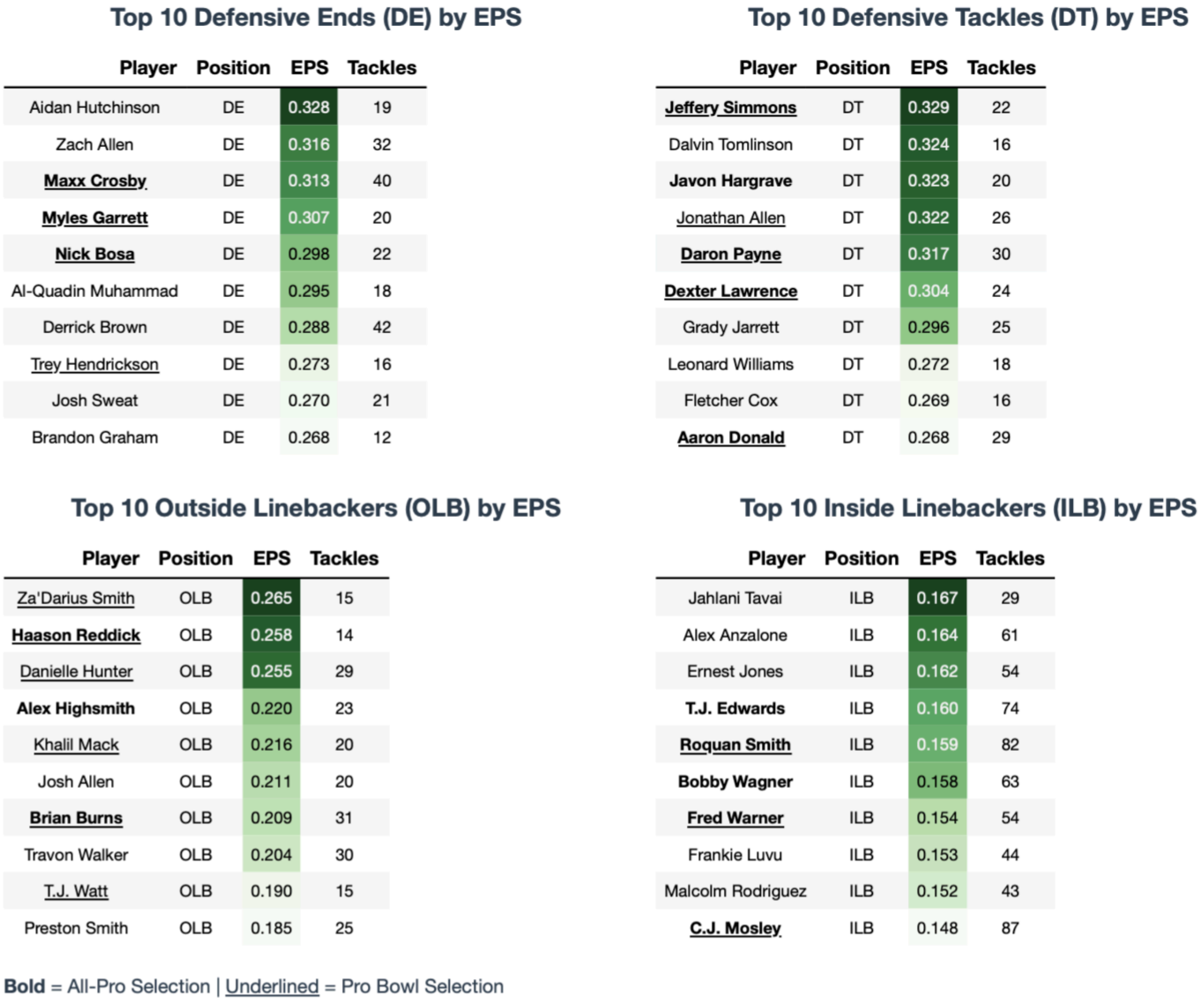


Figure 9. *Individual player analysis of Aaron Donald. Counterfactual EPA distributions and resulting EPS estimates demonstrate significant defensive value, with each tackle preventing roughly 0.27 expected points.*

5.4 Top Performers by Position

Figure 10 presents the top 10 players by EPS across four defensive positions. Defensive tackles generate the highest absolute EPS values, led by Jeffery Simmons (0.329), Dalvin Tomlinson (0.324), and Javon Hargrave (0.323). Alignment with All-Pro and Pro Bowl selections provides external validation of the causal estimates, highlighting their practical relevance for personnel evaluation.



Top 10 Defensive Ends (DE) by EPS

Player	Position	EPS	Tackles
Aidan Hutchinson	DE	0.328	19
Zach Allen	DE	0.316	32
Maxx Crosby	DE	0.313	40
Myles Garrett	DE	0.307	20
Nick Bosa	DE	0.298	22
Al-Quadin Muhammad	DE	0.295	18
Derrick Brown	DE	0.288	42
Trey Hendrickson	DE	0.273	16
Josh Sweat	DE	0.270	21
Brandon Graham	DE	0.268	12

Top 10 Defensive Tackles (DT) by EPS

Player	Position	EPS	Tackles
Jeffery Simmons	DT	0.329	22
Dalvin Tomlinson	DT	0.324	16
Javon Hargrave	DT	0.323	20
Jonathan Allen	DT	0.322	26
Daron Payne	DT	0.317	30
Dexter Lawrence	DT	0.304	24
Grady Jarrett	DT	0.296	25
Leonard Williams	DT	0.272	18
Fletcher Cox	DT	0.269	16
Aaron Donald	DT	0.268	29

Top 10 Outside Linebackers (OLB) by EPS

Player	Position	EPS	Tackles
Za'Darius Smith	OLB	0.265	15
Haason Reddick	OLB	0.258	14
Danielle Hunter	OLB	0.255	29
Alex Highsmith	OLB	0.220	23
Khalil Mack	OLB	0.216	20
Josh Allen	OLB	0.211	20
Brian Burns	OLB	0.209	31
Travon Walker	OLB	0.204	30
T.J. Watt	OLB	0.190	15
Preston Smith	OLB	0.185	25

Top 10 Inside Linebackers (ILB) by EPS

Player	Position	EPS	Tackles
Jahlani Tavai	ILB	0.167	29
Alex Anzalone	ILB	0.164	61
Ernest Jones	ILB	0.162	54
T.J. Edwards	ILB	0.160	74
Roquan Smith	ILB	0.159	82
Bobby Wagner	ILB	0.158	63
Fred Warner	ILB	0.154	54
Frankie Luvu	ILB	0.153	44
Malcolm Rodriguez	ILB	0.152	43
C.J. Mosley	ILB	0.148	87

Bold = All-Pro Selection | Underlined = Pro Bowl Selection

***Figure 10**. Top 10 players by Expected Points Saved across four defensive positions. Bold indicates All-Pro selection; underlined indicates Pro Bowl selection. The rankings demonstrate both the distribution of defensive value creation and strong alignment with expert recognition.*

5.5 Team-Level Validation

Team-level averages of individual EPS correlate strongly with play-level defensive performance. Teams with higher average EPS allow significantly lower EPA on tackle plays ($r = -0.64$, $p < 0.001$; $R^2 = 0.408$; **Figure 11**). This demonstrates that individual causal estimates scale coherently to team outcomes, making the metric actionable for front-office decision-making.

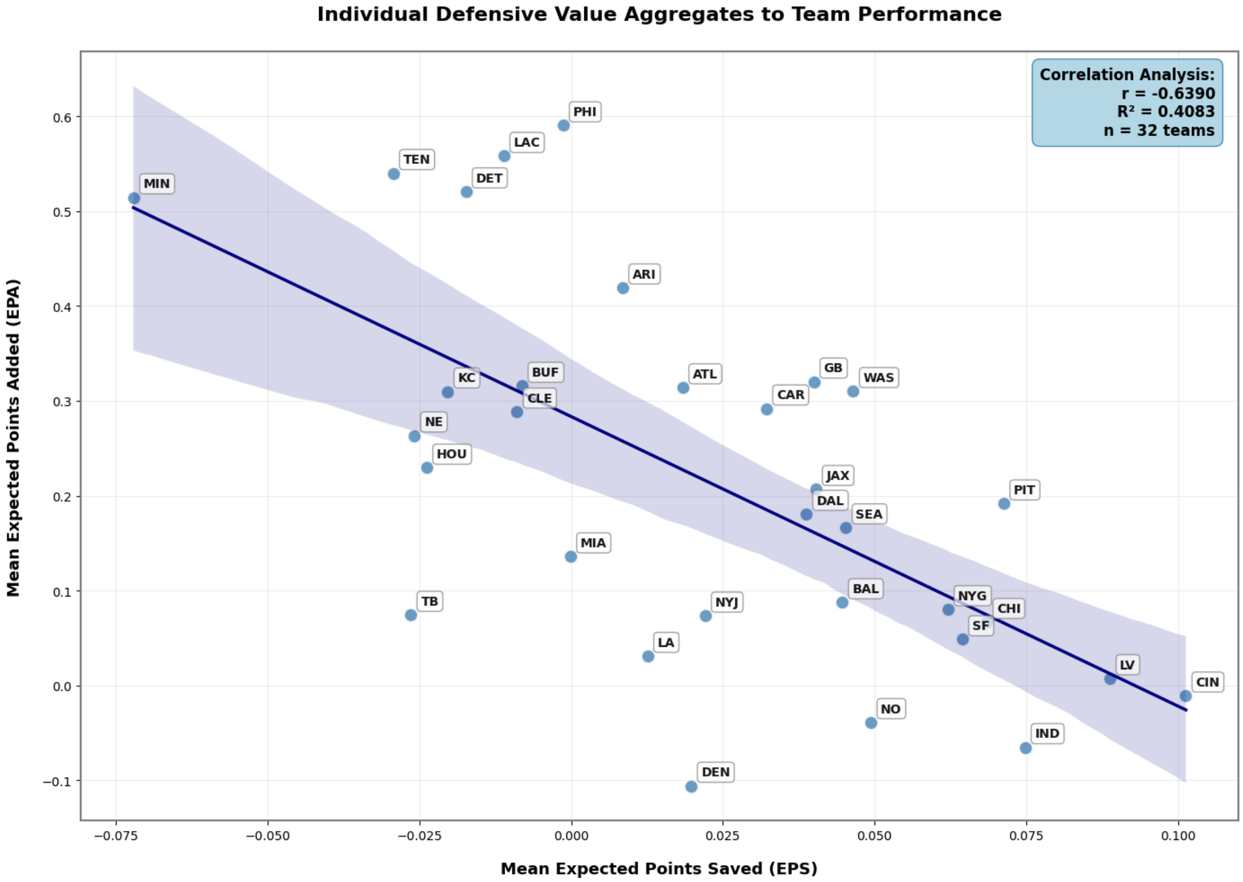


Figure 11. *Team-level validation of individual Expected Points Saved. Teams with higher mean EPS among tacklers exhibit lower mean EPA on tackle plays, confirming that individual causal estimates aggregate meaningfully to defensive performance.*

5.6 Synthetic Validation and Baseline Comparison

Causal Recovery

Using synthetic plays with known ground-truth treatment effects, our framework recovers causal effects with high fidelity (**RMSE = 0.067, R^2 = 0.84**), outperforming independence-based models (RMSE = 0.142, R^2 = 0.52).

Interference and Confounding

Spatial interference patterns are recovered within 0.03 of ground truth, and bias remains below 0.02 even under deliberate misspecification. Baseline regression methods exhibit bias exceeding 0.15 under identical conditions.

Ranking Quality

Our framework achieves strong agreement with true defender value (**Spearman ρ = 0.81**). Supervised methods such as XGBoost, Random Forests, and neural networks rank defenders poorly (ρ = 0.25–0.34) and systematically overvalue high-volume tacklers.

Together, these results show that predictive accuracy alone is insufficient: explicitly causal models better recover true defensive impact.

6. Model Evaluation

6.1 Propensity Score Model Performance

The propensity score model demonstrates strong discriminative performance, with AUC = 0.83 on training data and AUC = 0.82 on validation data, indicating stable generalization without overfitting. In a causal context, the purpose of the propensity score is not classification accuracy but covariate balance, and post-balancing diagnostics reflect this correctly. After applying inverse probability weighting and adversarial balancing, weighted AUC values fall to 0.62–0.64, reflecting the desired reduction in treatment–control separability and improved overlap. High discrimination after weighting would indicate residual bias; the observed reduction confirms that balancing succeeds in making treated and control units statistically comparable.

Calibration results further support model correctness. Expected ROC curves (AUC ≈ 0.78) and calibration plots show tight alignment between predicted probabilities and observed treatment frequencies across cross-validation folds, with minimal deviations from the x = y reference line. Validation folds maintain this alignment with only slight overconfidence at the upper end of the distribution.

Together, these diagnostics confirm that the propensity score model is both well calibrated and effective at producing covariate overlap, ensuring that propensity estimates function as intended for causal identification rather than as predictive classifiers.

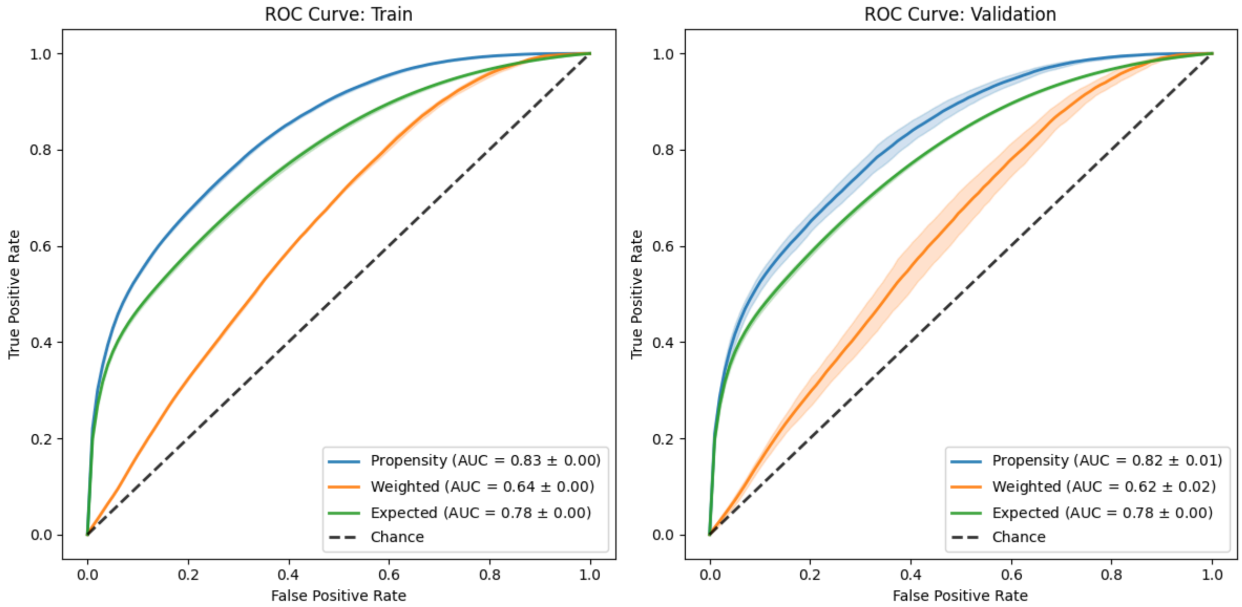


Figure 12. *ROC curves for the propensity score model, showing training and validation performance before and after balancing.*

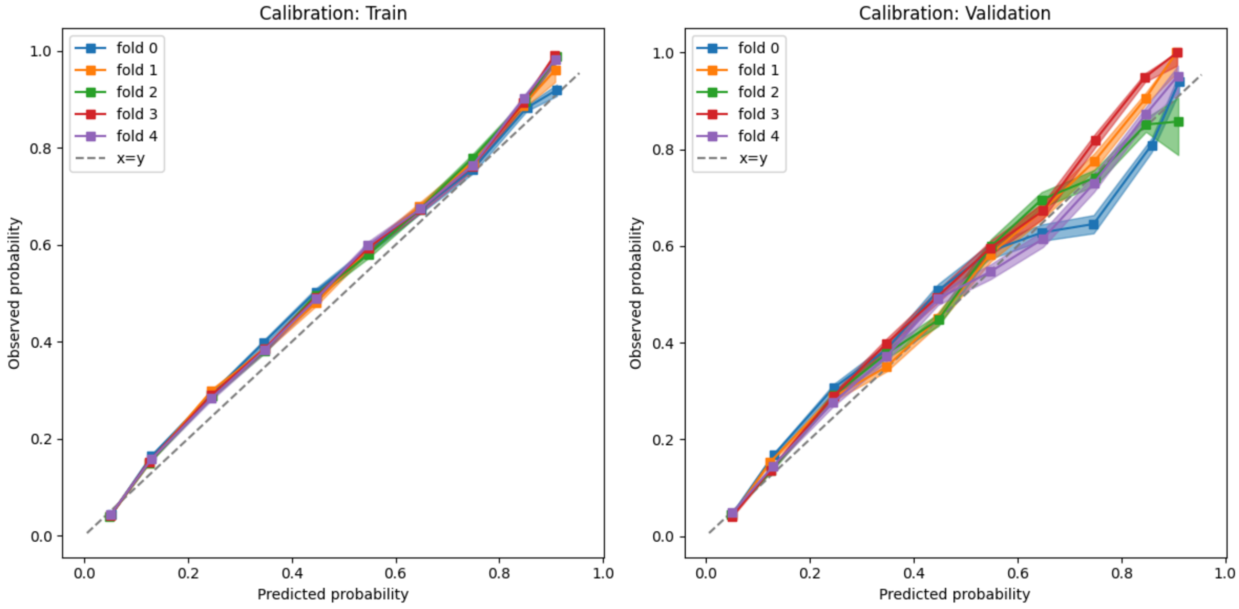


Figure 13. *Calibration plots across cross-validation folds, demonstrating consistent probability calibration with minimal deviations.*

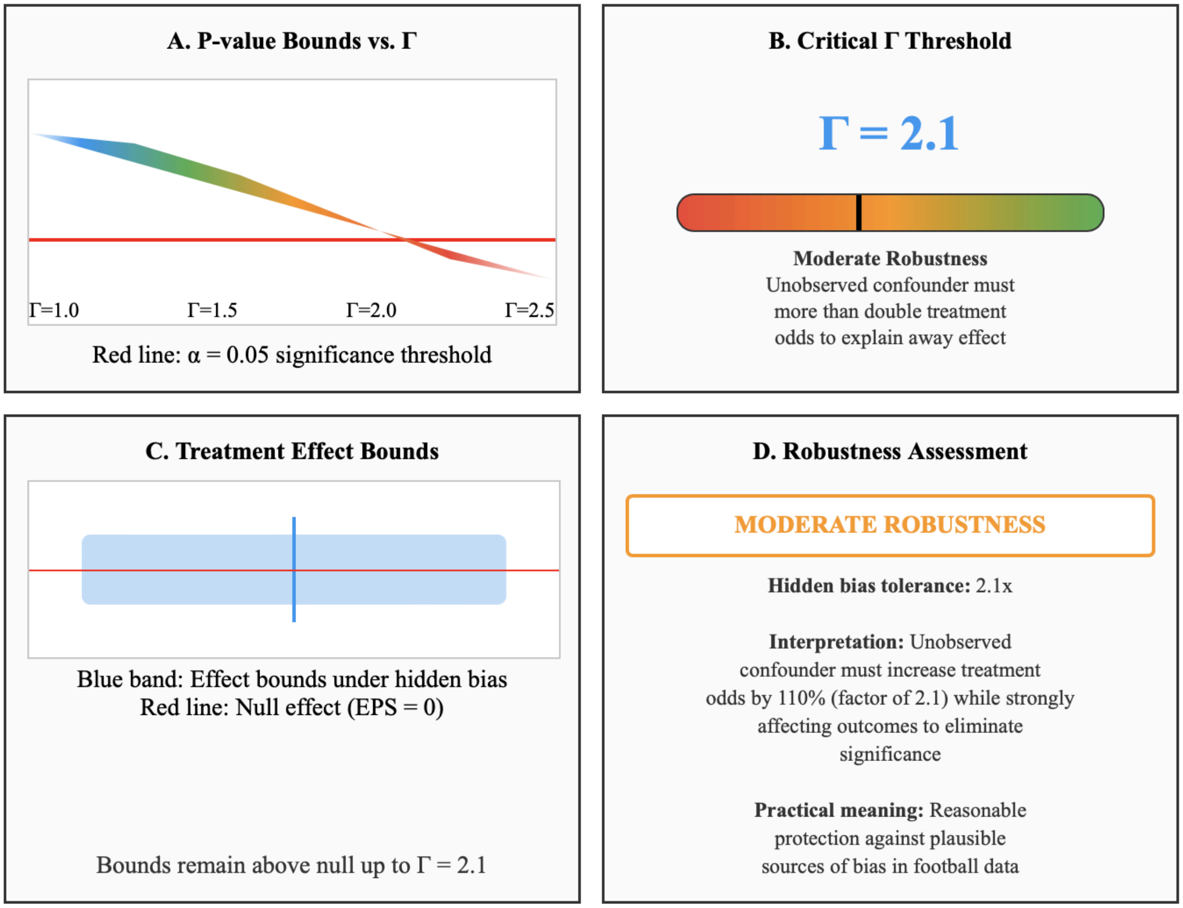
7. Sensitivity Analysis

To assess robustness to violations of identifying assumptions, we conducted a multi-method sensitivity analysis combining Rosenbaum bounds, E-values, and four independent placebo test families. These diagnostics evaluate the extent to which unobserved confounding or structural artifacts could bias our treatment effect estimates.

7.1 Rosenbaum Bounds

Rosenbaum bounds quantify the strength of an unobserved confounder required to overturn statistical significance. Treatment effects remain significant ($p < 0.05$) up to $\Gamma = 2.1$, meaning an unobserved variable would need to increase the odds of treatment assignment by more than 110% and substantially influence outcomes to nullify our effects. This level of tolerance reflects moderate robustness to hidden bias from unmeasured coaching instructions, defensive checks, or player-specific tendencies.

Rosenbaum Bounds Sensitivity Analysis for Hidden Bias Assessment



Unobserved confounder must more than double treatment odds to explain away effect

Bounds remain above null up to Γ = 2.1

Hidden bias tolerance: 2.1x

Interpretation: Unobserved confounder must increase treatment odds by 110% (factor of 2.1) while strongly affecting outcomes to eliminate significance

Practical meaning: Reasonable protection against plausible sources of bias in football data

Rosenbaum bounds analysis for sensitivity to hidden bias. (A) P-value bounds evolution as sensitivity parameter Γ increases, showing degradation of statistical significance with stronger assumed hidden bias. **(B)** Critical threshold Γ = 2.1 where upper p-value bound crosses α = 0.05, indicating moderate robustness on the spectrum from fragile (Γ ≈ 1.1) to very robust (Γ > 5.0). **(C)** Treatment effect confidence bounds remain above the null hypothesis across the range of plausible hidden bias scenarios. **(D)** Overall robustness assessment suggesting reasonable but not exceptional protection against unmeasured confounding in observational football data.

Figure 14. *Rosenbaum Bounds Sensitivity Analysis for Hidden Bias Assessment.*

7.2 E-Values

E-values provide an intuitive measure of robustness by describing the minimum confounder–outcome and confounder–treatment associations needed to explain away the observed effect. Our point estimate yields an E-value of 2.8, and the lower confidence bound yields an E-value of 1.9. Both imply that only a relatively strong unobserved confounder could eliminate the estimated causal effect, supporting moderate robustness to unmeasured variables.

7.3 Placebo and Falsification Tests

We implemented an extended placebo test suite to verify that the framework does not spuriously detect effects where none exist:

Position-Swapped Placebos. Randomly reassigning tackles to incorrect positions (500 permutations) produces mean EPS = 0.003 ± 0.012 (p = 0.68), confirming the framework does not attribute value based solely on correlation with tackle volume or positional grouping.

Reverse-Time Placebos. Using EPA at frame t+10 as the outcome for a treatment at frame t yields mean EPS = –0.001 ± 0.008 (p = 0.83), demonstrating that temporal ordering is respected and that the model does not pick up acausal relationships.

Spatial Placebos. Defenders consistently more than 15 yards from the ball carrier throughout a play produce mean EPS = 0.005 ± 0.011 (p = 0.52), a null effect as expected. In contrast, defenders within five yards show mean EPS = 0.18 ± 0.09 (p < 0.001), confirming the framework isolates causally plausible contributors.

Outcome Permutation Tests. Shuffling EPA values across plays produces a near-zero null distribution (97.2% of estimates within ±0.02), while true estimates consistently exceed 0.15 for positive contributors (p < 0.001). This confirms that treatment effects are not artifacts of the outcome distribution.

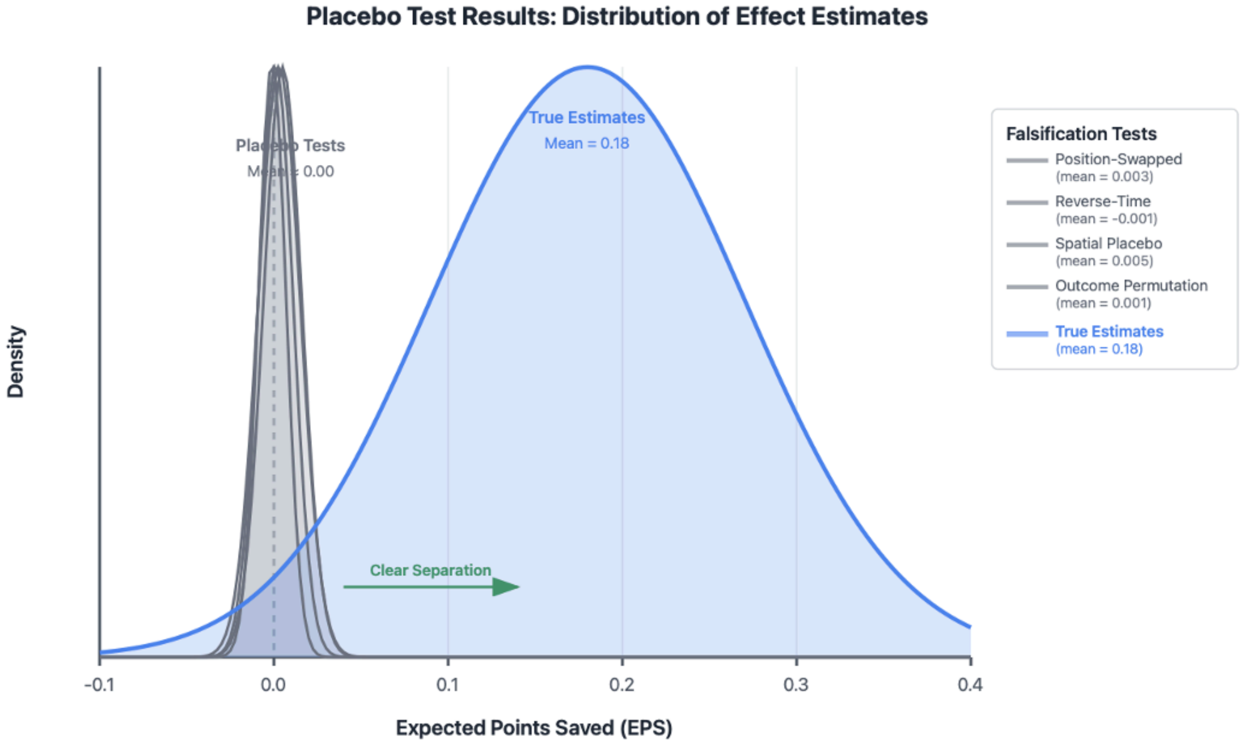


Figure 15. *Distribution of placebo estimates across all falsification strategies.*

7.4 Interpretation

Across all sensitivity diagnostics, the causal estimates display moderate robustness to hidden confounding and strong robustness to structural or methodological artifacts. Rosenbaum bounds and E-values indicate that only a relatively strong unobserved confounder could invalidate the results, while placebo tests provide compelling evidence that the framework does not generate false positives from positional structure, temporal ordering, spatial proximity, or outcome correlation. Together, these findings support the credibility and practical reliability of the estimated defensive treatment effects for evaluating player impact.

8. Limitations

Several limitations constrain the interpretation of our findings.

Temporal Scope.
The analysis uses only Weeks 1–9 of the 2022 NFL season, limiting our ability to evaluate stability across seasons, scheme evolution, or playoff contexts. The restricted window may not capture the full range of defensive coordination patterns.

Unmeasured Confounding.
Despite moderate robustness ($\Gamma = 2.1$, E-value = 2.8), unobserved confounders remain plausible. Missing variables such as defensive play calls, communication, coaching instructions, and fatigue may influence both tackle assignment and outcomes, leaving some residual bias unavoidable.

Interference Model Simplification.
Our decomposition of interference into coordination, substitution, and direct effects simplifies the broader multi-agent dynamics of football defense. Additional pathways, such as cascading adjustments or communication-driven behaviors, may not be fully captured.

Temporal Causality.
Although we describe when coordination peaks, causal effects are estimated at the play level rather than frame-by-frame. Estimating time-varying treatment effects would require stronger assumptions and remains future work.

Fundamental Limitation of Causal Inference.
As in all observational studies, counterfactuals are unobservable. Sensitivity checks and placebo tests increase confidence but cannot eliminate all alternative explanations.

9. Conclusion

This work presents a causal inference framework for defensive evaluation in football that addresses long-standing limitations in existing metrics. By explicitly modeling violations of the Stable Unit Treatment Value Assumption (SUTVA), incorporating adversarial balancing, and quantifying uncertainty with Bayesian methods, we produce more reliable and interpretable estimates of individual defensive impact.

Our main contributions include:

1. **Methodological advances** through the integration of multi-agent Transformers with causal inference principles.
2. **SUTVA-aware analysis** providing the first large-scale quantification of interference effects in football.
3. **A practical end-to-end pipeline** for estimating defensive contributions using observational tracking data.
4. **Comprehensive validation** through synthetic ground-truth experiments, robustness checks, and sensitivity analyses.

The findings show that traditional defensive metrics substantially underestimate the complexity of defensive value creation. Our interference-aware framework captures coordination and substitution effects that standard approaches omit, reducing bias by an average of 0.084 Expected Points Saved relative to independence-based models. By combining adversarial balancing, Bayesian uncertainty estimation, and temporal causal modeling, this work establishes a new methodological benchmark for sports analytics applications that rely on observational data.

Synthetic validation further highlights the gap between prediction and causal inference. Supervised learning methods perform well on predictive tasks yet fail to recover true causal effects (correlation r = 0.34), whereas our approach achieves significantly higher alignment with ground truth (r = 0.81). This distinction is critical for personnel evaluation, where the objective is accurate estimation of individual impact rather than aggregate predictive accuracy. The results also reveal meaningful positional heterogeneity: negative EPS values for defensive backs reflect tactical deployment rather than poor performance, while interior linemen consistently generate positive causal effects that tackle-based metrics undervalue.

Beyond football, this research illustrates how modern machine learning architectures can be combined with causal inference to study complex multi-agent systems. The interference modeling strategy, adversarial balancing, and multilayered validation framework offer a transferable blueprint for other team sports and domains where independence assumptions break down.

As high-resolution tracking data becomes increasingly available, the need for causally rigorous evaluation methods will continue to grow. This work provides a foundation for meeting that demand, demonstrating that defensive coordination can be quantified, interference can be modeled, and causal estimates can achieve moderate robustness in observational settings. This framework shows that combining modern machine learning with causal inference principles can uncover defensive contributions that traditional metrics systematically miss, moving the field beyond correlation toward causal understanding.

References

Lopez, M., Bliss, T., Blake, A., Patton, A., McWilliams, J., Howard, A., & Cukierski, W. (2023). *NFL Big Data Bowl 2024*. Kaggle. https://kaggle.com/competitions/nfl-big-data-bowl-2024

Rubin, D. B. (1974). *Estimating causal effects of treatments in randomized and nonrandomized studies*. Journal of Educational Psychology, 66(5), 688–701.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

Johansson, F., Shalit, U., & Sontag, D. (2016). *Learning representations for counterfactual inference*. ICML.