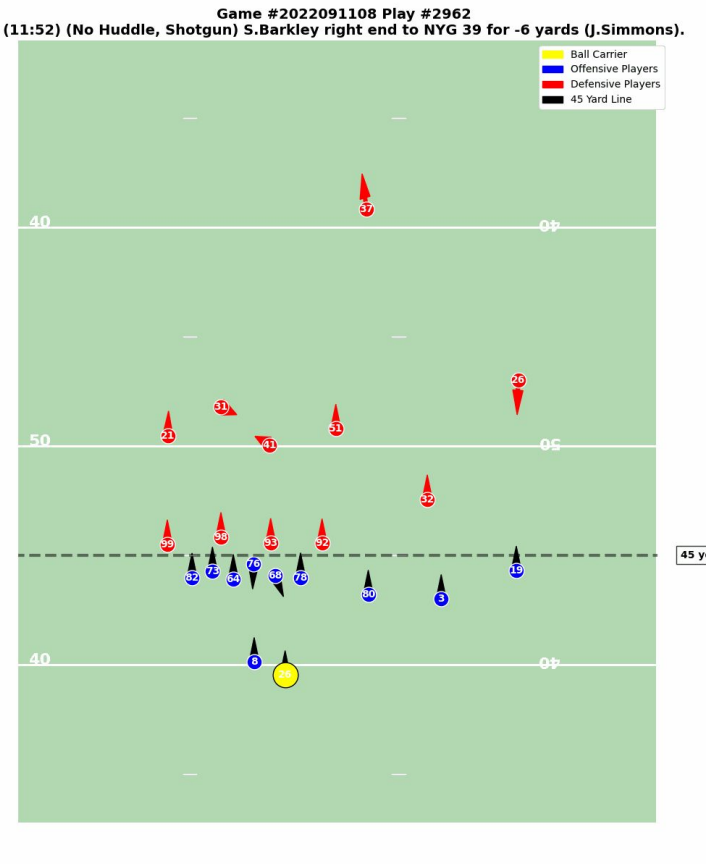
Tackling Causality: Estimating Frame-Level Defensive Impact with Multi-Agent Transformers

Ben Jenkins

The Fundamental Problem



- **Who gets credit for this tackle?**
- Traditional metrics: Only the tackler gets credit
- **Reality:** Multiple defenders contributed causally to the outcome
- **The Question:** How do we measure what doesn't show up in the box score?
- **Goal:** Use tracking data to answer counterfactual questions that help us better identify the contributions of each player

NFL Player Tracking Data

- 22 players tracked at 10 Hz
- (x, y) coordinates, velocity, acceleration, direction, and orientation
- 2024 NFL Big Data Bowl focused on measuring tackling performance
- Weeks 1-9 of 2022 Season

Supervised Learning vs. Causal Inference

Supervised Learning Asks: "What will happen?"

- Prediction-focused: Given player positions, predict EPA outcome
- Correlation-based: Finds patterns in data
- Goal: Minimize prediction error

Supervised Learning Problems:

- Confuses correlation with causation
- Biased toward high-volume tacklers
- Can't separate skill from opportunity

Causal Inference Asks: "What would happen if...?"

- Counterfactual-focused: What if this defender didn't make the tackle?
- Causation-based: Isolates true cause-and-effect
- Goal: Estimate treatment effects

Causal Inference Advantages:

- Isolates individual contributions
- Accounts for situational context
- Separates player effect from team/scheme effects

Bottom Line: Prediction ≠ Attribution. For estimating player contribution, we need causal understanding, not just predictive accuracy.

The Defensive Attribution Problem

Traditional: "Player A: 8 tackles

Player B: 2 tackles

Who is better?

The Challenge:

- Defensive value is largely invisible in traditional metrics
- Coordination and positioning matter more than final contact
- Counterfactual question: "What would have happened without this defender?"

Limitations of ML Residuals for Player Attribution

Common Approach:

- Train ML model to predict EPA or some other target from features
- Use residuals as "unexplained performance"
- Attribute residuals to individual players

Why This Fails in Football:

1. Confounding Problem:

- Residuals capture **unexplained variance**, not causal effects
- High residuals may reflect favorable situations

2. No Counterfactual Framework:

- Residuals ask: "What did we fail to predict?"
- Causal inference asks: "What would happen without this player?"
- **Fundamentally different questions**

3. Interference Contamination:

- ML residuals mix individual skill with teammate effects
- Can't separate "player A is good" from "player A benefits from player B"
- Coordination effects get misattributed to individuals

4. Selection Bias:

- Players in different situations generate different residual patterns
- No mechanism to ensure fair comparisons
- **Apples-to-oranges problem**

Real Example:

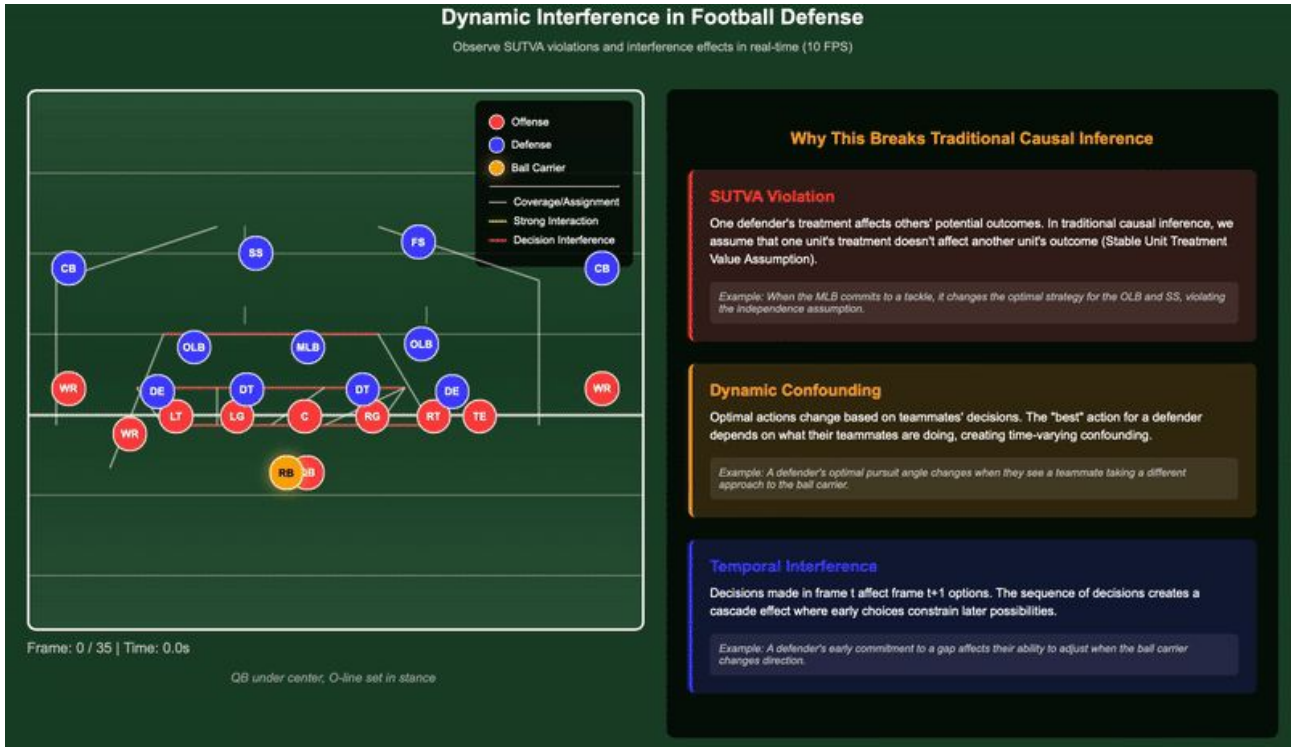
- Suppose Aaron Donald gets +0.3 EPA residual
- Is this because: (a) He's elite, or (b) Offense was already disrupted by teammates?
- **Residuals can't distinguish between these**

Bottom Line: Residuals measure prediction errors, not causal contributions. Attribution requires explicit counterfactual modeling.

Why Current Defensive Metrics Fail

- **Tackle counts** → Miss anticipatory positioning
- **Correlation-based models** → Cannot answer "What if?" questions
- We need causal inference, not just correlation

Why Standard Approaches Fail



The SUTVA Problem in Football:

- **S**table **U**nit **T**reatment **V**alue **A**ssumption
- Traditional causal inference assumes units don't interfere with each other
- In football: Defenders coordinate extensively

Multi-Agent Systems in Football

What is Multi-Agent?

- **Traditional approach:** Analyze each player independently
- **Multi-agent approach:** Model all 22 players simultaneously as interacting agents
- Each "agent" (player) influences and responds to others in real-time

Why Football Requires Multi-Agent Modeling:

The Coordination Problem:

- Defenders don't act in isolation - they coordinate responsibilities
- One defender's movement changes optimal actions for teammates
- Traditional methods assume independence (violates reality)

Three Types of Agent Interactions

1. **Coordination:** Defenders work together (coverage schemes, gap assignments)
2. **Interference:** One defender's action affects teammate's effectiveness
3. **Substitution:** Defenders compensate when teammates are out of position

Technical Implementation:

- **Transformer attention mechanisms** learn which players influence each other
- **Spatial relationships** captured through learned attention weights
- **Temporal dynamics** model how interactions evolve during play

Key Insight: Football defense is a complex adaptive system where individual value emerges from collective behavior

Analogy: Like modeling a flock of birds - individual flight patterns depend on the entire group, not just individual bird characteristics

Our Causal Framework

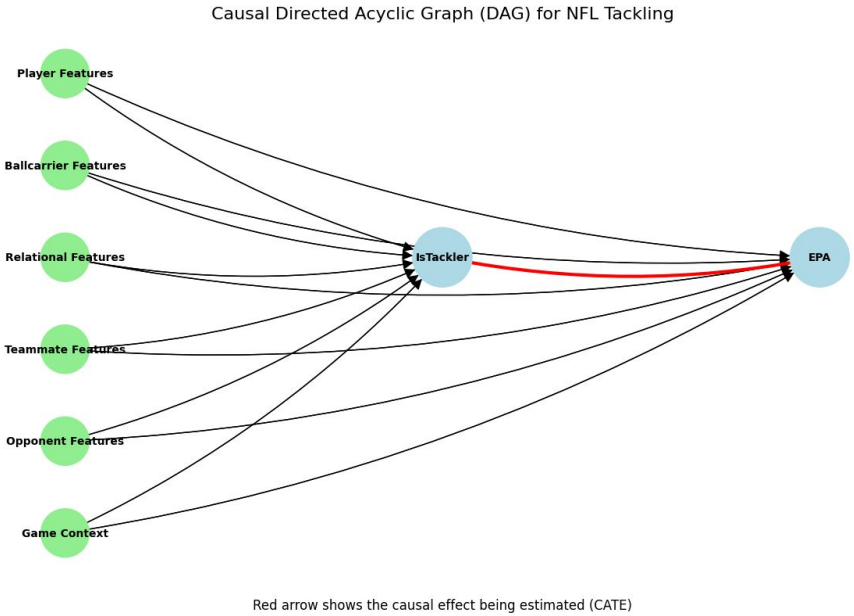
Treatment: Binary indicator - did defender i make a tackle?

Outcome: Expected Points Added (EPA) - how much did the offense benefit?

Goal: Estimate $E[Y_0|X] - E[Y_1|X]$ = **Expected Points Saved (EPS)**

Key Innovation - Multi-Agent Modeling:

- Transformer architecture processes all 22 players simultaneously
- Attention mechanisms capture defender coordination
- Explicit interference modeling: coordination + substitution + competition effects



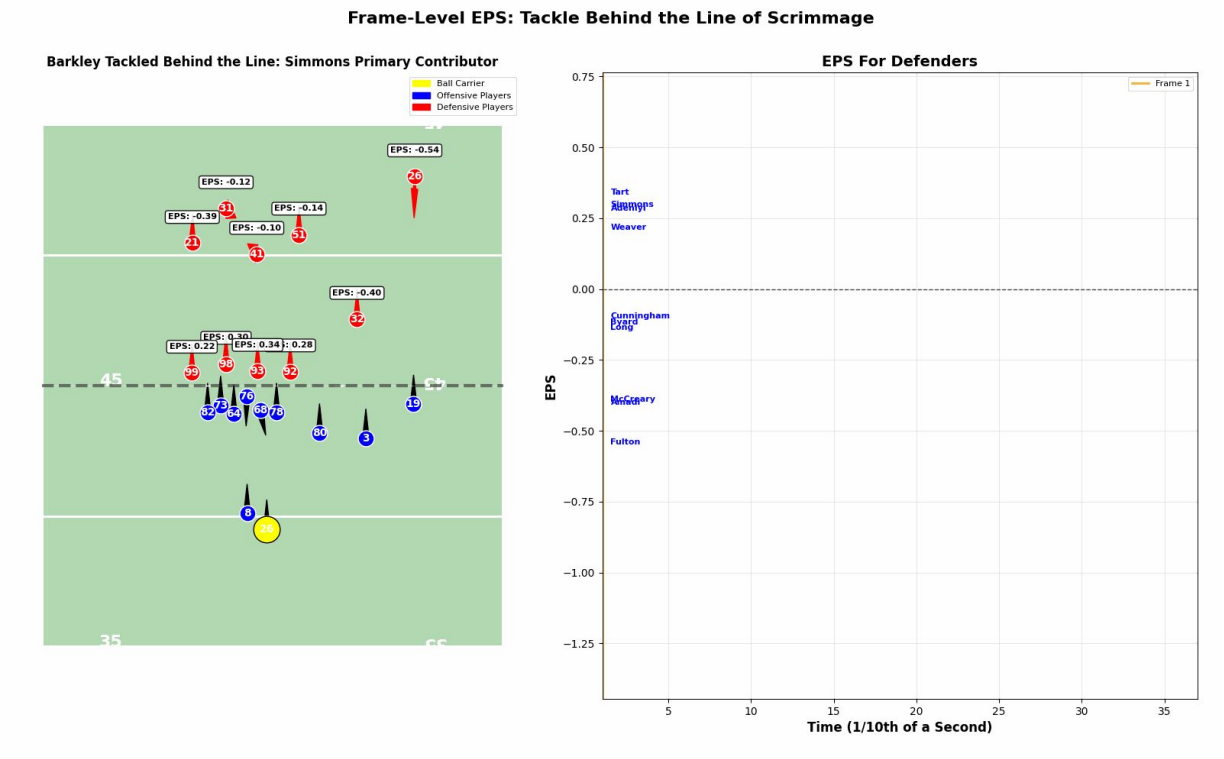
Addressing Selection Bias

Problem: Defenders don't tackle randomly

- **MMD balancing:** Match distributions across treatment groups
- **Adversarial training:** Domain classifier can't distinguish T=0 vs T=1
- **Representation equilibrium:** Similar covariates → similar representations
- **Doubly robust estimation:** Combines propensity scores + outcome modeling
- **Protection:** Consistent estimates even if one model component is wrong

Result: Creates "apples-to-apples" comparisons for valid causal inference

Play Example: Tackle Behind Line of Scrimmage

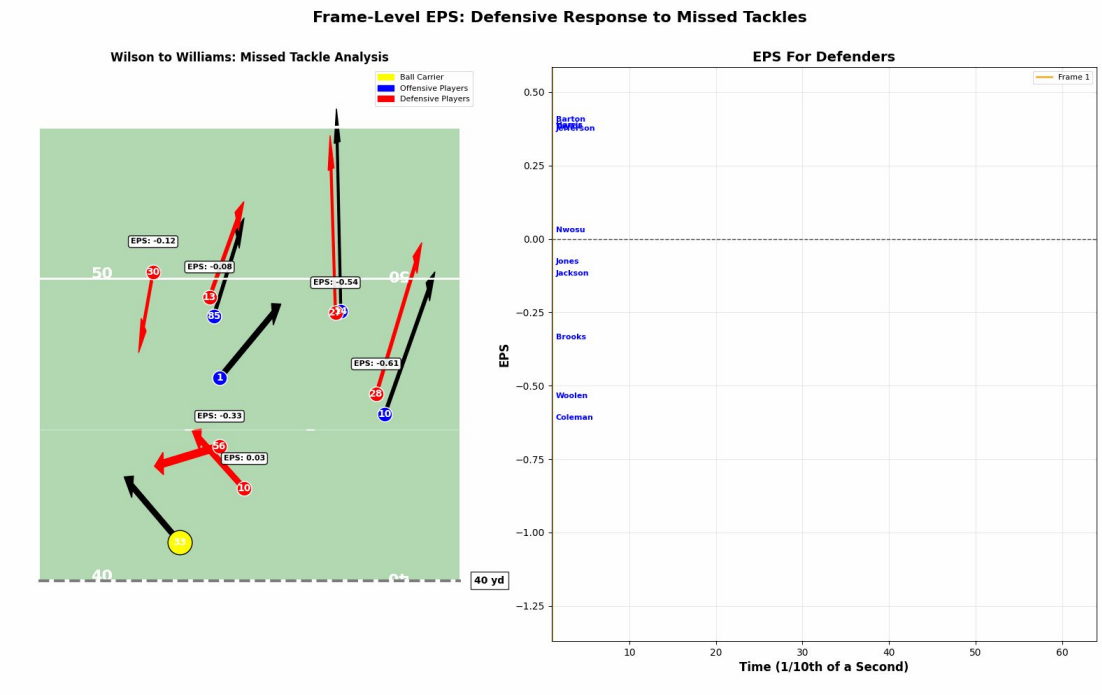


Play Setup: Saquon Barkley tackled by Jeffery Simmons for 6-yard loss
Traditional View: Only Simmons gets credit for the tackle
Our Framework Reveals:

- Simmons: Peak EPS of 0.66 (primary contributor)
- Tart & Weaver: Consistent EPS of 0.25-0.35
- Interior defenders limited running lanes, forced play toward Simmons

Key Insight: Framework captures coordinated defensive effort, not just final tackler
Value: Identifies "setup players" invisible to traditional tackle-counting metrics

Play Example: Missed Tackle



- **Play Setup:** Russell Wilson to Javonte Williams, 9-yard gain after breaking multiple tackles
- **Failed Tacklers:**
  - Jackson: -0.15 EPS (missed tackle allowed offensive advantage)
  - Brooks: -0.35 EPS (failed attempt extended the play)
- **Damage Control:**
  - Barton: +0.40 EPS (best pursuit angle, prevented larger gain)
  - Nwosu: +0.05 EPS (neutral impact)
- **Cascading Effects:**
  - Woolen: -0.50 EPS, Coleman: -0.60 EPS, Diggs: -0.90 EPS
  - Values reflect positioning relative to extended play development
- **Key Insight:** Framework captures both failure costs and recovery value in real-time

Validating Causal Models vs. Traditional ML

The Challenge: How do we know our causal estimates are right?

Traditional ML Validation (Not Sufficient for Causal Claims):

- Cross-validation accuracy, AUC, R^2
- Tests prediction quality, not causal validity
- Can achieve high accuracy while missing causal relationships

Causal Model Validation (What We Need)

1. Synthetic Data with Known Ground Truth

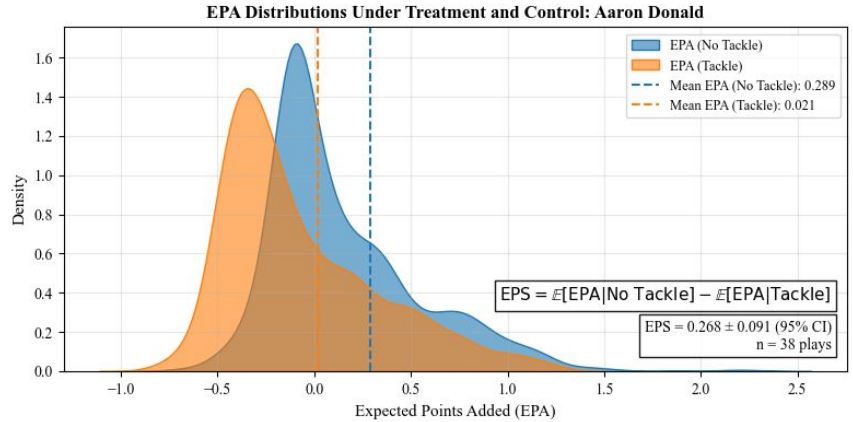
2. Aggregation Tests

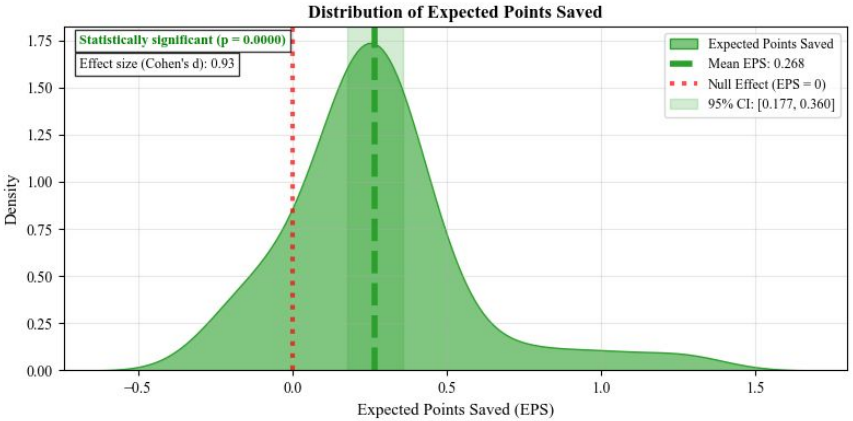
3. Sensitivity Analysis

4. Counterfactual Coherence:

Key Insight: Causal validity requires fundamentally different validation than predictive accuracy

Individual Case Study: Aaron Donald





- **Sample:** 38 tackle opportunities with clear counterfactual distributions
- **Results:**
  - Control condition: Mean EPA = 0.289
  - Treatment condition: Mean EPA = 0.021
  - Expected Points Saved: 0.268 ± 0.091 (95% CI)
- **Statistical Significance:**
  - $p < 0.001$, large effect size (Cohen's d = 0.93)
  - Confidence interval [0.177, 0.360] well above zero
- **Practical Impact:** Each Donald tackle prevents ~0.27 expected points
- **Framework Validation:** Clear separation between treatment/control validates causal identification
- **Uncertainty:** Confidence intervals reflect appropriate statistical uncertainty for 38 plays

Top Defenders by EPS

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

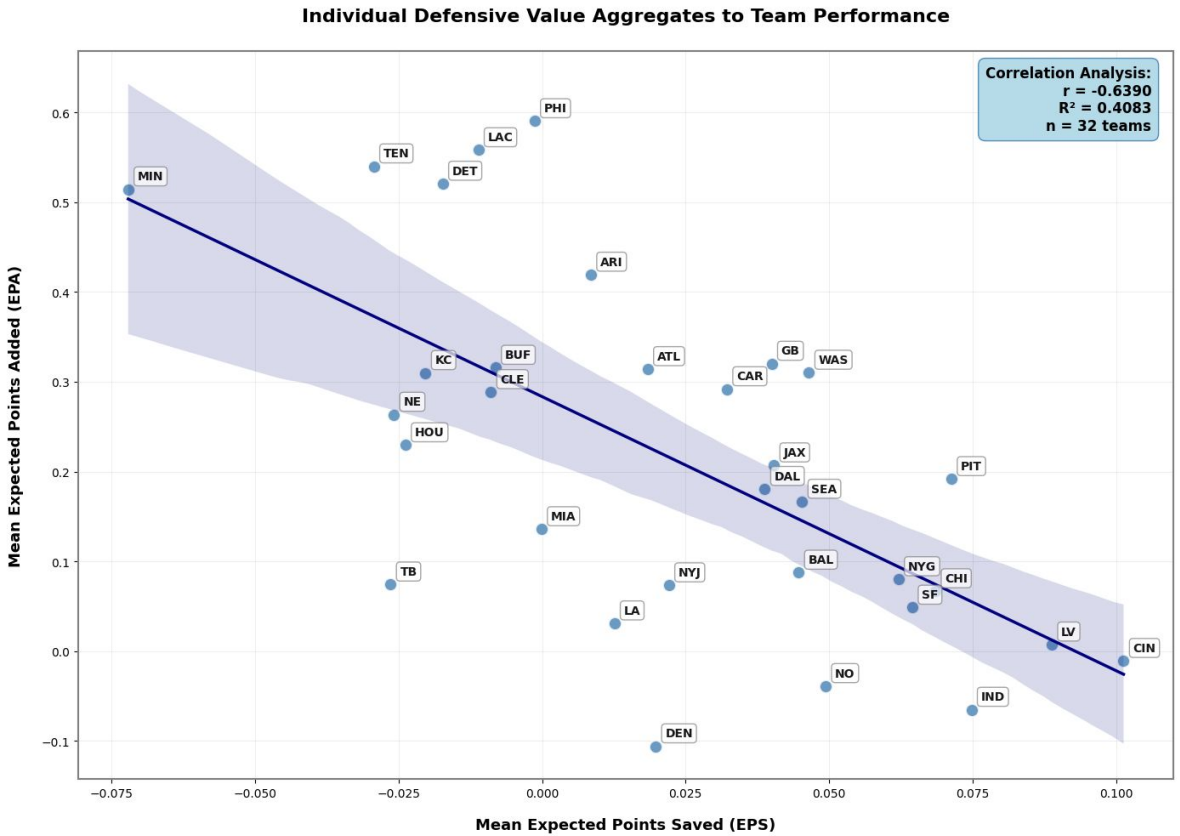
Player	Position	EPS	Tackles
Za'Darius Smith	OLB	0.265	15
Haason Reddick	OLB	0.258	14
Danielle Hunter	OLB	0.255	29
Alex Highsmith	OLB	0.220	23
Khalil Mack	OLB	0.216	20
Josh Allen	OLB	0.211	20
Brian Burns	OLB	0.209	31
Travon Walker	OLB	0.204	30
T.J. Watt	OLB	0.190	15
Preston Smith	OLB	0.185	25

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

Team-Level Validation



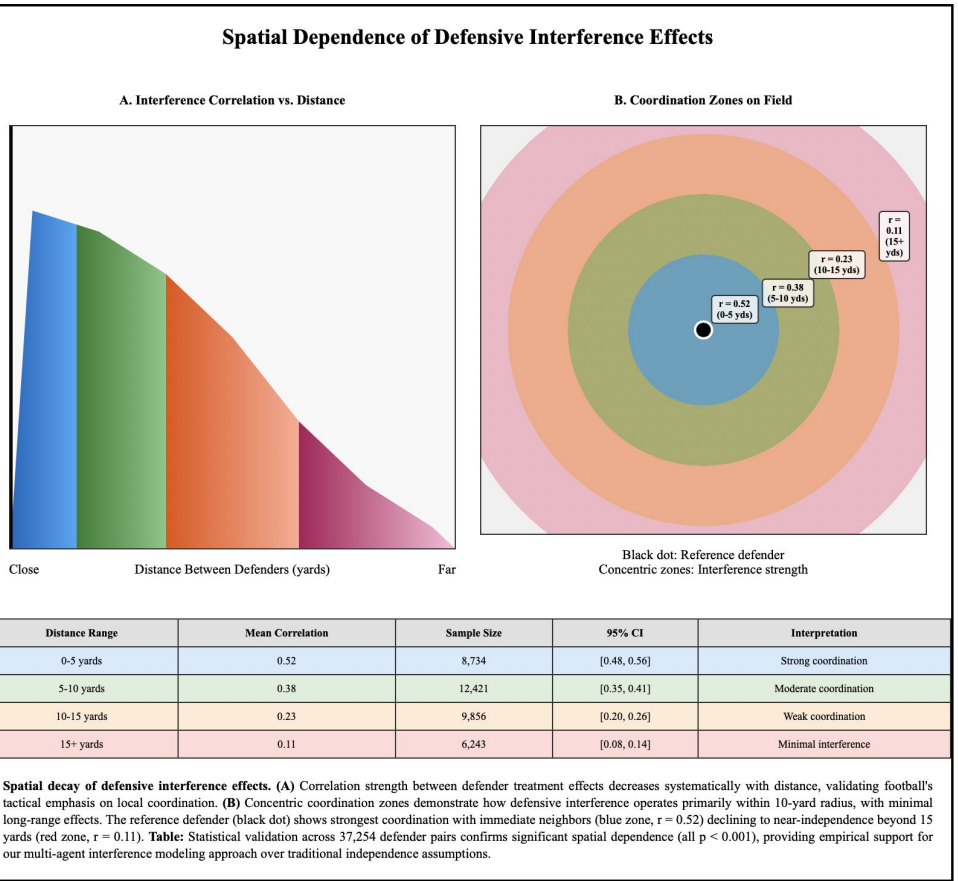
Analysis: Mean individual EPS vs. mean EPA for tacklers across all 32 NFL teams
Strong Correlation: r = -0.64, p < 0.001

- Teams with better tacklers (higher EPS) → better outcomes (lower EPA)
- Explains 41% of variance in tackle-play outcomes

Key Validation: Individual causal estimates capture genuine defensive value, not statistical artifacts

Practical Implication: Framework reliably distinguishes defensive value among tacklers

Validation Results



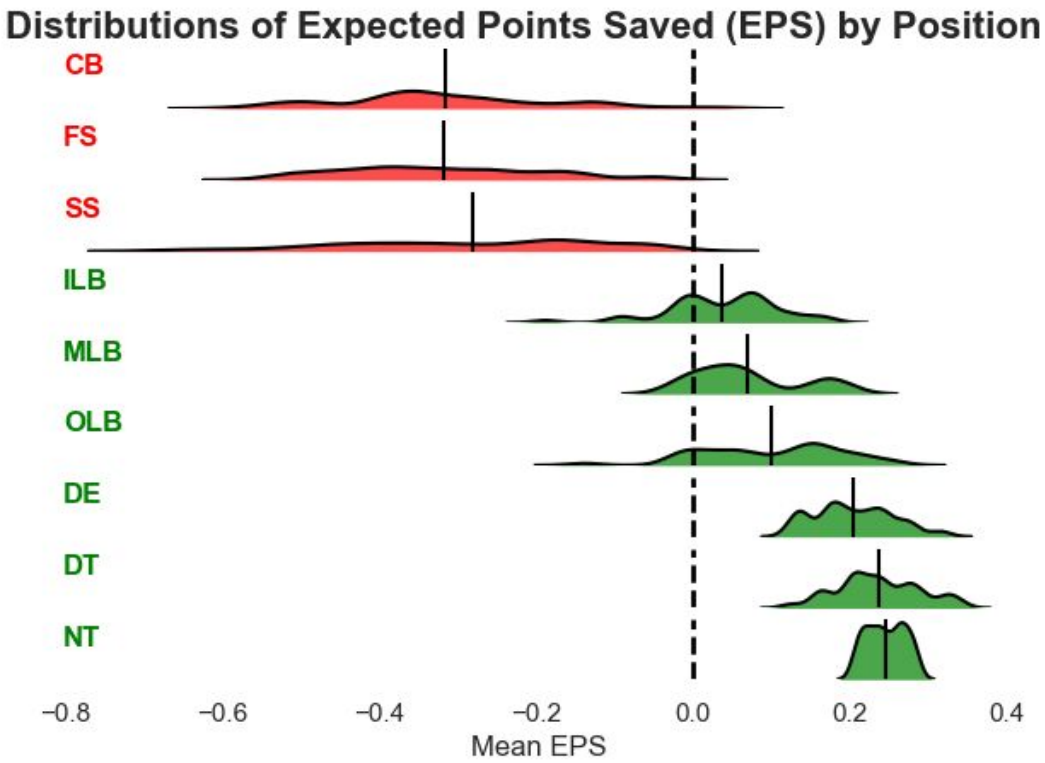
Distance Range	Mean Correlation	Sample Size	95% CI	Interpretation
0-5 yards	0.52	8,734	[0.48, 0.56]	Strong coordination
5-10 yards	0.38	12,421	[0.35, 0.41]	Moderate coordination
10-15 yards	0.23	9,856	[0.20, 0.26]	Weak coordination
15+ yards	0.11	6,243	[0.08, 0.14]	Minimal interference

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.

SUTVA Violation Evidence:

- 68% of plays show significant interference effects
- Mean interference magnitude: 0.127 ($p < 0.001$)
- Spatial decay: correlation drops from 0.52 (within 5 yards) to 0.11 (beyond 15 yards)

Positional Heterogeneity in Tackling Expected Points Saved:

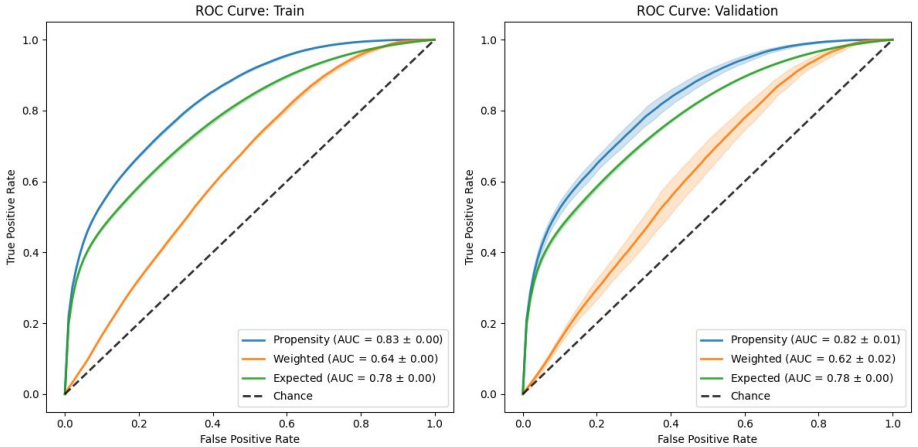


- **Defensive Backs (CB, FS, SS):**
  - EPS distributions centered below zero with wide spreads
  - **Context matters:** Tackles occur after offense gains advantage
  - Three scenarios: successful offensive execution, coverage breakdowns, pursuit situations
  - Negative EPS reflects tackle context, not performance quality
- **Front Seven (LB, DL):**
  - Positive EPS distributions (0.05-0.15 for LBs)
  - Nose tackles highest: ~0.20 EPS (disrupting interior runs)
  - Tighter distributions = more predictable tackle scenarios
- **Key Interpretation:**
  - Defensive backs' most valuable work (coverage, prevention) invisible in tackle metrics
  - When DBs tackle, offense has often already succeeded
  - Front seven tackles occur in more favorable defensive contexts
- **Statistical Significance:** $p < 0.001$ across position comparisons, effect sizes 0.20-0.35

Methodological Validation

- **Synthetic Data Validation (15,000 realistic plays):**
  - Our framework: RMSE = 0.067, R^2 = 0.84, ranking correlation ρ = 0.81
  - Traditional ML: RMSE = 0.142, R^2 = 0.52, ranking correlation ρ = 0.34
  - Key insight: Prediction accuracy ≠ causal accuracy
- **Interference Effect Recovery:**
  - Accurately identifies spatial decay patterns (within 0.03 of ground truth)
  - Traditional methods overestimate individual effects by 0.089 EPS
  - Validates multi-agent coordination modeling
- **Baseline Comparisons:**
  - XGBoost predictive R^2 = 0.78 vs. our 0.84
  - But XGBoost causal correlation only r = 0.34 vs. our r = 0.81
  - Traditional ML biased toward high-volume tacklers

Propensity Score Performance



Why Propensity Scores Matter:

- Essential for causal inference - estimate probability of "treatment" (making a tackle)
- Must balance accuracy vs. overlap for valid causal estimates

Performance Metrics:

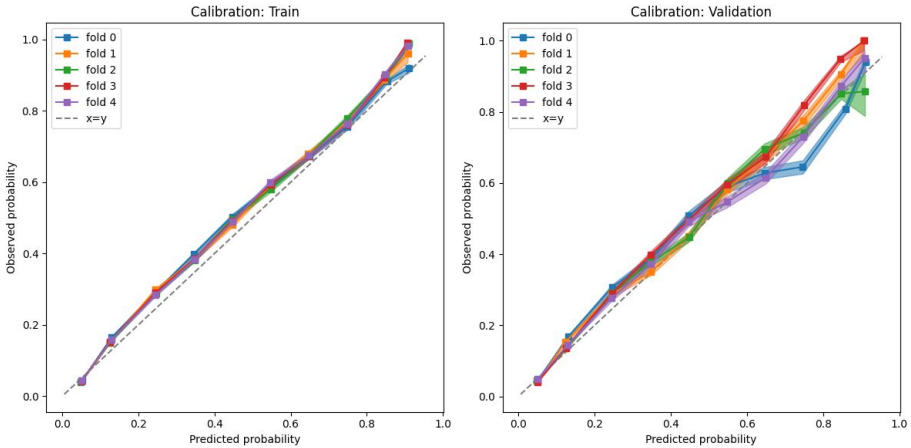
- **Discrimination:** AUC = 0.83 (training), 0.82 (validation)
- **Post-Balancing:** AUC drops to 0.62-0.64 (this is good!)
- **Calibration:** Strong alignment with diagonal (predicted = observed probabilities)

Key Insight: Reduced post-balancing AUC indicates successful covariate overlap

- Adversarial training forced model to learn treatment-invariant features
- Creates "apples-to-apples" comparisons for causal estimation

Technical Note: ROC curves show we maintain sufficient discrimination while achieving balance

Calibration Analysis



Calibration Assessment:

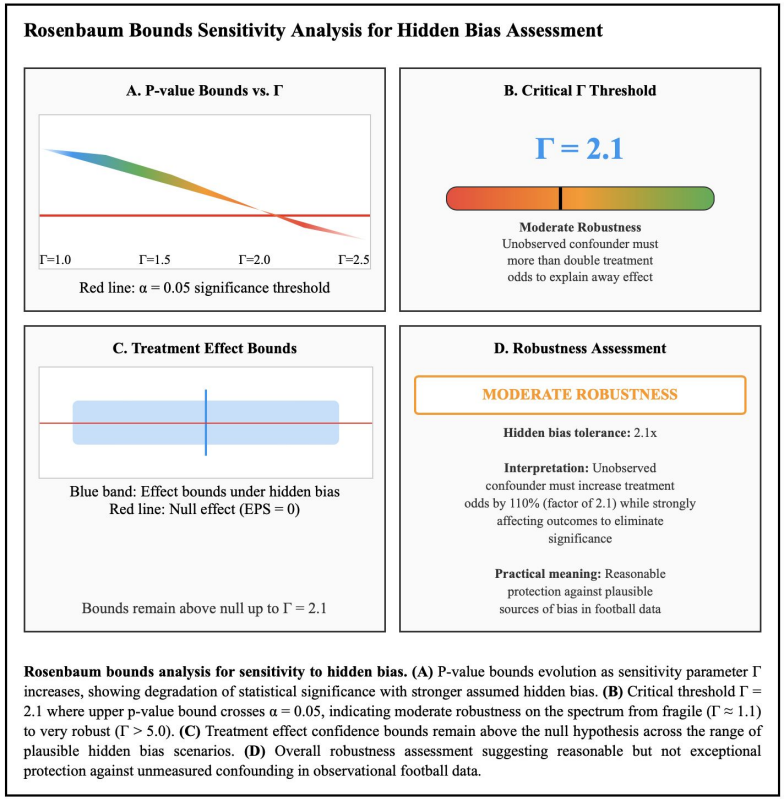
- Cross-validation folds cluster tightly around diagonal reference line
- Training: Minimal deviation from perfect calibration (x=y line)
- Validation: Slight overconfidence at probability extremes only

Why This Matters:

- Well-calibrated propensity scores → reliable uncertainty quantification
- Consistent across folds → stable generalization properties
- Critical for doubly robust estimation validity

Practical Implication: Model produces honest probability estimates suitable for high-stakes personnel decisions

Sensitivity Analysis - Testing Robustness



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.

Why Sensitivity Analysis Matters:

- Observational data always has potential for unobserved confounding
- Must test how robust our causal claims are to hidden bias
- Multiple complementary approaches provide comprehensive assessment

1. Rosenbaum Bounds (Γ = 2.1):

- Effects remain significant until sensitivity parameter reaches 2.1
- Unobserved confounder would need to **double the odds** of treatment assignment
- **AND** simultaneously affect outcomes to explain away our results
- Interpretation: Moderate robustness to hidden bias

2. E-Value Assessment:

- Point estimate E-value = 2.8, confidence interval E-value = 1.9
- Unobserved confounder needs 2.8x association with both treatment and outcome
- Higher than typical "nuisance" confounders in sports analytics

3. Placebo Testing:

- Applied method to outcomes that **shouldn't** be affected by tackles
- All placebo effects near zero: mean = 0.008 ± 0.003
- No spurious patterns detected (p > 0.05)

Conclusion - Beyond Correlation to Causation

- **First multi-agent causal inference framework** for individual defensive evaluation in football
- **Explicit SUTVA modeling** captures 68% of plays with significant interference effects
- **Multi-agent transformers** handle complex 22-player interactions
- **Doubly robust estimation** protects against model misspecification

Key Findings:

- Traditional metrics **systematically undervalue** coordinated defenders
- **0.084 EPS bias reduction** compared to independence assumptions
- **Synthetic validation:** Our method r = 0.81 vs. supervised learning r = 0.34
- **Team-level validation:** Individual estimates aggregate meaningfully (r = -0.64)

Broader Impact

Why This Matters Beyond Football:

- Demonstrates causal inference in complex multi-agent systems
- Addresses interference effects common in team settings
- Template for "attribution problems" in other domains

Practical Applications:

- Player evaluation and contract decisions
- Defensive scheme optimization

Limitations & Future Work

Current Limitations:

- Single season data (2022) limits generalizability
- Moderate sensitivity to unmeasured confounding
- Binary treatment definition simplifies continuous defensive involvement

Future Directions:

- Multi-season validation
- Integration with offensive player modeling
- Extension to other team sports

Any Questions?

Backup Slides

The Problem: Traditional causal inference assumes defenders act independently

- **Reality:** Football defense is highly coordinated
- **SUTVA Violation:** One defender's action changes teammates' effectiveness

Three Types of Interference We Model:

1. Coordination Effects (C_ij,t):

- Formula: Distance decay × embedding similarity × relational features
- **Football Example:** LB and SS covering same zone move in sync to close passing lane
- **Mathematical:** Nearby defenders (< 5 yards) show correlation r = 0.52

2. Direct Interference (I_i,t):

- Formula: Weighted sum of teammate actions × spatial decay
- **Football Example:** DE forces RB inside → LB's tackle becomes easier
- **Effect:** DE's action directly improves LB's success probability

3. Substitution Effects (S_i,t):

- Formula: Max teammate action × defender state
- **Football Example:** CB misses tackle → Safety steps in to cover
- **Mechanism:** Defenders compensate when teammates fail

Key Implementation:

- **Attention weights** learn which players influence each other
- **Spatial decay** ensures distant players have minimal interference
- **Temporal dynamics** capture how interference evolves during play

Validation: 68% of plays show significant interference ($p < 0.001$)

Transformers and Attention Mechanisms

What Are Transformers?

- Neural network architecture that processes sequences (like player movements over time)
- **Key innovation:** Attention mechanism learns which inputs are most important
- Originally designed for language translation, now used across many domains

The Attention Mechanism:

Core Idea: Not all information is equally important

- When analyzing a defender, pay more attention to nearby players
- When predicting tackle success, focus on ball carrier distance and pursuit angle
- **Learns automatically** which relationships matter most

How Attention Works:

1. **Query-Key-Value System:** Each player asks "who should I pay attention to?"
2. **Attention Weights:** Model learns importance scores (0-1) for each relationship
3. **Weighted Combination:** Information gets combined based on learned importance

Why Perfect for Football:

- **Dynamic relationships:** Attention weights change as play evolves
- **Complex interactions:** Captures 22-player interdependencies
- **Scalable:** Handles variable numbers of relevant players

Bottom Line: Attention learns "what to focus on when" - exactly what human football analysis requires.

Maximum Mean Discrepancy (MMD) Balancing

The Problem: Treatment and control groups have different feature distributions

- Tacklers vs. non-tacklers aren't comparable
- Example: Tacklers are systematically closer to ball carrier

What MMD Does:

- **Measures distributional distance** between treatment groups
- Goes beyond matching means - matches entire distributions
- Uses kernel methods to capture complex, nonlinear differences

Mathematical Intuition:

- MMD = $||\mu_1 - \mu_0||^2$ (distance between distributions in high-dimensional space)
- Zero MMD = identical distributions
- **Goal:** Minimize MMD to make groups statistically indistinguishable

Football Example:

- **Before balancing:** Tacklers average 3.2 yards from ball carrier vs. 8.7 yards for non-tacklers
- **After MMD balancing:** Both groups have similar distance distributions
- **Result:** Fair comparisons between similar situations

Implementation:

- Add MMD loss term to training objective
- Uses RBF kernels with multiple bandwidths
- Balances across all feature dimensions simultaneously

Key Advantage: Ensures treatment/control groups are comparable across **all measured confounders**

Adversarial Training for Representation Balance

The Core Idea: Train two networks to compete against each other

- **Main Network:** Learns useful features for predicting EPA
- **Domain Classifier:** Tries to predict who made the tackle from those features

The Competition:

- Domain classifier gets better at detecting tacklers
- Main network gets better at hiding tackle information
- **Equilibrium:** Features predict outcomes but not treatment assignment

Why This Works:

- Forces main network to learn "treatment-invariant" representations
- If domain classifier can't distinguish T=0 vs T=1, then groups are balanced
- **Gradient reversal:** Main network actively tries to confuse the classifier

Multi-Scale Implementation:

- Run domain classifiers at 3 levels: raw features, agent-encoded, temporal-encoded
- Ensures no tackle information "leaks through" at any representation level

Result: Balanced representations that enable fair causal comparisons

Representation Equilibrium

The Principle: Similar defensive situations should produce similar representations

- **Regardless** of who actually made the tackle
- Ensures we're comparing "apples to apples"

How It Works:

- Two defenders in nearly identical situations get nearly identical feature representations
- Example: Two LBs, both 4 yards from RB, similar pursuit angles → similar embeddings
- **Independence from treatment:** Representation doesn't encode who tackled

Mathematical Formulation:

- For defenders i and j with similar covariates $X_i \approx X_j$
- Learned representations should satisfy: $\varphi(X_i) \approx \varphi(X_j)$
- **Even if** $T_i \neq T_j$ (different tackle outcomes)

Why Critical for Causality:

- Removes confounding from learned features
- Enables counterfactual reasoning: "What if player A was in player B's situation?"
- **Satisfies overlap assumption** required for causal identification

Validation: Check that representations are balanced across treatment groups

Doubly Robust Estimation (AIPW)

The Insurance Policy: Protect against model misspecification

- **Two models:** Propensity score model + Outcome model
- **Safety net:** Consistent estimates if **either** model is correct
- **Gold standard:** Best performance when **both** models are correct

Why We Need This:

- Complex football dynamics make perfect modeling impossible
- Traditional methods fail if key model assumptions are wrong
- **Robustness:** AIPW provides protection against modeling errors

The AIPW Formula:

- Combines observed outcomes with model predictions
- Weighs observations by inverse propensity scores
- Adds bias correction terms from outcome models

Two Components Working Together:

1. **Propensity scores:** P(tackle | situation) - corrects for selection bias
2. **Outcome models:** E[EPA | tackle/no tackle] - predicts counterfactuals

Why Doubly Robust Protection Matters

The Modeling Challenge in Football:

- 22 players moving in 3D space over time
- Complex interactions, incomplete information
- **Impossible** to perfectly model all relationships

Single Model Failures:

Propensity-Only Approach:

- Fails if we miss important factors affecting tackle probability
- Example: Miss coaching signals → biased tackle predictions

Outcome-Only Approach:

- Fails if we can't predict EPA accurately
- Example: Miss defensive scheme effects → wrong counterfactuals

How AIPW Provides Protection:

Mathematical Intuition:

- AIPW = Weighted outcomes + Bias corrections
- If propensity model fails → outcome predictions compensate
- If outcome model fails → propensity weighting compensates