Causal Ad Incrementality without Experiments: Addressable-Audience CEM+DML at Billion-Impression Scale

Jaeyeong Kim
Tatari

Alexander J. Rasmussen
Tatari

Jake P. Reschke
Tatari

Abdullah M. Spall
Tatari

Abstract

Advertisers need to know how many conversions an ad campaign actually caused, but the standard tools—geo lift tests, public-service announcement controls, and ghost ads—are too slow, expensive, or privacy-restricted for routine use. We present an end-to-end observational pipeline that draws controls from the addressable audience, pairs each treated impression with controls under a symmetric per-pair attribution window, and estimates incrementality with coarsened exact matching plus double machine learning (or a sparsity-routed surrogate/IPW fallback). The novel construction draws controls from the **addressable audience** and pairs each treated impression with a **symmetric per-pair attribution window**; the remaining steps use standard estimators. Across eight production connected-TV deployments, the existing simulation-based incrementality estimator understates upper-funnel incrementality by **1.1**×–**33.8**× (DML as the reference), with the gap tracking the brand's existing-user pool. The pipeline agrees with independent synthetic-control geo lift tests within 95% CIs on three of four campaigns—including the only one sharp enough to discriminate—and passes all eight A/A placebos. On LaLonde–Dehejia–Wahba PSID-1, the R-learner under CEM matching is within **1.4**% of the experimental treatment effect—an order of magnitude tighter than the published state of the art (**15.7**% AIPW-GRF [13]); on the same PSID-1 data where Imbens and Xu's eleven propensity-trimmed estimators all yield negative ATT estimates, our post-CEM meta-learners retain 78% of the treated and stay positive near the $1,794 experimental ATT—suggesting their trim itself shifts the target effect.

Keywords

incrementality, causal inference, double machine learning, coarsened exact matching, advertising measurement, CTV, observational study, surrogate index

1 Introduction

Every advertiser faces the same question: *how many conversions would not have happened without the ad?* Deterministic attribution conflates correlation with causation, leaving a gap to true incrementality that is large, variable across industries, and unpredictable in direction [11]. Ghost ads [14] are user-level randomized controlled trials (RCTs) but require platform cooperation; public-service announcement (PSA) controls lose randomization symmetry under optimizer routing; geo lift tests [4] are quasi-experimental with synthetic-control counterfactuals [1, 10]. All three share four limits as a *primary* measurement tool: cost (holdouts sacrifice incremental revenue), speed (the unfavorable lift-vs-variance economics of ad measurement [19] imposes a duration floor), privacy (user-level randomization conflicts with cookie deprecation and GDPR), and scalability (a platform serving hundreds of clients cannot run concurrent geo lift tests for all), leaving most advertisers without a continuous causal measurement option. This paper introduces a **Coarsened Exact Matching plus Double Machine Learning (CEM + DML) pipeline** that estimates causal incrementality from impression logs alone, using the addressable audience as the source of controls and scaling to billions of impressions.

The simulation-based incrementality (SBI) estimator we benchmark against at Tatari [27] is a counterfactual baseline simulated from a 5-minute post-impression traffic window, scaled by a calibrated multiplier. Such simulation-based methods produce campaign-level totals but cannot deliver household-level treatment effects or principled confidence intervals (CIs). Table 3 (§3.2) quantifies how the SBI estimate diverges from the DML causal estimate on the upper-funnel (visit) estimand (causal target). Even where RCTs are feasible, the unfavorable lift-vs-variance economics produce CIs too wide to be informative [19]; where synthetic-control geo lift *is* available, our DML estimates land at or below the geo midpoint on every measurable-lift test (§3.1). Because no observational study fully rules out unmeasured confounding, we triangulate against three references: geo lift agreement, 8/8 A/A placebos covering zero, and public-data RCT reconstruction on LaLonde–Dehejia–Wahba.

This paper's novelty is an integrated, four-stage pipeline for observational causal incrementality: **(1) exposure logs → (2) addressable-audience controls (from other advertisers' impressions) → (3) CEM matching with per-pair attribution windows → (4) DML / surrogate / inverse propensity weighting (IPW) estimation** [7, 22]. Stages 1 and 4 use standard estimators; to our knowledge, **the addressable-audience plus per-pair attribution-window construction (stages 2–3) is novel** among published ad-measurement pipelines. Each treated impression's per-pair attribution window runs to the same user's next impression or to study end, whichever comes first, and is copied onto every matched control (symmetric, non-overlapping within an individual). Stage 2 is data-source agnostic (impression logs, platform user data, eligibility logs, or identity-graph cooperation). **Existing observational ad-measurement work covers parts of this pipeline along three threads, but no published method combines all four stages.** Diemert et al. [9] apply DML to uplift modeling (stage 4) but assume the treated/control split is given; Gordon et al. [11] benchmark observational matching against Facebook RCTs (stages 3–4) without addressable-audience derivation or per-pair windows; and Sapp

et al.'s *Near Impressions* [25] match on upstream auction events and credit all downstream conversions, lacking per-impression window truncation and tied to auction-derived inputs unavailable in streaming or linear CTV. CEM on pretreatment activity and demographics enforces empirical comparability across treated and control cohorts; outcome sparsity is routed automatically between direct DML, surrogate index [2], and IPW.

Contributions. (1) Production deployment finding. On eight anonymized campaigns, the SBI understates upper-funnel causal incrementality by **1.1**× to **33.8**× (DML as the reference), with the magnitude tracking the brand's existing-user pool (§4.1). All eight upper-funnel A/A tests pass, and the operational pipeline agrees with three of four geo lift tests on the conversion estimand—including the only test sharp enough to discriminate among plausible point estimates. **(2) Methodological identification.** We introduce audience-level CEM with impression-time anchoring—an observational construction using only impression logs and an addressable-audience set. On LaLonde–Dehejia–Wahba, every cross-fit meta-learner under CEM stays within \$1,374 (PSID-1) and \$1,139 (CPS-1) of the experimental average treatment effect on the treated (ATT), with R- and DR-learner co-leading at 1.4% and 3.8% error on PSID-1 (§3.3). **(3) System.** We implement the pipeline natively on PySpark for scalability, composing CEM, cross-fit DML, and the sparsity-routed surrogate/IPW fallback, and reusing a shared cross-fit propensity across all outcomes. It is the only implementation that completes at $n{=}10^7$ rows per group, where the single-node reference EconML [5] times out, and is **8**× faster even at $n{=}10^6$ where both complete (Table 1).

2 Method

2.1 Control Group: CEM from the Addressable Audience

The core challenge in estimating ad incrementality from observational data is constructing a valid counterfactual: *what would exposed users have done without the ad?* Estimating an average treatment effect requires the standard triple of *consistency* (each unit's observed outcome equals its potential outcome under the treatment received), *exchangeability* ($\{Y(0), Y(1)\} \perp T \mid X$, where $Y(t)$ is the potential outcome under treatment $t \in \{0, 1\}$, T the realized treatment, and X the pretreatment covariates—no unmeasured confounding given X), and *positivity* (every covariate cell contains both treated and untreated units; $0 < P(T{=}1 \mid X{=}x) < 1$ for all x). Naive comparison of ad-exposed to all unexposed users violates positivity: many unexposed users never appeared to the auction or differ on unobservable selection criteria, and in every covariate cell where the unexposed pool is empty there is no common support—no propensity-score model can rescue identification by reweighting across regions with no training data [16]. We instead draw controls from the **addressable audience**: users who received *other* advertisers' impressions during the same period but were not exposed to the target campaign. These users share the same selection mechanism (being in the ad-serving pool) and are active on the same platform at the same times, which substantially reduces the primary source of selection bias. Population-level positivity is not guaranteed: some addressable users still have $P(T{=}1 \mid X) = 0$ for any given campaign (targeting filters, frequency caps, creative restrictions), so CEM enforces *empirical* positivity by dropping covariate cells with no treated or no control overlap. The resulting estimand is the matched-subset ATT, a directly actionable budget target.

We match treatment and control users via **CEM** [12] at the device level on pre-treatment brand activity (1/7/28-day rolling sessions on the target advertiser's site), cross-platform activity (sessions on other advertisers' sites), and third-party demographics. CEM is preferred over propensity score matching (PSM) and nearest-neighbor matching on four grounds. (i) CEM vectorizes as a groupby-bucket-join on Spark; PSM and nearest-neighbor require $O(N^2)$ pairwise distance computation or sequential greedy pairing, both infeasible at billion-impression scale. (ii) It makes positivity violations explicit by dropping cells with no available controls rather than smoothing them via a propensity fit. (iii) CEM belongs to the Monotonic Imbalance Bounding class [12], so the user-chosen coarsening puts an explicit upper bound on post-match covariate imbalance, sidestepping the King–Nielsen propensity-score paradox [16] in which PSM can *increase* imbalance on non-propensity covariates. (iv) CEM separates overlap enforcement (upstream) from nuisance residualization (downstream DML), so the propensity inside the R-learner is not doubly burdened by also driving treatment–control pairing.

CEM bins each covariate via *zero-aware dyadic-tail quantile binning*: a dedicated zero bin plus a positive tail at geometrically halving quantile edges $p_k = 1 - 2^{-k}$. A cascade coarsening starts at 6 bins and steps down to 4, stopping at the first bin count for which the unmatched treatment fraction falls below 2%. Within each surviving cell, matching is $1{:}k$ ($k{=}5$ default): each treated unit takes k controls by modular stride, with replacement when the cell has fewer than k per treated unit; the duplication ratio is logged. After matching, each treatment impression's anchor and per-pair attribution-window endpoint are copied to all k matched controls (symmetric windows within pair), and impression-time rolling features are computed relative to the anchor. Augmented inverse-propensity-weighted (AIPW) standard errors (SEs) use resolved-id cluster-robust variance and the row A–B bootstrap is IP-clustered for the with-replacement dependence.

2.2 Double ML Estimation

Given matched anchors with features X, binary treatment T, and outcome Y (with treated sample size n_T and control sample size n_C), we estimate conditional average treatment effects (CATEs) within the DML framework of Chernozhukov et al. [7]. DML with cross-fitting yields $\sqrt{n}$-consistent estimates and standard normal-theory CIs when the nuisance models $\hat{e}(x)$ (propensity) and $\hat{\mu}(x)$ (outcome) converge faster than $n^{-1/4}$ in probability—flexible ML estimators suffice where regression, PSM, and IPW would require correctly specified parametric models. A *meta-learner* turns an arbitrary regression algorithm into a treatment-effect estimator by fitting separate nuisance models for treatment and outcome and combining them; the R, S, T, X, and DR variants differ in how they combine. We adopt the **R-learner** [22], a meta-learner related to the causal-forest framework [3] that is naturally suited to heterogeneous-effect estimation by residualized regression:

(1) **Propensity** $\hat{e}(x) = \hat{P}(T{=}1 \mid X{=}x)$ fitted once via $K{=}3$ cross-fitting and shared across all outcomes.
(2) **Outcome** $\hat{\mu}(x) = \hat{E}[Y \mid X{=}x]$ fitted per outcome via cross-fitting.
(3) **CATE** via out-of-fold residuals $\hat{T}_{\text{res},i} = T_i - \hat{e}(X_i)$, $\hat{Y}_{\text{res},i} = Y_i - \hat{\mu}(X_i)$ and the weighted regression

$$\hat{\tau} = \arg\min_{\tau \in \mathcal{F}} \sum_i \hat{T}^2_{\text{res},i} \left(Z_i - \tau(X_i)\right)^2, \quad Z_i = \hat{Y}_{\text{res},i} / \hat{T}_{\text{res},i},$$

with $\mathcal{F}$ a flexible function class (a cross-fit Spark Random Forest in our implementation).

The $\hat{T}^2_{\text{res}}$ weighting down-weights observations with weak identification. ATE/ATT estimates are accompanied by AIPW confidence intervals [24] from a single influence-function (IF; analytic SE) pass; total incremental impact is $\hat{\text{ATT}} \times n_T$. The same fit also returns user-level $\hat{\tau}(x)$ for uplift modeling at the marginal cost of one extra weighted regression [3, 9, 17]. We benchmark R against S-, T-, X- [17], and DR-learners [15] in §3.3 and adopt R as the operational default on runtime grounds.

2.3 Three-Way Sparse Outcome Routing

Down-funnel conversions (purchases, sign-ups) are often extremely rare: the control-group conversion rate p is typically 10^{-5} to 10^{-3}, which makes direct DML unreliable. The framework automatically routes each outcome by its treatment-group conversion count into one of three regimes. **Dense** outcomes use **Direct DML** (full R-learner). **Sparse** outcomes use **Surrogate DML**, where an upstream signal M (the mediator—a site-visit indicator in our setting) feeds a surrogate model $\hat{g}(x,t,m) = \hat{P}(Y{=}1 \mid X, T, M)$ that replaces the sparse binary conversion outcome with a dense continuous proxy [2]; DML is then applied to the surrogate outcome. **Very sparse** outcomes use **IPW**, Hajek-normalized (weights divided by their sum for stability), with Fisher's exact test and bootstrap CIs. Thresholds are tuned per deployment and held fixed across reported runs. The surrogate-index approach substantially reduces variance for rare events [6] under a surrogacy condition (the surrogate M fully mediates the treatment-to-outcome path)—in its strongest form the Prentice criterion $Y \perp T \mid M, X$, or the composite surrogacy-plus-comparability condition of Athey et al. [2]. Surrogacy is approximately satisfied for the ad → site-visit → conversion chain; brand-awareness paths that bypass the observed surrogate [20] attenuate surrogate-routed estimates toward zero, and the routing rule restricts surrogate use accordingly. Formal calibration of the routing thresholds and a surrogacy diagnostic (comparing surrogate-DML to direct DML on dense outcomes where both routes are valid) are left to future work.

2.4 Implementation

The pipeline is implemented natively on PySpark, with the shared cross-fit propensity reused across every outcome model (~40% runtime reduction in the multi-outcome regime) and per-outcome fits parallelized via a driver-side `ThreadPoolExecutor`. A typical analysis samples 5M impressions per treatment/control group; at this scale, EconML [5] exhausts memory and wall-clock budgets (Table 1). Table 1 reports wall-clock on a Nie–Wager [22] Setup B synthetic data-generating process (DGP) across three sample sizes on a fixed 10× `i3.4xlarge` cluster: EconML times out at $n = 10^7$ while the PySpark R-learner completes in ~18 minutes. Precision in estimation of heterogeneous effects (PEHE) and |ATT err| agree within ±15% and a factor of 1.3 at the overlap scales, so the PySpark implementation is methodologically equivalent where direct comparison is feasible and the only feasible option at larger scales.

Table 1: Wall-clock comparison on Nie–Wager Setup B synthetic DGP, fixed 10× `i3.4xlarge` cluster. Mean runtime across 5 replications.

n per group	EconML (single-node)		PySpark
	`NonParamDML`	`DRLearner`	R-learner
10^5	91 s	92 s	59 s
10^6	996 s	1,013 s	127 s
10^7	timeout	timeout	1,098 s

3 Empirical Results

We validate the framework on CTV advertising data from anonymized advertisers (mostly direct-to-consumer, DTC, brands) in ten industries. The SBI comparison and A/A placebo cover eight advertisers on the upper-funnel estimand; the geo lift test validation uses DTC Fashion (all-purchase and new-customer cohorts), DTC Pet Food, and DTC Vitamins on the conversion estimand, where independent geo lift tests built on the synthetic control method (SCM) are available. We present the geo lift test validation first to establish DML credibility against an external benchmark before adjudicating the SBI.

3.1 Geo-Test Validation

We validate against independent geo lift test estimates on three advertisers (Table 2). In addition to the operational pipeline (row D, CEM + R-learner), we report the full 2 × 3 {sampling × estimator} grid so that geo-CI agreement decomposes into the contributions of CEM, cross-fit residualization, and IPW reweighting. The comparison geo lift tests use an SCM with elastic-net donor weights (Doudchenko–Imbens variant) [1, 10]; treated designated market areas (DMAs) are advertiser-selected (quasi-experimental), and we treat the SCM lift as an independent second estimator rather than ground truth. The geo 95% CIs are percentile bounds from a stationary block bootstrap [23] of pre-period residuals (Politis–White automatic block length), pooled across $D{=}100$ random donor subsamples.

Headline (operational pipeline, row D). Only DTC Fashion (all-purchase) has a geo CI tight enough to discriminate among plausible point estimates; the operational pipeline lands inside it at −13.6% from the midpoint. The two near-null tests (DTC Pet Food, DTC Vitamins) have CIs spanning zero or factor-of-13 wide, where any reasonable observational estimator clears the agreement bar by default. The DTC Fashion (new-customer) cohort is the one informative miss (7.5% below the geo CI lower bound), recovered inside the CI by both non-residualized CEM variants (B, F)—localizing the miss to the cross-fit residualization stage, not CEM.

Table 2: Observational methodology sweep vs. SCM-based geo lift test on incremental conversions over the per-pair attribution window; row D is the operational pipeline. Cells: *point [low, high]*; ✓/✗ marks inclusion in the geo 95% CI. Observational CIs: 90% IP-cluster bootstrap (n_{boot}=500, rows A–B); 95% AIPW IF [24] (rows C–D); 95% IPW IF [21] (rows E–F).

Method (engineer pipeline)	CI method	DTC Pet Food ($n_T \approx 8.6$M)	DTC Fashion (all-purchase) ($n_T \approx 3.8$M)	DTC Fashion (new-customer) ($n_T \approx 3.8$M)	DTC Vitamins ($n_T \approx 2.2$M)
Geo SCM (95% CI)	Block boot.	−66 [−240, +123]	987 [572, 1,413]	479 [308, 603]	61 [9, 116]
A. Naive + Diff-in-means	Bootstrap	419 [375, 466] ✗	2,909 [2,767, 3,068] ✗	1,279 [1,195, 1,358] ✗	99 [80, 119] ✓
B. CEM + Diff-in-means	Bootstrap	246 [191, 297] ✗	1,043 [888, 1,207] ✓	409 [313, 508] ✓	77 [56, 98] ✓
C. Naive + R-learner	AIPW IF	−85 [−257, +87] ✓	781 [411, 1,151] ✓	28 [−245, +301] ✗	69 [31, 108] ✓
D. CEM + R-learner (op.)	AIPW IF	−88 [−192, +16] ✓	853 [655, 1,050] ✓	285 [158, 412] ✗	26 [−14, +66] ✓
E. Naive + IPW	IPW IF	−52 [−224, +120] ✓	912 [541, 1,282] ✓	14 [−259, +288] ✗	74 [35, 112] ✓
F. CEM + IPW	IPW IF	−7 [−112, +97] ✓	931 [733, 1,128] ✓	342 [214, 469] ✓	37 [−4, +77] ✓

Methodology decomposition (rows A–F). The sweep isolates three contributions. **CEM matching (A→B) is the first-order correction:** it collapses the difference-in-means (DiM) overcount on every non-null test, falls inside the geo CI on 3/4, and identifies selection bias as the first-order error (§3.3). **Cross-fit residualization (B→D) refines but does not replace matching:** it shifts the appreciable-lift estimates toward and below the geo midpoint and pulls DTC Pet Food to a near-null inside the zero-covering CI, at mildly wider AIPW CIs. **IPW vs. direct DML under CEM (D vs. F) agree to within** 11–81 **units.**

Why DML lands at or below the geo midpoint. DML is at or below the geo midpoint on all four tests because SCM-based geo lift tests count conversions from *all* users in treated DMAs while the operational pipeline restricts to addressable, matched users, and pre-treatment-activity matching attenuates the matched-subset ATT below marginal lift. We therefore report geo agreement as directionally consistent. All six observational rows share the per-pair window, so the sweep isolates CEM and estimator contributions but does not ablate the window itself; a CEM + naive-window comparison is left to future work.

3.2 Operational-Stack Validation: A/A and SBI Comparison

Having established that DML agrees with the SCM-based geo lift test on the conversion estimand (§3.1), we now turn to the operational-stack validation on the upper-funnel (site visit) estimand for eight production-scale advertisers. Table 3 reports two checks per industry: an A/A placebo test (does the pipeline produce a near-zero estimate under no real treatment?) and an SBI comparison.

Estimand and protocol. The SBI's 5-minute baseline + short-tail multiplier estimates *upper-funnel* (visit) incrementality only, so this validation uses visit lift and reserves conversion lift for the geo lift test validation (§3.1). Each A/A splits a 20% control holdout into pseudo-treatment/control through the full DML pipeline and passes if the AIPW 95% CI covers zero. Table 3 sets the SBI's 28-day estimate [27] alongside the DML causal estimate at the matched $\hat{\text{ATT}} \times n_T$ scale.

Filter rule for the two starred advertisers. DTC Wellness and Mobile App report DML on a post-filter impression population that excludes a small set of 2-flag-intersection ID-resolution cells producing implausibly negative treated lift at the 7-day window; the same filter is applied to the A/A run (an identity-graph alternative is in §5).

Headlines. All eight A/A tests pass; every 95% CI covers zero. The absolute placebo impact is at most about 100 visits against operational DML estimates of 9K–2.04M visits, ruling out A/A false positives. DML reports between **1.1**× and **33.8**× the SBI's 28-day estimate across the eight advertisers—a quantified disagreement between two internal estimators, not a ground-truth refutation of the SBI. Unlike the geo comparison (DML within the CI band, at or below its midpoint), the SBI disagreement is one-sided: DML is uniformly higher, and its size appears to track the brand's existing-user pool relative to new-user reach (a hypothesized mechanism; §4.1).

3.3 LaLonde–Dehejia–Wahba No-Control Reconstruction

LL-DW [8, 18] is the canonical public-data instance of the no-control problem that CTV measurement faces: a treated population's natural controls are discarded and replacements are drawn from a structurally non-overlapping pool, directly analogous to drawing CTV controls from the addressable audience. Experimental ground truth comes from the National Supported Work (NSW) randomized trial (ATT $1,794 on 185 treated men). LL-DW discards the experimental controls and swaps in external observational pools (PSID-1, n_C=2,490; CPS-1, n_C=15,992). We run all seven meta-learners on both variants under three conditions—*raw*, *CEM-cut* (matched-subset ATT), and *CEM-weighted* (full-pool ATT via smoothed inverse-frequency weighting)—plus a DiM baseline, 100 bootstrap replications (Table 4).

CEM is doing most of the work here. Matching alone reduces the DiM |ATT err| by 84%/81% on PSID-1/CPS-1, isolating the correction from any model-based step. Adding CEM-cut on top of each meta-learner cuts mean |ATT err| across the seven cross-fit and EconML estimators by another 49%/43% and compresses the cross-estimator spread 4×/3×—every estimator now within $1,374/$1,139 of the experimental ATT, with EconML's `NonParamDML` on PSID-1 as the lone exception (it degrades slightly). The matched-sample reduction is also variance-friendly: CEM retains 78%/85% of NSW-treated but only a small fraction of external controls (Table 4, n used row), removing high-variance no-overlap cells with median Δstd +28%/+9%—well below the $\sqrt{12} \approx 3.5\times$ inflation expected from random subsampling.

Table 3: Operational-pipeline validation on eight production-scale advertisers; upper-funnel (28-day visit) estimand. SBI: scaled view-through (VT) counts and the simulation-based incremental estimate (5-min baseline × default 2.2 [27]). DML: $\hat{\text{ATT}} \times n_T$, post-filter for ∗-marked advertisers. A/A: 28-day relative visit lift on the placebo split.

Advertiser	SBI 7d VT	SBI 28d VT	SBI 28d Incr. (sim.)	DML 7d Incr.	DML 28d Incr.	DML/SBI (28d)	A/A 28d Rel. Lift
Fintech A	5.34M	7.13M	78.9K	1.10M	2.04M	25.8×	+0.92%
Fintech B	705K	1.08M	116.2K	188K	226K	1.9×	−4.70%
DTC Health	3.46M	5.60M	534.3K	1.21M	1.90M	3.6×	+2.17%
DTC Furniture	300K	612K	123.1K	93K	147K	1.2×	+7.84%
DTC Wellness*	1.62M	2.43M	126.6K	115K	137K	1.1×	+1.12%
B2B SaaS	392K	695K	6.5K	154K	220K	33.8×	+2.87%
Mobile App*	179K	339K	3.3K	3.9K	8.9K	2.7×	−0.39%†
DTC Retail	533K	825K	47.4K	188K	286K	6.0×	+1.74%

† Mobile App's A/A subsample triggers a 100% filter dropout (small-N cascade), so its A/A is reported unfiltered—a diagnostic failure mode of the post-hoc filter and one motivation for the identity-graph extension (§5).

Table 4: LaLonde–Dehejia–Wahba no-control reconstruction; mean across 100 bootstrap reps. Top: ATT mean (truth $1,794); bottom: |ATT err| (mean absolute error per rep). CEM-c = CEM-cut, CEM-w = CEM-weighted (§3.3). R/S/T/X/DR-learner are PySpark cross-fit; NonParamDML and DRLearner (EM) are EconML references. "n used (T+C)" is the across-rep mean of CEM-matched rows; it lies below the deterministic raw-pool CEM count (291/2,031 on PSID-1/CPS-1) because matched cells may lose treatment or control under resampling.

	PSID-1			CPS-1		
Estimator	Raw	CEM-c	CEM-w	Raw	CEM-c	CEM-w
ATT mean ($, truth = $1,794)						
DiM baseline	−15,177	−986	−8,397	−8,486	−166	−5,542
R-learner	939	1,819	1,645	897	1,998	1,872
S-learner	−808	420	−169	−115	669	124
T-learner	−1,513	1,372	543	−481	1,708	1,035
X-learner	−1,137	1,336	529	501	1,852	1,230
DR-learner	−69	1,726	1,352	1,057	1,895	1,502
NonParamDML	2,156	1,359	1,908	1,589	2,384	2,246
DRLearner (EM)	611	886	612	1,106	1,870	970
|ATT err| ($, lower is better)						
DiM baseline	16,971	2,787	10,191	10,280	1,984	7,336
R-learner	1,028	922	848	1,004	648	658
S-learner	2,602	1,374	1,963	1,909	1,139	1,670
T-learner	3,307	863	1,280	2,275	703	849
X-learner	2,931	860	1,292	1,298	643	708
DR-learner	1,873	810	868	892	673	732
NonParamDML	825	1,078	828	1,109	822	792
DRLearner (EM)	1,248	1,089	1,526	964	741	1,244
n used (T+C)	2,675	227	2,675	16,177	1,343	16,177

On the resulting matched subset, R- and DR-learner co-lead at 1.4% and 3.8% on PSID-1 and 11.4% and 5.6% on CPS-1; the four post-CEM cross-fit meta-learners (R, T, X, DR) are statistically indistinguishable on |ATT err|. Against Imbens and Xu's 15.7% AIPW-GRF on PSID-1 [13][1], **our R-learner result on PSID-1 is a 4–11× point-estimate lead** at n_T=185, where all CIs are wide. The matched-population picture is sharper still: on the same PSID-1 data, Imbens and Xu's propensity trim drops a substantial fraction of NSW-treated, shifting the experimental ATT from $1,794 to $306, and every one of their eleven post-trim estimators then yields a *negative* ATT estimate (their Figure 2 reports nine; the tutorial's Table B1 adds two more) [13]. CEM-cut keeps 78% of NSW-treated on PSID-1 (Table 4, n used row), and every post-CEM cross-fit meta-learner stays positive, with R- and DR-learner landing within $25 and $68 respectively of the $1,794 experimental ATT. On CPS-1 our R/DR and their NN matching (3.6%) bootstrap CIs symmetrically subsume each other's point: a competitive tie.

Weighted-CEM robustness check. Under weighted CEM (full-pool estimand, directly comparable to Imbens and Xu), R-learner attains **8.3**%/**4.4**% ATT error on PSID-1/CPS-1—beating their best PSID-1 result (15.7% AIPW-GRF) and tying their best CPS-1 result (3.6% NN matching). R- and DR-learner are robust to both CEM variants; R-learner is the operational default on runtime grounds.

4 Discussion

4.1 When SBI Fails Most

The SBI-vs-DML gap in Table 3 sorts the eight advertisers into three clusters consistent with brand-level heterogeneity in TV ad effectiveness [26]: the gap tracks the brand's existing-user pool relative to its new-user reach. The mechanism is in the SBI's baseline: the 5-minute pre-impression window samples the advertiser's existing engagement pool, whose event density exceeds the addressable audience's, so the inflated baseline depresses simulated lift and the global ×2.2 drag multiplier cannot absorb brand-specific delayed re-engagement. The DML/SBI bands (∼ 1–3×, ∼ 3–6×, ∼ 25–34×) align with brand-pool size; no global multiplier calibrates across this 30× spread, and the conversion-side gap is bounded above by the visit-side gap, making geo agreement (§3.1) consistent with the SBI finding.

4.2 Generalizability, Privacy, and Scope

The addressable audience is built from *cross-advertiser* impressions, so this is a platform-level method: it fits operators that see many advertisers' impressions (CTV measurement platforms, DSPs, walled gardens), not a standalone advertiser. Such platforms gain continuous incrementality measurement for every client without a

[1]Estimates from their replication tutorial at https://yiqingxu.org/tutorials/lalonde/.

holdout that forgoes that client's bidding opportunity. We validate on CTV; DSP-served display extends naturally. On privacy, the pipeline uses only logs already collected for ad delivery—no new data collection, randomization, or spend manipulation—runs in-platform on resolved household/device IDs, and fixes the cohort at ingest so identifier churn needs no retroactive re-resolution.

4.3 Limitations

Five limitations bear on interpretation. **Unmeasured confounding:** identification relies on exchangeability given the addressable-audience covariates; if device-level intent correlates with both addressability and outcome, ATT is biased upward. Four facts bound the residual: (a) DML sits at or below the geo midpoint on every measurable-lift test (§3.1), opposite to upward bias; (b) 8/8 A/A placebos cover zero (§3.2); (c) the active-user mechanism [19] inflates the SBI's pre-impression baseline, not our activity-matched estimator, so it drives the SBI understatement we observe; and (d) on LL-DW PSID-1 CEM + R recovers the experimental ATT within 1.4% (§3.3). A formal Rosenbaum sensitivity analysis on production data is left to future work; Imbens and Xu's [13] PSID-1 placebo analyses partially substitute. **LL-DW → CTV transfer:** the public benchmark establishes CEM-DML where ground truth exists; transfer to CTV is by structural analogy. **Third-party data noise:** geo-IP vendor disagreement attenuates the geo lift estimate, and the post-hoc 2-flag-intersection ID filter (Table 3 starred rows) is a coarse stand-in for the continuous identity-graph features future work will fold into the nuisance models; in small-N subsamples it can cascade to 100% dropout (Mobile App A/A, Table 3 footnote). **Cross-advertiser exposure as the only audience definition:** we validate using other advertisers' impressions; identity-graph and platform-user-pool variants share the input contract but are not benchmarked here. **Open ablations:** a CEM + naive-window comparison, more geo tests, routing-threshold sensitivity, and formal surrogacy diagnostics are deferred to future work.

5 Conclusion

We presented an integrated CEM + DML pipeline for observational ad incrementality. On LL-DW PSID-1, the R-learner under CEM beats the published state-of-the-art AIPW-GRF result (15.7% [13]) under both CEM-cut (1.4%, matched-subset) and CEM-weighted (8.3%, full-pool); on the same PSID-1 data, Imbens and Xu's eleven propensity-trimmed estimators all turn negative while ours stay positive near the $1,794 experimental ATT. All eight A/A placebos pass, the SBI understates upper-funnel incrementality by up to 33.8×, and the operational pipeline agrees with the only discriminating geo lift test. Future work: three-way sample splitting for the surrogate, continuous-treatment extensions, and identity-graph features in the DML nuisance models.

Acknowledgments

The authors used Anthropic's Claude Opus 4.7 for drafting.

References

[1] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *J. Amer. Statist. Assoc.* 105, 490 (2010), 493–505.

[2] Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. 2025. The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely. *The Review of Economic Studies* (2025), rdaf087.

[3] Susan Athey, Julie Tibshirani, and Stefan Wager. 2019. Generalized Random Forests. *The Annals of Statistics* 47, 2 (2019), 1148–1178.

[4] Joel Barajas, Tom Zida, and Mert Bay. 2020. Advertising Incrementality Measurement Using Controlled Geo-Experiments: The Universal App Campaign Case Study. In *Proceedings of the AdKDD Workshop at KDD.*

[5] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. 2019. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/py-why/EconML.

[6] Jiafeng Chen and David M Ritzwoller. 2023. Semiparametric Estimation of Long-Term Treatment Effects. *Journal of Econometrics* 237, 2 (2023), 105545.

[7] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal* 21, 1 (2018), C1–C68.

[8] Rajeev H Dehejia and Sadek Wahba. 1999. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *J. Amer. Statist. Assoc.* 94, 448 (1999), 1053–1062.

[9] Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. 2018. A Large Scale Benchmark for Uplift Modeling. In *Proceedings of the AdKDD and TargetAd Workshop at the 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).*

[10] Nikolay Doudchenko and Guido W Imbens. 2016. Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. https://arxiv.org/abs/1610.07748. arXiv:1610.07748 [stat.ME] NBER Working Paper No. 22791.

[11] Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2019. A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook. *Marketing Science* 38, 2 (2019), 193–225.

[12] Stefano M Iacus, Gary King, and Giuseppe Porro. 2012. Causal Inference Without Balance Checking: Coarsened Exact Matching. *Political Analysis* 20, 1 (2012), 1–24.

[13] Guido W Imbens and Yiqing Xu. 2024. LaLonde (1986) after Nearly Four Decades: Lessons Learned. arXiv:2406.00827 [econ.EM]

[14] Garrett A Johnson, Randall A Lewis, and Elmar I Nubbemeyer. 2017. Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness. *Journal of Marketing Research* 54, 6 (2017), 867–884.

[15] Edward H Kennedy. 2020. Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects. arXiv:2004.14497 [math.ST]

[16] Gary King and Richard Nielsen. 2019. Why Propensity Scores Should Not Be Used for Matching. *Political Analysis* 27, 4 (2019), 435–454.

[17] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 4156–4165.

[18] Robert J LaLonde. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review* 76, 4 (1986), 604–620.

[19] Randall A Lewis and Justin M Rao. 2015. The Unfavorable Economics of Measuring the Returns to Advertising. *The Quarterly Journal of Economics* 130, 4 (2015), 1941–1973.

[20] Jura Liaukonyte, Thales Teixeira, and Kenneth C Wilbur. 2015. Television Advertising and Online Shopping. *Marketing Science* 34, 3 (2015), 311–330.

[21] Jared K Lunceford and Marie Davidian. 2004. Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study. *Statistics in Medicine* 23, 19 (2004), 2937–2960.

[22] Xinkun Nie and Stefan Wager. 2021. Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *Biometrika* 108, 2 (2021), 299–319.

[23] Dimitris N Politis and Joseph P Romano. 1994. The Stationary Bootstrap. *J. Amer. Statist. Assoc.* 89, 428 (1994), 1303–1313.

[24] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *J. Amer. Statist. Assoc.* 89, 427 (1994), 846–866.

[25] Stephanie Sapp, Jon Vaver, Jon Schuringa, and Steven Dropsho. 2017. *Near Impressions for Observational Causal Ad Impact.* Technical Report. Google Inc.

[26] Bradley T Shapiro, Günter J Hitsch, and Anna E Tuchman. 2021. TV Advertising Effectiveness and Profitability: Generalizable Results from 288 Brands. *Econometrica* 89, 4 (2021), 1855–1879.

[27] Tatari, Inc. 2024. System and Method for Quantification of Latent Effects on User Interactions with an Online Presence Resulting from Content Distributed through a Distinct Content Delivery Network. U.S. Patent No. 12,165,169.