Incrementality Prediction: Synergizing Past Experiments for Intervention Personalization [*]

Ta-Wei Huang   Eva Ascarza   Ayelet Israeli

November 15, 2025

Abstract

Firms routinely conduct randomized experiments to evaluate the causal impact of marketing interventions. Yet these experiments are often analyzed in isolation, estimating treatment effects separately for each intervention. This limits firms' ability to leverage accumulated experimental knowledge. We develop a two-stage causal machine learning framework that integrates data across experiments to estimate the *conditional incrementality function* — treatment effects expressed as a function of intervention and customer attributes. In the first stage, we construct doubly robust scores within each experiment to obtain unbiased treatment effect signals. In the second, we employ a deep representation learning architecture that predicts these scores from intervention features and customer covariates, balancing flexibility with control of model complexity. The framework improves targeting efficiency and generalizes treatment effect predictions to new or untested interventions.

Combining data from 362 promotion experiments involving more than seven million customers, we show that synergizing these experiments through our framework substantially outperforms conventional approaches that analyze them separately. We further develop empirical tools to assess attribute shift and concept shift when generalizing to new offers, finding both to be modest in our setting. By synthesizing information across experiments, our framework transforms historical experimentation into a powerful resource for personalization.

Keywords: Personalization, Heterogeneous Treatment Effect, Field Experiments, Deep Learning, Representation Learning, Promotion

[*] Ta-Wei Huang is an Assistant Professor of Marketing at the National University of Singapore (email: thuang@nus.edu.sg). Eva Ascarza is a Professor of Business Administration at Harvard Business School (email: eascarza@hbs.edu). Ayelet Israeli is the Marvin Bower Associate Professor of Business Administration at Harvard Business School (email: aisraeli@hbs.edu). The authors thank the anonymous company for providing the data and Noah Ahmadi for assistance with data preparation. They are also grateful for valuable feedback from Bruce Hardie, Jeremy Yang, faculty and students in the HBS Marketing Unit, and members of the Customer Intelligence Lab. In addition, the authors thank participants at the 2024 American Causal Inference Conference, the 2024 Interactive Marketing Research Conference, CODE@MIT 2024, 2025 EMAC Spring Conference, as well as seminar participants at NUS, SMU, USC, UBC, and the RSM Rotterdam for their helpful comments. The authors received no financial support for the research, authorship, or publication of this article and declare no conflicts of interest. Generative AI (ChatGPT) was used solely to correct grammar and improve writing clarity; all ideas, analyses, and content are entirely the authors' own.

1 Introduction

Firms are increasingly adopting personalized interventions to strengthen customer relationships and improve marketing return on investment (Lemmens et al. 2025). In practice, this requires allocating marketing interventions to customers with the highest potential for incremental gains. A central challenge, however, is accurately capturing heterogeneity in customers' sensitivity to such interventions. For instance, in the context of retail promotions, some customers increase their spending meaningfully in response to modest discounts (e.g., 10% off), whereas others only change their behavior when offered deep price cuts (e.g., 40% off). Capturing this heterogeneity is essential for effective personalization — without it, firms risk eroding margins by offering discounts that fail to generate sufficient returns or by foregoing profit opportunities from customers who would have responded to smaller incentives.

At the core of this challenge lies the need to estimate how intervention effects vary across customers. A widely adopted framework for this purpose is *conditional average treatment effect* (CATE) modeling, , which quantifies the incremental impact of an intervention conditional on observed customer covariates. In practice, firms typically operationalize this approach through three main steps: (i) conduct a randomized controlled experiment for a specific intervention, (ii) train a CATE model to predict incremental effects based on pre-treatment covariates, and (iii) target new customers with the highest predicted CATE based on their covariates. By moving beyond average treatment effects, CATE models provide a systematic framework for identifying which customers are most likely to benefit from an intervention, thereby enabling firms to allocate marketing resources more efficiently. It has been successfully applied across domains such as promotional activities (Simester et al. 2020, Zhang and Misra 2025, Daljord et al. 2023), product experience (Ye et al. 2025), communications (Ellickson et al. 2023), and free trial design (Yoganarasimhan et al. 2023).

While current methods help firms decide which customers to target with a specific intervention, they provide limited guidance when managers face a broader set of decisions — such as identifying which customers would respond best to alternative and even untested interventions. One approach to address this challenge is to conduct experiments that include multiple intervention conditions and estimate separate CATE models for each (e.g., Zhao and Harinen 2019, Ellickson et al. 2023). However, such models cannot ensure generalization to interventions that were not explicitly tested. A more general solution would conduct a single experiment with a large number of intervention variants and use it to estimate the incremental impact as a joint function of intervention design features and customer covariates (e.g., Ellickson et al. 2024). Although appealing in theory, such experiments are rarely feasible in practice due to the implementation costs associated with simultaneously testing a large number of treatment variants and substantial sample size requirements. In addition, firms typically face business constraints such as limited campaign budgets, competing promotional priorities, and operational limits on how many interventions can be deployed or monitored at once, all of which make large-scale multifactorial experimentation difficult to execute in real-world settings.

In contrast, firms routinely conduct randomized controlled trials that assess a single intervention at a time. Over time, these efforts generate a rich archive of experiments encompassing diverse intervention designs and customer contexts. Despite their potential, such historical experiments are usually analyzed in isolation, leaving valuable information about heterogeneity in customer responsiveness across interventions underutilized. This gap underscores the need for a framework that systematically synthesizes past experiments to inform personalization decisions for both previously tested and novel interventions.

This research introduces a novel approach that leverages firms' existing experimental data to enable personalized interventions. Specifically, we propose a causal machine learning framework for estimating the *conditional incrementality function* (CIF) — treatment effects expressed as a function of both intervention features and customer covariates — using data from past experiments. By treating each intervention as a structured object characterized by its design features, CIF estimation enables personalization that jointly informs targeting decisions (whom to serve) and design decisions (what to offer). We show that by systematically pooling information across multiple experiments, our framework substantially improves estimation accuracy by leveraging existing experimental data. It also enables managers to generalize insights beyond previously tested interventions, design more effective targeting strategies for new offers, and avoid the need for simultaneous large-scale experimentation.

Although pooling information from past randomized experiments can yield richer insights than analyzing them individually, it also requires additional assumptions to guarantee valid joint estimation. From an identification standpoint, the standard assumptions that guarantee causal interpretability within a single experiment, *overlap*, *unconfoundedness*, and *no interference*, are insufficient for identifying the true CIF as if all interventions were tested simultaneously. We therefore introduce two additional assumptions necessary to identify the true CIF from sequentially conducted A/B tests: (i) unobserved intervention features are not systematically related to customer outcomes (*ignorability of unobserved intervention features*), and (ii) customer behavior remains consistent across experiments (*stability*). Although the stability assumption may appear strong, it embodies the same foundational premise underlying any predictive task: that behavioral patterns observed in past data generalize to future customers.

To model the CIF using data from past randomized controlled experiments, we propose a two-stage estimation framework. In the first stage, we compute a *doubly robust* (DR) score (Kennedy 2023) for each observation within each experiment. Under the identification assumptions outlined above, the DR score serves as an unbiased signal of the true treatment effect conditional on both intervention features and customer covariates. In the second stage, we train a predictive model to capture the relationship among intervention features, customer covariates, and the DR score. We provide formal theoretical guarantees showing that this two-stage framework yields an asymptotically correct estimator of the true CIF with high probability, provided that (i) firms experiment across the joint intervention–customer space with non-zero probability, and (ii) the second-stage model is sufficiently expressive but appropriately regularized to avoid overfitting. Under these conditions, the estimator converges to the true CIF as the number of experiments increases.

The first condition implies that as long as the firm conducts experiments involving diverse combinations of intervention design features and customer covariates, the two-stage estimation framework can recover the true CIF given a sufficient number of experiments. This requirement is relatively mild, as it does not necessitate orthogonal experimental designs. Our empirical results further show that even when past experiments are not orthogonally designed, jointly leveraging them yields substantial gains in targeting profitability compared with modeling them separately. In general, the importance of experimental design depends on how smoothly customer responses vary across offers and on the diversity of the firm's past experiments.

The second condition directly guides our choice of the second-stage model. The model must be flexible enough to capture heterogeneity arising from both intervention design features and customer covariates, yet constrained to prevent overfitting spurious patterns. To achieve this balance, we develop a specialized deep representation learning architecture to predict the DR score across all experiments. The architecture comprises two key components: (i) separate encoding networks for intervention features and customer covariates, which transform high-dimensional inputs into low-dimensional representations that capture sources of treatment effect heterogeneity; and (ii) a unified prediction network that integrates these representations to predict the DR score jointly across experiments. This design captures complex forms of heterogeneity while controlling model complexity, enabling the learned representations to generalize effectively even when the number of past experiments and sample sizes within each experiment are limited.

We assess the proposed framework in the context of promotional offers for consumer packaged goods (CPG). Leveraging data from 362 randomized controlled experiments encompassing over seven million customers on a leading North American customer engagement platform, we find that the CIF model delivers markedly stronger performance than conventional methods that analyze each experiment in isolation. Importantly, even when specific experiments are excluded from the model's training set, the CIF model continues to predict treatment effects in those held-out experiments more accurately than individual CATE models trained directly on the data from those excluded experiments. This finding highlights the model's ability to synthesize generalizable patterns across experiments and to extrapolate effectively to new, untested interventions.. Our empirical analysis also shows that the deep representation learning architecture embedded in the CIF model provides additional performance gains over standard deep learning approaches, underscoring the practical importance of architectural design in achieving the right balance between flexibility and complexity.

Furthermore, we empirically investigate two practical challenges that arise when generalizing treatment effect predictions beyond the conditions observed in historical experiments: (i) new interventions or customer profiles may differ substantially from those in prior experiments (*attribute shift*), and (ii) customer sensitivity may change over time in ways that violate the stability assumption (*concept shift*). Regarding attribute shift, we find that substantial differences in promotion features for untested offers can reduce the model's targeting effectiveness, whereas variations in customer covariates do not have a statistically significant impact on performance. This pattern suggests that, for the focal company, generalizing to new interventions requires particular attention to offer design features, as large deviations from previously tested promotions can hinder targeting performance. By contrast, we find no evidence of meaningful performance degradation when the distribution of customer covariates changes.

Regarding concept shift, we develop an empirical test that examines whether residual variation in treatment effects (i.e., variation unexplained by the CIF model) can be explained by the timing of each experiment. The results reveal no systematic relationship between treatment timing and residual effects, indicating that concept shift is unlikely to pose a major concern in our empirical setting. More broadly, these empirical assessments offer managers practical guidance on when and how insights from past experiments can be reliably leveraged to personalize future marketing interventions.

Our contributions are as follows. Methodologically, we integrate ideas from causal inference and machine learning to develop a framework that leverages past experiments to estimate treatment effects as a joint function of intervention features and customer covariates. We establish formal identification conditions under which past experiments allow recovery of the true CIF. By pooling information across experiments, the framework improves targeting efficiency for tested interventions relative to estimating treatment effects separately within each experiment. Moreover, it enables prediction for interventions not previously tested and outperforms models trained solely on the experimental data from those interventions. In addition, we introduce empirical strategies to evaluate two practical challenges when applying such models in real-world settings: attribute shift (differences in intervention features or customer profiles across contexts) and concept shift (changes in customer sensitivity over time).

Managerially, we demonstrate that firms' existing archives of randomized experiments, routinely collected but rarely exploited beyond single-offer evaluation, can be transformed into a scalable asset for personalization. By leveraging these experiments, managers can better align interventions with customers and improve marketing return on investment without resorting to costly, large-scale multi-treatment trials. Moreover, we empirically examine concerns that attribute shift and concept shift could limit the applicability of models trained on past experiments. Our analysis shows that both shifts are modest in our context and have no material impact on targeting effectiveness. Together, these findings demonstrate that synergizing past experiments is not only theoretically sound but also practically feasible and valuable for firms seeking to personalize marketing interventions.

The remainder of the paper is structured as follows. Section 2 reviews the related literature. Section 3 introduces the proposed two-stage estimation framework and discusses its identification requirements. Section 4 describes the empirical context and data. Section 5 outlines the implementation of our approach and alternative benchmarks. Section 6 presents results on targeting existing interventions and generalizing to new ones. Section 7 illustrates how firms can apply the framework to design personalized promotion offers. Section 8 examines the extent of attribute shift and concept shift in our setting. Finally, Section 9 concludes with key findings and directions for future research.

2 Related Literature

Our paper contributes to several streams of literature.

First, we contribute to the literature on heterogeneous treatment effect (HTE) estimation and targeted marketing interventions. In statistics, most existing work focuses on estimating HTEs within a single experiment involving a binary treatment condition (e.g., Wager and Athey 2018, Nie and Wager 2021, Semenova and Chernozhukov 2021, Kennedy 2023). In marketing, most studies likewise analyze one experiment at a time with a binary treatment condition (e.g., Ascarza 2018, Simester et al. 2020, Hitsch et al. 2023, Huang and Ascarza 2024) or a limited number of treatment variants, typically estimating separate models for each variant (Yoganarasimhan et al. 2023, Ye et al. 2025, Ellickson et al. 2023). More recently, researchers have begun to examine single experiments with high-dimensional treatments, such as representing multiple text-based interventions using embeddings derived from large language models and estimating treatment effects as a function of those embeddings (Ellickson et al. 2024, Imai and Nakamura 2024). Our work contributes to this literature by showing how firms can leverage *multiple single-variant randomized controlled experiments* to estimate treatment effects in high-dimensional treatment settings. Specifically, we establish formal identification conditions and develop a practical estimation framework that recovers treatment effects as a joint function of intervention design features and customer covariates using an archive of historical experiments, each consisting of a high-dimensional treatment condition and a control condition. In addition, we demonstrate that synergizing data from multiple experiments enhances targeting performance, even for untested treatment variants, compared with relying solely on data from each experiment in isolation.

Second, our work contributes to the literature on the generalizability of causal effect estimation, which investigates how treatment effects can be extrapolated to new populations or interventions. Prior research in statistics has primarily focused on transporting treatment effects to new populations with different covariate distributions (e.g., Crump et al. 2009, Petersen et al. 2012, Nethery et al. 2019, Khan et al. 2023, Zivich et al. 2024, Zhu et al. 2023). Within marketing, Rafieian (2023) proposes a matrix completion method across multiple experiments to estimate treatment effects for customer segments that are not jointly observed in both treatment and control conditions within any tested single intervention. Simester et al. (2020) empirically examine how machine learning–based targeting models perform when generalizing to new customers. Beyond differences in covariate distributions (referred to as *attribute shift*), they also highlight the potential issue of changes in customer responsiveness (referred to as *concept shift*). Our work contributes to this stream by offering both theoretical and empirical characterizations of attribute shift and concept shift in the context of HTE estimation for high-dimensional treatments. Theoretically, we extend classic domain adaptation theory from supervised learning (Ben-David et al. 2010) to formally characterize the generalization challenges inherent in HTE estimation. Practically, we develop an empirical framework to detect and quantify both attribute shift and concept shift using real-world experimental data.

Third, our research relates to the literature on multi-task and transfer learning (see Zhuang et al. 2020, Zhang and Yang 2021, for surveys). Most literature in computer science focuses on supervised prediction tasks (e.g., Kini and Manjunatha 2020, Bastani 2021, Xu et al. 2021). Recent work in marketing has shown increasing interest in transferring information across experiments to improve treatment effect estimation. For example, Timoshenko et al. (2020) propose a matrix factorization approach that leverages data from related campaigns to improve targeting performance in a focal campaign with limited experimental observations. Our research contributes to this emerging stream in two key ways. First, we demonstrate that synergizing past experiments not only improves sample size efficiency but also enables generalization to interventions that have not been directly tested. Second, we extend the theoretical foundations of multi-task learning, originally developed for supervised prediction (Maurer et al. 2016, Tripuraneni et al. 2020), to the setting of treatment effect estimation. Building on this extension, we further operationalize these ideas into an empirical framework that enables estimation of treatment effects across multiple experiments.

Fourth, our research broadly relates to the field of marketing customization. The standard literature emphasizes delivering relevant content or product recommendations by predicting click or purchase probabilities (e.g., Ansari and Mela 2003, Yoganarasimhan 2020, Gabel and Timoshenko 2022, Liberali and Ferecatu 2022). More recent studies have examined the treatment effects of customizable components within interventions (Wang et al. 2016, Ellickson et al. 2023, Daljord et al. 2023), but these typically focus on simplified intervention spaces and estimate average treatment effects across coarse customer segments using linear models. Our research develops a flexible machine learning framework that estimates heterogeneous treatment effects across high-dimensional intervention design features and customer covariates. This allows firms to design and target interventions based on incremental impact and enables more precise and granular personalization.

Finally, this paper contributes to the growing literature on representation learning in marketing. Prior research has primarily applied representation learning to compress or integrate high-dimensional product and customer data. For example, Dew et al. (2022) develop a multimodal framework that combines diverse data sources such as logos, Gabel et al. (2019) and Chen et al. (2022) adapt Word2Vec (Mikolov et al. 2013) to learn product embeddings from co-purchasing behavior, and Ma et al. (2025) combines multiple customer signals to create state representations that guide personalization decisions. We extend this stream of research by applying multi-task representation learning (Maurer et al. 2016, Tripuraneni et al. 2020) to the challenge of predicting treatment effects across many separate A/B experiments. We adapt the theoretical framework of Tripuraneni et al. (2020) to establish guarantees for our two-stage CIF estimation problem, where the prediction target is a noisy proxy (the DR score), and the objective is to recover a unified treatment effect function across many separate A/B experiments. Our analysis formally characterizes the importance of model flexibility and complexity in this setting, providing guidance for constructing treatment effect prediction models.

3 Model

3.1 Problem Setup

Consider a company that seeks to improve a business outcome by delivering a marketing intervention. For instance, the company might send promotional offers to customers to increase their spending on the promoted items. Each intervention is characterized by a vector of design features $\mathbf{Z} \in \mathcal{Z}$, such as discount levels or promoted product categories, and each customer is described by a vector of pre-treatment covariates $\mathbf{X} \in \mathcal{X}$, such as their demographics or prior purchase behaviors.

The company is interested in designing personalized interventions aimed at maximizing their incremental impact on key business outcomes, such as spending or profit from promoted items. A key step toward this objective is to estimate each customer's sensitivity to an intervention, conditional on both the intervention design features and the customer covariates . We define this response function as the *Conditional Incrementality Function* (CIF):

$$\tau(\mathbf{Z}, \mathbf{X}) \equiv \mathbb{E}[Y(\mathbf{Z}) - Y(\mathbf{0})|\mathbf{X}], \tag{1}$$

where $Y(\mathbf{Z})$ is the potential outcome of the business objective (e.g., spending on promoted items) for a customer should they receive an intervention with design features $\mathbf{Z}$, and $Y(\mathbf{0})$ is the potential outcome in the absence of the focal intervention. In other words, the CIF captures how much more a given customer is expected to respond to a specific intervention relative to not receiving it.

While the CIF is conceptually related to the conventional CATE framework, it differs in an important respect. Standard CATE approaches are typically designed to evaluate the effect of a single treatment relative to a control condition. In contrast, the CIF extends this framework by representing each intervention as a structured object characterized by its design features, $\mathbf{Z}$. The objective is to quantify the incremental impact of alternative offer designs, thereby informing both targeting (whom to serve) and design (what to offer) decisions within a unified modeling framework.

3.1.1 Archive of Historical Experiments

To estimate the CIF directly, a firm could, in principle, implement a single multi-treatment experiment that simultaneously tests a wide range of offers defined by different combinations of design features (analogous to a conjoint or factorial experimental design) together with a no-intervention control group. However, such an approach is rarely used in practice. First, such an experiment is often difficult because deploying thousands of intervention variants requires significant coordination across marketing, IT, and product teams. Second, data and budget constraints make such experiments costly and often underpowered — reliable estimates for many treatments would require far larger samples than most firms can afford. Third, managers tend to prefer simplicity: testing one intervention variant at a time is easier to manage, less costly, and allows faster rollout.

As a result, instead of attempting large-scale factorial experiments with thousands of intervention variants, firms typically run a sequence of smaller experiments, each focused on a single treatment. For example, a retailer might test whether offering a 20% discount increases spending by comparing a treatment group that receives the coupon to a control group that does not. Such tests are often limited to specific customer segments, such as recent purchasers or loyalty program members, to ensure relevance and limit costs.

The standard approach in the literature for leveraging such experimental data has been to estimate a separate CATE model for each individual experiment (e.g., Wager and Athey 2018, Ascarza 2018, Hitsch et al. 2023, Haushofer et al. 2025). While widely used, this strategy has important limitations. First, estimating CATEs independently for each experiment fails to exploit the underlying structure linking intervention design features to customer responses. Consequently, this approach cannot be used to make informed decisions about untested variants without running new experiments. Second, individual experiments often contain too few observations to support precise estimation of CATEs. This sparsity reduces statistical efficiency and limits the reliability of the resulting models. Third, when experiments are restricted to a narrow customer segment, the resulting CATE estimates have limited external validity. In such cases, predictions for customers outside the tested segment may be inaccurate or biased.

To overcome these limitations, we propose a modeling framework that integrates data from *multiple single-variant experiments* to estimate the CIF defined in Equation (1). This framework directly addresses the shortcomings of estimating separate CATE models for each experiment. First, by explicitly incorporating intervention design features, it allows firms to make targeting and design decisions for new, untested offers. Second, by pooling data across experiments, the framework substantially increases statistical power, yielding more precise and reliable treatment effect estimates, particularly in settings where individual experiments are too small to provide credible estimates on their own. Third, by combining information from multiple experiments that span different customer segments, the framework improves generalizability beyond the narrow customer pools of a single experiment.

3.1.2 Identification of CIF from Multiple Past Experiments

To formalize the idea of synergizing past experiments, consider a firm that has previously conducted K experiments, each designed to evaluate a distinct marketing intervention. In experiment $k \in \{1, 2, \cdots, K\}$, a unique intervention characterized by design features $\mathbf{Z}_k \in \mathcal{Z}$ was tested on a selected pool of customers. Each experiment included both a treatment group, which received the intervention, and a control group, which did not, enabling valid comparisons of outcomes between treated and untreated customers.

Let $W_{k,i} \in \{0, 1\}$ denote the treatment assignment for the i-th observation in experiment k, where $W_{k,i} = 1$ indicates assignment to the treatment condition and $W_{k,i} = 0$ indicates assignment to the control condition. For each observation in an experiment, the firm observes a vector of pre-treatment covariates $\mathbf{X}_{k,i}$, with support $\mathcal{X}_k \subseteq \mathcal{X}$, as well as the realized outcome of interest $Y_{k,i}$. We define the potential outcomes as $\{Y_{k,i}(W_{k,i} = 1), Y_{k,i}(W_{k,i} = 0)\}$, representing the outcomes for observation i in experiment k if they were to receive the intervention ($W_{k,i} = 1$) or not receive the intervention ($W_{k,i} = 0$), respectively.

We begin by outlining the assumptions under which the CIF can be pointwise identified using data from past single-variant experiments. The first three assumptions align with the standard conditions for identifying causal effects within a single experiment (Imbens and Rubin 2015).

Assumption 1 (Overlap) *For each experiment $k = 1, \cdots, K$, the assignment of treatment is subject to random variation, that is,* $0 < \mathbb{P}(W_{k,i} = 1|\mathbf{X}_{k,i}) < 1,\ \mathbf{X}_{k,i} \in \mathcal{X}_k$.

Overlap requires that each unit in the observed covariate space $\mathcal{X}_k$ has a positive probability of receiving either treatment or control, ensuring that treatment effects can be estimated from the observed data.

Assumption 2 (Unconfoundedness) *For each experiment $k \in 1, \cdots, K$, the treatment assignment is free from unobserved confounders, that is,* $\{Y_{k,i}(W_{k,i} = 1), Y_{k,i}(W_{k,i} = 0)\} \perp W_{k,i} \mid \mathbf{X}_{k,i}$.

Unconfoundedness ensures that treatment assignment is as good as random once we condition on the observed customer covariates. This assumption is satisfied by design in randomized experiments and can also be extended to observational causal inference settings, provided that treatment assignment depends solely on observed customer characteristics.

Assumption 3 (No Interference) *We assume that the potential outcomes for observation i in response to intervention k are independent of the treatment assignments of all other customers and all other interventions. Formally,* $\{Y_{k,i}(W_{k,i} = 1), Y_{k,i}(W_{k,i} = 0)\} \perp \{W_{k',i'}\}_{(k',i') \neq (k,i)}$.

No Interference rules out spillover effects, ensuring that a customer's outcome depends only on their own treatment and is not influenced by treatments assigned to other customers or the presence of other interventions.[1] Together, these conditions ensure that we can identify the causal effect of each specific intervention.

To ensure that pooling data across experiments and modeling them jointly yields a valid estimation of the CIF (evaluated at the design features and covariates observed in past experimental data), we introduce two additional assumptions.

[1] In our empirical setting, the likelihood of intervention interference is minimal for two reasons. First, customers did not receive multiple promotional interventions for the same items simultaneously, and the outcome variable captures only purchases of the items promoted by the focal intervention. This ensures that any observed sales lift can be attributed to a single intervention. Second, because the company implemented complete randomization, both treatment and control groups had an equal chance of receiving other unrelated interventions. As a result, any differences in outcomes between the two groups can be causally attributed to receiving the focal intervention. However, when experiments are not fully randomized, the risk of intervention interference increases.

Assumption 4 (Ignorability of Unobserved Design Features) *The unobserved aspects of the intervention — that is, features not captured by the design vector $\mathbf{Z}_k$ — are assumed not to systematically influence the conditional mean outcome given the customer covariates. Formally, we assume*

$$\mathbb{E}[Y_{k,i}(W_{k,i}=1) \mid \mathbf{Z}_k, \mathbf{X}_{k,i}] = \mathbb{E}[Y_{k,i}(\mathbf{Z}_k) \mid \mathbf{X}_{k,i}] \quad \textit{and} \quad \mathbb{E}[Y_{k,i}(W_{k,i}=0)|\mathbf{Z}_k, \mathbf{X}_{k,i}] = \mathbb{E}[Y_{k,i}(\mathbf{0}) \mid \mathbf{X}_{k,i}].$$

Ignorability of Unobserved Design Features requires that any unobserved design features not captured in $\mathbf{Z}_k$ do not systematically influence customer responses once we condition on the observed design features and customer covariates. For example, if $\mathbf{Z}_k$ excludes visual imagery elements in a promotional offer, the assumption implies that the quality of such imagery is not systematically correlated with other observed design attributes (e.g., product category or discount level). Otherwise, for product categories that tend to feature higher-quality images, the expected treated outcome ($\mathbb{E}[Y_{k,i}(W_{k,i}=1) \mid \mathbf{Z}_k, \mathbf{X}_{k,i}]$) would exceed the counterfactual expectation under the observed design features alone ($\mathbb{E}[Y_{k,i}(\mathbf{Z}_k) \mid \mathbf{X}_{k,i}]$). Note that if all relevant design attributes are observed, this assumption holds automatically.[2]

Assumption 5 (Stability) *Customer behavior is assumed to be stable in the sense that, for any experiment $k \in 1, \ldots, K$, the conditional expectation of the potential outcome consistently represents the true outcome of interest, that is,*

$$\mathbb{E}[Y_{k,i}(\mathbf{Z}_k) \mid \mathbf{X}_{k,i} = \mathbf{X}] = \mathbb{E}[Y(\mathbf{Z})|\mathbf{X}] \quad \textit{and} \quad \mathbb{E}[Y_{k,i}(\mathbf{0}) \mid \mathbf{X}_{k,i} = \mathbf{X}] = \mathbb{E}[Y(\mathbf{0})|\mathbf{X}].$$

Stability requires that the relationship between customer covariates, intervention design, and outcomes remains consistent across experiments. In other words, the potential outcomes observed in each experiment should reflect draws from the same underlying behavioral process that the firm aims to model. Violations of this assumption, due to shifts in customer preferences or systematic differences in populations across experiments, would require additional modeling assumptions to account for behavioral changes over time.

Together, these five assumptions ensure that: (i) treatment effect estimates from past experiments can be given a causal interpretation (Assumptions 1 to 3), and (ii) synergizing data

[2] This is a relaxation of Assumption 2 in Ellickson et al. (2024), which requires that text embeddings fully capture all relevant features of text content.

across experiments and modeling them jointly enables pointwise identification of the CIF (Assumptions 4 to 5). The following theorem establishes the nonparametric identification of the CIF using data from past experiments.

Theorem 1 (CIF Identification) *Under Assumption 1 to Assumption 5, for any design feature* $\mathbf{Z} \in \mathcal{Z}$ *that was tested in at least one experiment (i.e., there exists* $k \in \{1, \cdots, K\}$ *such that* $\mathbf{Z}_k = \mathbf{Z}$*) and any individual with covariates* $\mathbf{X} \in \mathcal{X}$*, the CIF defined in Equation* (1) *is point-wise identified such that* $\tau(\mathbf{Z}_k, \mathbf{X}_{k,i}) = \mathbb{E}[Y_{k,i}|W_{k,i} = 1, \mathbf{X}_{k,i}] - \mathbb{E}[Y_{k,i}|W_{k,i} = 0, \mathbf{X}_{k,i}]$.

The proof of Theorem 1 is provided in Web Appendix A.1. This theorem implies that, instead of running a single large experiment that tests all interventions simultaneously, a company can combine multiple independently conducted experiments to learn how different types of customers respond to various intervention designs, without relying on parametric assumptions about the underlying relationships.

3.2 Two-stage Framework for Incrementality Prediction

Next, we propose a two-stage framework for estimating the CIF by synergizing past experiments, building on the transformed outcome regression approach developed for CATE estimation (e.g., Semenova and Chernozhukov 2021, Nie and Wager 2021, Kennedy 2023). In the first stage, we construct an unbiased proxy for the conditional average treatment effect within each experiment using the doubly robust (DR) score (Kennedy 2023). In the second stage, we train a machine learning model to predict this score using intervention design features and customer covariates. This two-stage approach enables flexible, data-efficient estimation of the CIF across a broad set of intervention and customer combinations that fall within the range covered by past experiments, even when individual experiments are limited in scope.

3.2.1 Stage 1: Doubly Robust Score Construction

Theorem 1 shows that when Assumptions 1 through 5 are satisfied, the standard CATE function within a given experiment is equivalent to the CIF evaluated at the design features of the intervention tested in that experiment. This equivalence justifies using the DR score computed within each single experiment as a valid prediction target for estimating a CIF model *across* multiple experiments. The following corollary formalizes this result, showing that the DR score from each experiment provides an unbiased signal for the true CIF.

Corollary 1 (Unbiased Score) *Let $\mu_{k,w}(\mathbf{x}) = \mathbb{E}[Y_{k,i}|W_{k,i} = w, \mathbf{X}_{k,i} = \mathbf{x}]$ be the conditional expected outcome for intervention k, and let $\pi_k(\mathbf{x}) = \mathbb{P}[W_{k,i} = 1|\mathbf{X}_{k,i} = \mathbf{x}]$ denote the propensity score of being treated in experiment k. Then, under Assumption 1 to Assumption 5, the DR score is an unbiased signal for the CIF defined in* (1)*:*

$$\tau(\mathbf{Z}_k, \mathbf{X}_{k,i}) = \mathbb{E}\left[\mu_{k,1}(\mathbf{X}_{k,i}) - \mu_{k,0}(\mathbf{X}_{k,i}) + \tfrac{W_{k,i}-\pi_k(\mathbf{X}_{k,i})}{\pi_k(\mathbf{X}_{k,i})[1-\pi_k(\mathbf{X}_{k,i})]}\left[Y_{k,i} - \mu_{k,W_{k,i}}(\mathbf{X}_{k,i})\right]\middle|\, \mathbf{Z}_k, \mathbf{X}_{k,i}\right].$$

The proof is provided in Web Appendix A.2. Corollary 1 shows that the conditional expectation of the DR score, when constructed separately for each tested intervention, is equal to the CIF evaluated at the corresponding intervention design features and customer covariates. This result suggests a two-stage estimation strategy: first, compute the empirical DR score *within* each experiment, and then fit a model that estimates the CIF by learning the relationship between the DR scores, intervention design features, and customer covariates *across* experiments.

To empirically construct the DR score, we apply the standard cross-fitting procedure (Semenova and Chernozhukov 2021, Nie and Wager 2021). Specifically, for each experiment k, we build a machine learning model $\widehat{\mu}_{k,w}^{[-i]}(\mathbf{X}_{k,i})$ that predicts the expected outcome $Y_{k,i}$ for each treatment condition $W_{k,i} = w$, conditional on the covariates $\mathbf{X}_{k,i}$.[3] The superscript $[-i]$ indicates that the model is trained on data excluding observation i. For the propensity score, we leverage the fact that our empirical setting involves fully randomized experiments. In such cases, the true propensity score is known and corresponds to the empirical proportion of observations assigned to each treatment arm. Importantly, when the true propensity score is known, either through randomization, as in our setting, or via a known treatment rule, the DR score remains

[3] Nie and Wager (2021) recommend selecting the conditional outcome model that minimizes prediction error, as this indicates that the model captures baseline outcome variation more effectively.

unbiased even if the outcome model is misspecified (Funk et al. 2011) (see Web Appendix A.2 for a proof).

In observational studies, where treatment assignment depends on covariates and the true propensity score is unknown, the same cross-fitting procedure can be applied to estimate the propensity score using a machine learning model. In such a setting, it is critical that both the outcome and propensity score models are estimated with sufficiently high quality. Specifically, their mean squared errors must converge to zero at appropriate rates to ensure consistent estimation of treatment effects (Chernozhukov et al. 2018, Semenova and Chernozhukov 2021).

3.2.2 Stage 2: CIF estimation

After constructing the DR scores for each experiment, we pool them across experiments to create a unified dataset for estimating the CIF. While any regression-type models can be used as the second-stage model, there are two important properties an ideal second-stage model should have. First, the model should be sufficiently flexible to capture treatment effect heterogeneity across high-dimensional design features and customer covariates. In this regard, flexible machine learning approaches such as deep neural networks offer a plausible solution, as they can learn complex nonlinear relationships without requiring strong parametric assumptions.

Second, the model should maintain constrained complexity to prevent overfitting and ensure stable generalization. In practice, real-world experiments often have small sample sizes, providing limited information to estimate heterogeneous effects within specific designs or customer segments. Firms also tend to test only a narrow subset of potential interventions, resulting in sparse coverage of the broader design space. Moreover, marketing outcomes typically exhibit high noise and weak signals (Huang and Ascarza 2024), which increases the risk that highly flexible models (such as deep neural networks) capture spurious rather than meaningful patterns. These challenges make it essential to control model complexity so that estimated relationships can generalize beyond the observed experiments.

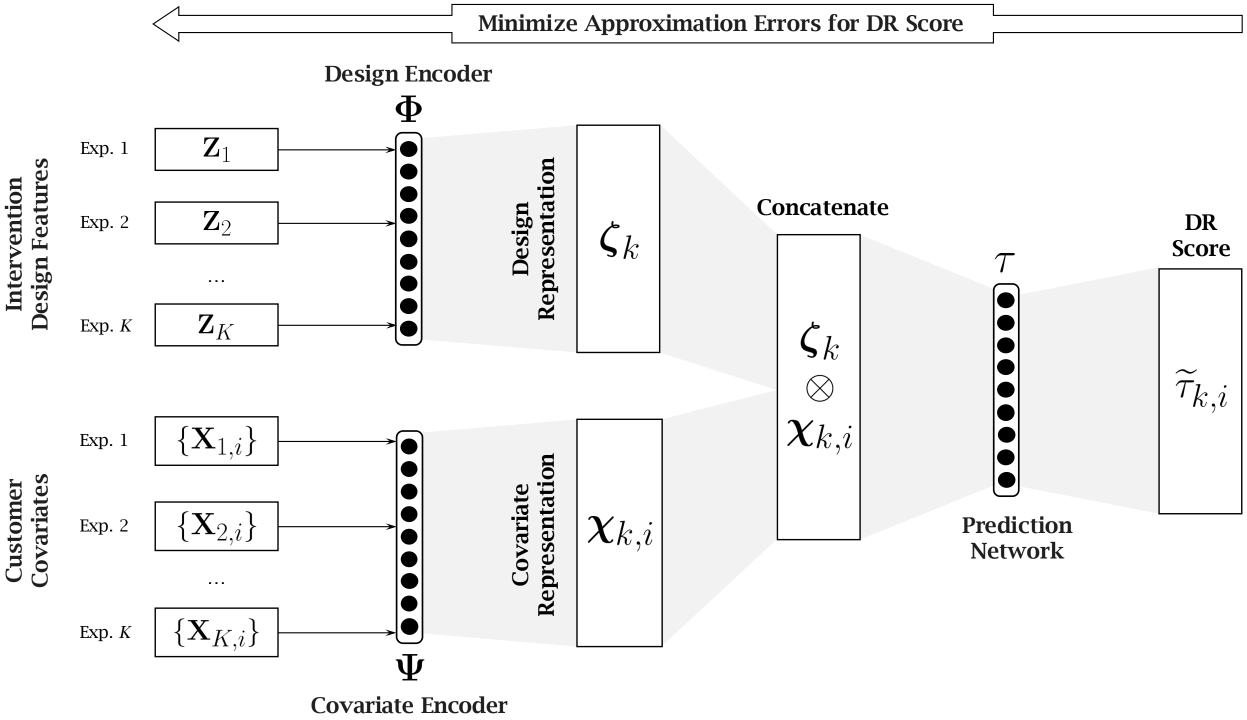
Together, these two considerations are important when selecting the second-stage model for CIF estimation: the model should be expressive enough to capture rich heterogeneity while sufficiently constrained to prevent overfitting. Guided by these principles, we develop a customized deep learning architecture that balances flexibility with constrained complexity.

Model Architecture. To balance model expressiveness with the need to control model complexity, we implement a *deep representation learning* (DRL) architecture as the second-stage model. This approach jointly learns low-dimensional representations of intervention design features and customer covariates while predicting DR scores across all tested interventions. The architecture is designed to accomplish two objectives. First, it extracts compact and dense representations from high-dimensional inputs that effectively capture heterogeneous treatment effects. Second, it learns a shared predictive function that maps these representations to DR scores across experiments.

Figure 1 presents the proposed model architecture. The model first maps high-dimensional design features $\mathbf{Z}_k$ and customer covariates $\mathbf{X}_{k,i}$ into low-dimensional representations: $\boldsymbol{\zeta}_k$ for intervention design and $\boldsymbol{\chi}_{k,i}$ for customer covariates through two separate deep neural networks: the design encoder $\boldsymbol{\Phi}$ and the covariate encoder $\boldsymbol{\Psi}$. The resulting representations are concatenated and fed into a prediction network, which predict the DR score (denoted as $\tilde{\tau}_{k,i}$) for each observation in experiment k.

The proposed representation learning architecture offers two practical advantages for managing model complexity. First, it performs effective dimensionality reduction by compressing high-dimensional and potentially irrelevant inputs into compact representations that retain the information most predictive of treatment effects. Second, by learning design and covariate representations separately, the model introduces an inherent form of regularization, since the representation learning stage prevents direct interactions between design features and customer covariates. This separate structure helps disentangle the respective contributions of design features and customer covariates to treatment effects and mitigates overfitting to spurious feature-covariate interactions.

Figure 1: Proposed Model Architecture for Incrementality Prediction



In addition, learning separate representations facilitates downstream analyses that deepen understanding of treatment effect heterogeneity. For example, clustering the design representations reveals groups of interventions that produce similar treatment effect patterns, offering actionable guidance for identifying which types of interventions the model views as functionally equivalent. Similarly, clustering the customer representations uncovers segments of customers with comparable treatment responsiveness, helping managers identify which covariates are most useful for segmentation when the objective is to target customers based on expected incremental impact.

Once the representations ζ_k and $\chi_{k,i}$ are extracted, we construct a prediction network that takes their concatenation as input to predict DR scores across all experiments. Importantly, rather than adopting the common multi-task learning approach of building separate prediction networks for each intervention (Maurer et al. 2016, Tripuraneni et al. 2020), we employ a single prediction network with parameters shared across experiments. This choice is motivated by two considerations. First, intervention-specific networks rely on the data from individual experiments to estimate parameters, which prevents them from predicting treatment effects to new, untested designs. Second, individual experiments typically have limited sample sizes, making the estimation of neural network models unstable. By pooling information across experiments and sharing parameters, the model produces more accurate and stable predictions.

Model Calibration. We jointly calibrate the three neural networks ($\mathbf{\Phi}$, $\mathbf{\Psi}$, and τ) simultaneously so that $\mathbf{\Phi}$ and $\mathbf{\Psi}$ learn representations most informative for predicting treatment effects across interventions. To do so, we pool data from all experiments and estimate $\mathbf{\Phi}$, $\mathbf{\Psi}$, and τ by minimizing the squared error between the model's predictions and the DR scores:

$$\arg\min_{\tau,\mathbf{\Phi},\mathbf{\Psi}} \sum_{k=1}^{K} \frac{\gamma_k}{N_k} \sum_{i=1}^{N_k} \left[\tau(\mathbf{\Phi}(\mathbf{Z}_k), \mathbf{\Psi}(\mathbf{X}_{k,i})) - \tilde{\tau}_{k,i_k}\right]^2, \tag{2}$$

where N_k denotes the number of observations in experiment k, and γ_k is the relative importance weight assigned to each intervention. For simplicity, we set $\gamma_k = 1$ for all interventions in our empirical application. In practice, however, companies may choose to assign different weights to align with their specific objectives. For example, if they wish to prioritize accuracy for certain types of interventions, they can assign higher weights to those interventions. Alternatively, the weights can be treated as tuning parameters to optimize an objective function.

3.3 Theoretical Guarantees

In Web Appendix B, we provide theoretical guarantees for the proposed CIF estimation framework. Specifically, we show that if the firm's experimentation procedure ensures that all regions of the joint design–covariate space are sampled with nonzero probability, and the model is sufficiently expressive to approximate the true CIF, then the mean squared error (MSE) of the two-stage CIF estimator converges to zero with high probability as the number of experiments (and thus the total sample size) increases.

To establish this result, we extend the theoretical framework of Tripuraneni et al. (2020), which analyzes multitask learning in standard supervised learning settings, to our two-stage CIF estimation framework. The resulting upper bound on mean squared error depends on two main factors: (i) the number of tested interventions and the sample size within each experiment, and (ii) the Gaussian complexity (Bartlett and Mendelson 2002) of the second-stage model, which captures its capacity to overfit random noise. The latter term can be further decomposed into the complexity of the representation learning modules that encode intervention designs and customer covariates, and the complexity of the prediction network that maps these representations into CIF estimates.

We then derive an upper bound on the MSE of the second-stage estimator when using deep neural networks (DNNs) by applying the general complexity bound for DNNs established in Golowich et al. (2018). The result suggests that the MSE converges to zero with high probability when the number of experiments approaches infinity. Besides, it demonstrates how our architectural choices — using separate encoding networks for interventions and customers, and sharing parameters across experiments during prediction — serve as regularization mechanisms that constrain model complexity. Consequently, the proposed architecture is less prone to overfitting than standard DNNs.

Note that the theoretical guarantees are not tied to the specific deep learning architecture we propose. Any model class that is sufficiently expressive yet adequately controlled in complexity can, in principle, recover the true CIF given enough experimental data. While our empirical analyses show that the proposed architecture performs particularly well in the current setting, alternative models may prove superior in different empirical contexts. For this reason, we recommend that firms evaluate a range of candidate model classes and select the one that delivers the strongest holdout performance prior to implementation.

3.4 Generalization to New Data

Beyond predictive accuracy on past experiments, the ability of any CIF model trained on historical data to generalize to new contexts depends on two critical factors (Simester et al. 2020). First, the degree of distributional differences in design features or customer covariates between past and new data, commonly referred to as *attribute shift*. Second, changes in the underlying CIF function across contexts, which represent violations of Assumption 5 and are commonly termed *concept shift*. In Web Appendix C, we formalize how these factors affect generalization by applying the domain adaptation framework of Ben-David et al. (2010) to our setting.

To assess whether these concerns present significant practical challenges, we develop an empirical approach to evaluate the three factors outlined above using the experimental data from our application. First, in Section 8.1, we evaluate the impact of attribute shift by holding out entire experiments and examining how model performance varies with the similarity between the training and holdout sets. Specifically, we measure the distance between the intervention design features and customer covariates of the held-out experiments relative to those in the training data. For design features, we observe a clear inverse relationship: as held-out interventions become more dissimilar from those previously tested, targeting profitability declines. This pattern indicates that attribute shift in intervention design features can be a boundary condition in our empirical setting. In contrast, we find no systematic relationship between covariate distance and targeting profitability once the CIF model is jointly estimated across experiments, suggesting that shifts in customer covariate distributions are less problematic in this context.

Second, in Section 8.2, we provide evidence that concept shift is unlikely to be a significant concern in our empirical context. Concept shift arises when the underlying relationship among design features, customer covariates, and treatment effects evolves over time. To evaluate this possibility, we examine whether the residual variation in the DR score, that is, the portion not explained by intervention design features or customer covariates, systematically varies with the timing of intervention launches. We find no such pattern: intervention timing does not account for any residual variation in the DR score. This result indicates that concept shift is not a key issue in our setting and supports the validity of using past experimental data to inform future targeting decisions.

In conclusion, although both attribute shift and concept shift are potential challenges when leveraging past experiments, our empirical evidence suggests that they are not material concerns in this application. Importantly, we provide a practical approach that firms can apply to diagnose whether attribute or concept shift may limit the usefulness of past experiments in their own contexts. By systematically evaluating these risks, managers can determine whether past experimental data can be reliably leveraged for personalizing future interventions.

4 Empirical Context and Data

4.1 Empirical Context

We evaluate our method using data from a North America–based consumer engagement platform. On this platform, customers earn reward points by uploading shopping receipts, which they can redeem for gift cards or products. The platform generates revenue by providing customer relationship management services to consumer packaged goods (CPG) companies and retailers (collectively referred to as partners), who pay commissions in exchange for promotional exposure. A central feature of the platform is the deployment of promotional offers on behalf of these partners (e.g., "Buy two bottles of wine from Brand W and receive 2,000 points"), which are designed to incentivize purchases. These offers are prominently displayed on the platform's discovery page, and customers automatically receive the specified rewards when they submit receipts that meet the offer conditions.

While the platform benefits from increased customer engagement and overall receipt uploads, the success of its operations is evaluated at the offer level, as partners pay to run specific promotional campaigns. Therefore, the platform's primary objective in promotions is to help each partner identify which customers should receive each offer to drive incremental spending and, in turn, profits.

4.2 Data Description

To quantify the incremental impact of promotional offers, the platform routinely conducts randomized controlled experiments for each offer. Our analysis draws on 362 such experiments conducted between October 2022 and April 2023, each testing a distinct promotion on samples ranging from 10,000 to 200,000 customers. In total, the dataset comprises 7,591,203 unique customers and 30,701,442 offer–customer observations across these 362 experiments.

4.2.1 Experiments

For each experiment, the platform tested a specific promotional offer on a predefined pool of eligible customers. This pool was typically determined based on partner-specific requirements, such as including all customers, prior purchasers of the brand, or customers with certain demographic characteristics. Customers in the eligible pool were then randomly assigned to treatment (90%) and control (10%) groups. Randomization was conducted independently of login activity and included all eligible customers, regardless of whether they accessed the platform during the campaign period. Only the treatment group received the focal promotion; the control group received no alternative offer.

Aside from the focal offer being tested, all customers continued to receive baseline promotional offers that were unrelated to the study. Because randomization was orthogonal to these baseline offers, exposure to unrelated promotions was balanced across treatment and control groups. Therefore, any differences in spending on the promoted items can be causally attributed to the focal offer.[4] Detailed randomization checks are provided in Web Appendix D.1.

The platform's primary objective is to increase total dollar spending on promoted items during the effective period of the offers. Customer spending was measured based on uploaded receipts. If a customer did not upload a receipt, the platform recorded no purchase, and their spending was coded as zero.[5] Figure 2 shows the ATE of each offer along with its 95% confidence interval. Using a two-tailed t-test at the 5% significance level, we find that only 33% of offers produce a statistically significant increase in spending on promoted items[6]. This sug-

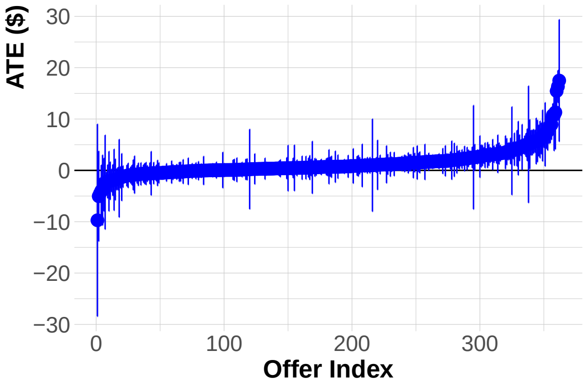
[4] Although a small subset of customers were eligible for multiple experiments and could receive more than one treatment offer simultaneously, we believe that any potential interference is minimal and find no evidence of significant interference effects. First, on average, only 7% of customers received multiple experimental offers at the same time, and none of these offers promoted the same items. Second, customers in both the treatment and control groups of the focal offer were equally likely to receive other experimental offers, minimizing the risk of systematic bias. Finally, in Web Appendix D.2, we show that the number of concurrent experimental offers — regardless of whether they belonged to the same product category — does not significantly affect the treatment effect of the focal offer on spending for the promoted items.

[5] For customers in the treatment group, the observed increase in spending may partly reflect a higher likelihood of receipt submission rather than a pure increase in consumption. Since the platform records purchases only when receipts are uploaded, we cannot fully disentangle whether the effect is driven by greater purchase activity or by a higher propensity to submit receipts. This caveat should be kept in mind when interpreting the treatment effect estimates.

[6] We also note that while most offers did not produce statistically significant changes in spending, only three offers (0.8%) produced statistically significant reductions in spending on promoted items.

gests that the pool of eligible customers was likely not systematically selected based on prior expectations of offer effectiveness.

Figure 2: Average Treatment Effects Across 362 Experiments



Note. Each point reports the average treatment effect of a promotional offer together with its 95% confidence interval.

4.2.2 Offer Features

Each promotional offer is defined by several key design features: (i) the promoted item(s) and their product categories, (ii) the point-earning criteria, such as minimum purchase amounts or quantities, (iii) the number of points awarded upon meeting these criteria, and (iv) additional details, including redemption limits or specific promotional tactics. In total, our dataset includes fourteen design features (six numerical and eight categorical) which expand to 46 variables after dummy encoding. Table 1 provides a summary of these features.

4.2.3 Customer Covariates

For each observation in an experiment, we construct 40 pre-treatment covariates using only data recorded before the experiment began. This ensures that these covariates describe the customer's characteristics and behaviors prior to treatment assignment, and are not influenced by any outcomes that occurred during or after the experiment. Table 2 presents summary statistics for selected customer covariates. The full list of covariates is provided in Web Appendix D.1. The covariates are organized into three groups:

Table 1: Summary of Offer Design Features

DISCRETE DESIGN FEATURES

Variable	Unique Value	Top Counts
Is the offer a multi-transaction offer?	2	False: 211, True: 151
Is the offer stackable?	2	False: 41, True: 321
Criteria: Purchase from selected items	2	False: 311, True: 51
Criteria: Bulk purchase	2	False: 238, True: 124
Criteria: Available for club members only	2	False: 343, True: 19
Criteria: Actions for reward claim	2	Quantity: 314, Spending: 48
Promoted product category	19	Personal Care: 70, Baby: 64, Beer: 60, Wine: 46
Promotion tactic	14	Dollar_or_unit: 170, Frequency: 71

NUMERICAL DESIGN FEATURES

Variable	# Missing	Mean	S.D.	Min	P25	Median	P75	Max
Points awarded	0	2,540	2,345	350	1000	2,000	3,000	20,000
Minimum required spending ($)	313	17.6	11.1	5.0	8.0	15.0	30.0	50.0
Minimum required quantity	49	1.38	0.62	1	1	1	2	4
Times the offer can be redeemed	0	4.19	7.02	1	1	1	3	25
Maximum redemption times per transaction	0	1.50	1.22	1	1	1	1	5
Duration of the promotion (in days)	0	46.0	30.6	5	30	31	63	176

Note. Consider the offer "Spend $30 on selected sizes of diapers and earn 3,000 points" as an example. It includes the criterion *purchase from selected items* and specifies a *spending* threshold to qualify for the reward. Besides, it is classified as a single-transaction, non-stackable offer. The variable "Is the offer stackable?" indicates whether qualifying purchases can count toward other offers simultaneously. The "Promotion tactic" variable captures the strategic intent of the offer. For example, "Dollar_or_unit" means that the offer is designed to increase total spending or unit sales of specific items, while "Frequency" suggests that the offer encourages more frequent purchases of those items.

Demographic and Account Attributes: Demographic and account-level information, including gender, age, platform tenure (measured at the time the offer was received), and whether the customer joined via a referral.

Non-contextual Behaviors: Overall shopping behavior during the 60 days before the experiment. This includes:

- *Recency:* Whether a receipt was uploaded in the last 7, 30, or 60 days.
- *Frequency:* Total number of receipts uploaded.
- *Monetary value:* Average order value and average item count per receipt.
- *Category and merchant preferences:* Share of receipts that include items from a specific product category or from a particular merchant.

Contextual Behaviors: Purchasing behavior related to the items promoted in the offer, measured over the 60 days before the experiment. Metrics include how recently the customer purchased these items, how often they purchased them, and how much they spent on them.

Table 2: Summary of Selected Covariates

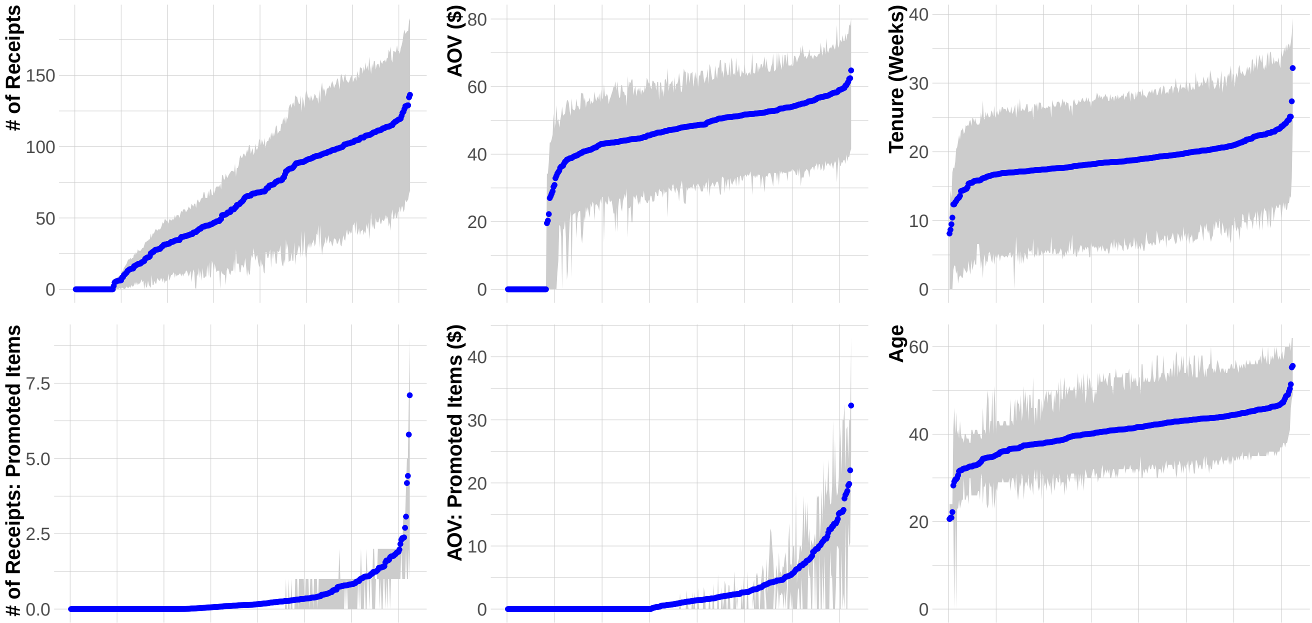
Variable	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Customer Attributes						
Is from referral?	0.65	0.48	0	0	1	1
Gender = male	0.14	0.48	0	0	0	1
Gender = female	0.78	0.48	0	1	1	1
Age	40	16	0	30	51	101
Tenure of membership (weeks)	19	15	6.1	6.5	28	646
NON-CONTEXTURAL BEHAVIORS (ALL RECEIPTS)						
# of receipts	58	69	0	3	86	2,903
Average order value (AOV)	54	90	0	18	59	5,145
Items per order	7	12	0	3	9	622
Uploaded a receipt in the past 7 days	0.63	0.48	0	0	1	1
Uploaded a receipt in the past 30 days	0.76	0.43	0	1	1	1
Uploaded a receipt in the past 60 days	0.78	0.42	0	1	1	1
Beverages	0.03	0.04	0.00	0.00	0.05	1.00
Candy	0.01	0.02	0.00	0.00	0.01	1.00
Dairy	0.02	0.03	0.00	0.00	0.03	1.00
CONTEXTURAL BEHAVIORS (RECEIPTS RELATED TO PROMOTED ITEMS)						
# of receipts	0.34	1.3	0	0	0	406
Average order value (AOV)	2.7	7.7	0	0	0	220
Items per order	0.24	0.55	0	0	0	12
Uploaded a receipt in the past 7 days	0.085	0.28	0	0	0	1
Uploaded a receipt in the past 30 days	0.14	0.34	0	0	0	1
Uploaded a receipt in the past 60 days	0.17	0.38	0	0	0	1

4.2.4 Eligible Pools

In each experiment, the platform tested a promotional offer on a specific pool of eligible customers. This design introduces heterogeneity in covariate distributions across experiments. Figure 3 illustrates the distributional differences of selected covariates across experiments. Each blue point represents the mean value for a given experiment, while the gray area shows the interquartile range (from the first to the third quartile) of each covariate. For example, in the top-right panel, each point indicates the average number of uploaded receipts for all customers within a specific experiment.

The results suggest that each individual experiment focuses on a relatively narrow customer pool. Therefore, combining data from multiple experiments provides broader coverage of the covariate space and may enhance the model's external validity. In Section 8, we show that

Figure 3: Distribution of Selected Covariates Across Experiments



Note: Each point represents the mean value for a given experiment, while the gray area indicates the interquartile range (from the first to the third quartile) of each covariate. Experiments are sorted by the average value of each selected covariate.

when the proposed model is trained on 181 randomly selected experiments, it can accurately predict treatment effects for the other 181 holdout experiments whose customer covariate distributions differ from those in the training set. This finding suggests that it is not necessary for each experiment to span all customer types; aggregating multiple targeted experiments can collectively achieve sufficient coverage of the covariate space.

5 Model Implementation

In this section, we describe the implementation of the proposed CIF model, outline benchmarks for comparison, and introduce the evaluation metrics.

5.1 Implementation of the Proposed Model

We implement the proposed two-stage CIF estimation framework as follows.

5.1.1 Stage 1: Doubly Robust Score Construction

For each experiment, we construct the DR score defined in Equation (1) using a ten-fold cross-fitting procedure as described in Chernozhukov et al. (2018). To estimate the conditional outcome models, we consider three candidate model classes (elastic net, random forest, and XGBoost) and select XGBoost given its best predictive accuracy (see Section Web Appendix E.1 for model selection details). For the propensity scores, we use the empirical treatment proportion within each experiment, as treatment assignment was completely randomized. The predicted outcomes and estimated propensity scores are then substituted into Equation (1) to obtain the DR score for each offer.

5.1.2 Stage 2: Model Specification and Calibration

To implement the proposed DRL architecture as the second-stage model, we construct a five-layer fully connected neural network for both the design encoder and the covariate encoder. The input layers accept 46 dimensions for design features and 40 dimensions for customer covariates, respectively. Each network includes three hidden layers with 10 units per layer. The output layers of the encoding networks produce 10-dimensional representations for both design features and customer covariates. We employ the Rectified Linear Unit (ReLU) activation function, $\sigma(x) = \max(0, x)$, for all hidden units. For the prediction network, we employ a five-layer fully connected neural network. The input to this network is the concatenation of the 10-dimensional representations from the design and customer encoders, yielding an input layer of 20 dimensions. Each of the three hidden layers contains 10 units. The output layer generates the CATE prediction. We apply the ReLU activation function to all hidden units and use a linear activation function in the output layer to produce the final prediction.[7]

We calibrate the proposed DRL model by optimizing the loss function in Equation (2) using the `Adadelta` algorithm (Zeiler 2012). This adaptive learning rate method dynamically adjusts the step size based on historical gradient information, thereby eliminating the need to manually tune a global learning rate and improving training stability. To further enhance computational efficiency and stabilize convergence, we employ mini-batch training with 30 observations per batch and train the network for 10 epochs.

[7] In Web Appendix F, we also test shallower and deeper network architectures as well as alternative representation dimensionalities. Overall, we find that the results are qualitatively consistent across these variations.

5.2 Methods for Comparison

We evaluate the performance of the proposed CIF model against two sets of benchmarks.

Individual CATE Models. The first benchmark compares our approach with the conventional method of estimating separate models for each experiment in isolation. Specifically, we estimate an independent deep neural network (DNN) for each experiment to predict the DR score. Because design features do not vary within a given offer, they are excluded from these models. This setup corresponds to a standard DR-learner for CATE estimation (Kennedy 2023) implemented with a standard fully-connected DNN.

Alternative Deep Learning Architectures for CIF Modeling. The second benchmark evaluates whether the proposed DRL architecture enhances performance relative to alternative neural network architectures used as the second-stage model within the CIF framework. Specifically, we compare its performance against (i) a standard DNN without the proposed representation structure and (ii) a standard DNN that excludes design features altogether. The first comparison isolates the performance gains attributable to the representation learning structure, whereas the second assesses the benefit of synergizing past experiments — even without incorporating design feature information — relative to modeling each experiment independently. All models are constructed with the same network depth and a comparable number of parameters as the DRL architecture.

5.3 Performance Metrics

We evaluate model performance using two key targeting metrics. First, we assess the model's ability to prioritize treatment allocation among customers with high treatment effects, that is, how effectively it ranks customers by their predicted responsiveness. Second, we evaluate the resulting profitability when the model is used to determine which customers should be targeted.

Treatment Prioritization. We assess treatment prioritization by evaluating each model's ability to distinguish observations with high treatment effects from those with low treatment effects. Specifically, we employ the *Area Under the Targeting Operating Characteristic Curve* (AUTOC; Yadlowsky et al. 2025), a standard metric for assessing how well a model ranks individuals by their treatment effects. AUTOC quantifies the incremental impact a model achieves by prioritizing the most responsive individuals at different proportions of the treated population, relative to random targeting at the same treatment rate. It then integrates performance across all possible targeting thresholds. A higher AUTOC value indicates stronger ranking performance.

Profitability. We evaluate profitability by comparing how different treatment models guide the platform's decision on which customers should receive an offer. Specifically, we consider the targeting policy $\pi_{\widehat{\tau}}$, under which the platform targets only those customers whose predicted incremental profitability for their partners, $m \cdot \widehat{\tau}(\mathbf{Z}_k, \mathbf{X}_{k,i})$ (where m denotes the partner's gross margin) exceeds the cost c_k of offer k.[8] $\pi_{\widehat{\tau}}(\mathbf{Z}_k, \mathbf{X}_{k,i}) = \mathbb{1}\big[m \cdot \widehat{\tau}(\mathbf{Z}_k, \mathbf{X}_{k,i}) > c_k)\big]$.[9]

We then estimate the expected profit on the holdout set under this policy using the inverse probability weighted (IPW) policy value estimator (Yoganarasimhan et al. 2023):

$$V(\pi_{\widehat{\tau}}) = \frac{1}{\sum_{k=1}^{K} N_k} \sum_{k=1}^{N} \sum_{i_k=1}^{N_k} \frac{\mathbb{1}\big[W_{k,i_k} = \pi_{\widehat{\tau}}(\mathbf{Z}_k, \mathbf{X}_{k,i_k})\big]}{\widehat{\mathbb{P}}\big[W_{k,i_k} = \pi_{\widehat{\tau}}(\mathbf{Z}_k, \mathbf{X}_{k,i_k})\big]} \big[m \cdot Y_{k,i_k} - \pi_{\widehat{\tau}}(\mathbf{Z}_k, \mathbf{X}_{k,i_k}) \cdot c_k\big],$$

where $\widehat{\mathbb{P}}\big[W_{k,i_k} = \pi_{\widehat{\tau}}(\mathbf{Z}_k, \mathbf{X}_{k,i_k})\big]$ denotes the estimated propensity score that the observed treatment assignment in the experiment coincides with the assignment prescribed by $\pi_{\widehat{\tau}}$.[10]

[8] Since the dataset does not report either the gross margin of the promoted item or the precise cost of each campaign, we approximate margin m by adopting the average gross margin of household products in the United States (51.32%; Damodaran (2025)). Regarding the cost of each campaign, we define the cost per customer, c_k, as the total number of points awarded multiplied by 0.001, since 1,000 reward points are approximately equivalent to $1.0 in value for consumers. This cost estimate assumes that brand clients bear the full promotion cost, consistent with the platform's current business practice.

[9] While this decision rule primarily aligns with the objectives of the platform's partners, it is also critical for the platform itself. Partners choose to collaborate with the platform because it demonstrates that its promotional campaigns effectively drive incremental benefits. Therefore, targeting decisions that maximize partners' profitability also strengthen the platform's value proposition and long-term business success.

[10] Since all experiments are randomized controlled trials, we follow Yoganarasimhan et al. (2023) and use the empirical proportion of cases in which the actual treatment assignment coincides with the policy assignment as the propensity score.

To protect the platform's confidentiality, we normalize the profit values of all methods by dividing them by the profit achieved under the uniform treatment policy, which is the norm for rolling out offers in industry and used frequently as benchmark in research (e.g., Yoganarasimhan et al. 2023, Simester et al. 2020): extending an offer to all customers if its average incremental profit (i.e., the ATE on spending multiplied by m) exceeds its cost (i.e., c_k). Under this normalization, a profit measure greater than 1 indicates that a given targeting policy improves the profitability, while the deviation from 1 reflects the difference in profit relative to the standard uniform treatment policy.

6 Empirical Performance

We evaluate both the internal and external validity of the proposed DRL-based CIF model using two managerially relevant scenarios. The first assesses internal validity by testing whether the model can accurately identify which customers should receive an offer that has already been tested, using data drawn from the same distribution as past experiments. The second examines external validity by evaluating the model's ability to recommend which customers should receive offers that have not been tested before, applied to customer populations whose covariate distributions may differ from those in the past experimental data.

6.1 Targeting for Tested Offers

We first consider a standard customer targeting setting, where firms aim to decide which customers should receive an offer that has already been tested on customers drawn from the same covariate distribution. This problem is central to much of the targeting literature (e.g., Ascarza 2018, Simester et al. 2020, Zhang and Misra 2025, Yoganarasimhan et al. 2023, Hitsch et al. 2023, Huang and Ascarza 2024), where the goal is to allocate limited marketing resources efficiently by identifying customers with the highest incremental response to an intervention.

Evaluation Procedure. We implement a bootstrap sample-splitting scheme, similar to the approach in Ascarza (2018) and Hitsch et al. (2023), to evaluate model performance. Specifically, for each of the 362 offers, we generate $B = 20$ random splits of the data into training sets (70%) and test sets (30%). For each split b, we train a CATE model $\widehat{\tau}^{(b)}$ on the training set and evaluate its performance on the corresponding test set using RMSE($\widehat{\tau}^{(b)}$) and AUTOC($\widehat{\tau}^{(b)}$). Repeating this process B times allows us to compute bootstrap means and standard deviations for the key performance metrics. This sample-splitting procedure mirrors the way firms make targeting decisions in practice: targeting rules are developed from past experimental data (training) and then applied to a new pool of customers (test). In this respect, the scheme offers a realistic assessment of how different models perform in standard targeting settings.

Results. Table 3 reports performance on holdout observations across four modeling approaches. The first column lists the model specifications. The second column reports the average AUTOC value across 20 train–holdout splits, with standard deviations in parentheses. The third column shows the percentage of splits in which the CIF model with our proposed DRL architecture outperformed each benchmark on AUTOC. The fourth column presents the average normalized profit across the same splits (standard deviations in parentheses), and the fifth column shows the percentage of splits where the proposed DRL model outperformed the benchmark on profit.

Table 3: Targeting Performance on Holdout Observations

Model	AUTOC	% Outperf.	Profit	% Outperf.
Joint: Proposed DRL	0.47 (0.12)	—	1.37 (0.11)	—
Joint: Standard DNN	0.36 (0.11)	85%	1.16 (0.19)	90%
Joint: Standard DNN w/o Design Features	0.37 (0.09)	100%	1.33 (0.13)	85%
Individual: Standard DNN	0.11 (0.07)	100%	0.96 (0.07)	100%

Note: We report the mean AUTOC and normalized profit across 20 train–holdout splits. Standard deviations are shown in parentheses. The third and fifth columns ("% Outperf.") report the percentage of splits in which the proposed model (first row) outperformed each benchmark in AUTOC and normalized profit, respectively. The results demonstrate the effectiveness of DNN-based models. For the complete set of results from alternative prediction approaches, see Web Appendix F.1.

First, all joint models (Rows 1–3) outperform the individual CATE model (Row 4), both in terms of AUTOC and profits, highlighting the value of synergizing information across experiments. Importantly, the proposed DRL model (Row 1) increases profitability by about 37% relative to the uniform treatment policy, whereas the individual model (Row 4) reduces profits by roughly 4%. This demonstrates that synthesizing past experiments enables substantially more effective targeting of marketing interventions. Second, the proposed DRL architecture (Row 1) outperforms both standard DNN benchmarks (Rows 2 and 3) across AUTOC and profit. Interestingly, the standard DNN without design features (Row 3) performs better than the version that includes them (Row 2). This result underscores that incorporating high-dimensional design features into a complex DNN without proper structural regularization can reduce performance, whereas a carefully designed architecture can translate rich inputs into meaningful improvements.

6.2 Targeting for Untested Offers

Next, we consider a setting in which the platform seeks to target a new offer that has not been previously tested. This scenario is particularly relevant in practice, as firms often wish to introduce new interventions that do not have existing experimental results, while still limiting them to customers who are most likely to generate incremental profit.

Evaluation Procedure. To evaluate the ability of the proposed method to predict treatment effects for new offers, we implement the following bootstrap procedure. In each replication, we randomly divide the 362 experiments into 181 training experiments and 181 holdout experiments. The training experiments are used to construct both the proposed model and the other two joint CIF models. For each holdout experiment, we further split the data into two subsets using a 70%-30% partition. The larger (70%) subset is used to construct an individual CATE model, mimicking the standard practice in which the company directly tests a new promotion and builds a CATE model from it. The remaining 30% of the data is reserved for evaluating the performance of both the individual CATE models and the joint models. We repeat this procedure across 20 random splits and report the summary statistics of the resulting performance metrics. This setup allows us to directly compare the performance of our proposed approach with that of the conventional "test-and-optimize" strategy for new offers, in which firms first

Table 4: Targeting Performance on Holdout Offers

Model	AUTOC	% Outperf.	Profit	% Outperf.
Joint: Proposed DRL	0.47 (0.14)	—	1.19 (0.08)	—
Joint: Standard DNN	0.21 (0.13)	100%	0.97 (0.06)	100%
Joint: Standard DNN w/o Design Features	0.37 (0.11)	75%	1.11 (0.09)	90%
Individual: Standard DNN	0.02 (0.08)	100%	0.85 (0.03)	100%

Note: We report the mean AUTOC and normalized profit across 20 train–holdout offer splits. Standard deviations are shown in parentheses. The third and fifth columns ("% Outperf.") report the percentage of splits in which the proposed model (first row) outperformed each benchmark in AUTOC and normalized profit, respectively. The results demonstrate the effectiveness of DNN-based models. For the complete set of results from alternative prediction approaches, see Web Appendix F.1.

conduct experiments on the focal offers and then estimate a separate CATE model from each experiment to guide targeting decisions.

Results. Table 4 summarizes the model performance when predicting treatment effects for new offers that were not used in training. Similar to the earlier findings, the joint models (Rows 1–3) continue to outperform the individually estimated CATE models (Row 4). The proposed joint DRL model (Row 1) improves profitability by 19% relative to the uniform treatment benchmark, whereas the individual model leads to a profit loss of about 15%. In addition, the DRL architecture (Row 1) again surpasses the standard DNN benchmark (Row 2) across both evaluation metrics. Interestingly, omitting design features (Row 3) leads to better performance than including them without sufficient regularization (Row 2), suggesting that adding more input variables can actually harm generalization when the model architecture is not properly structured. Collectively, these results reinforce our key arguments: (i) synergizing information across experiments yields more robust targeting models, and (ii) architectural design choices play an important role in targeting performance for new offers.

6.3 Robustness Checks for Alternative Model Classes

In Web Appendix F, we conduct robustness checks using alternative second-stage models, including XGBoost and Elastic Net, to demonstrate that the proposed two-stage CIF estimation framework outperforms the conventional approach of estimating separate CATE models for each experiment. Overall, we find that the XGBoost-based CIF model also significantly outperforms the individual CATE modeling approach, though its performance is slightly below that of the proposed DRL architecture. By contrast, the Elastic Net models deliver limited targeting value regardless of whether experiments are synergized or modeled individually, suggesting that nonlinear models are essential for capturing treatment effect heterogeneity.

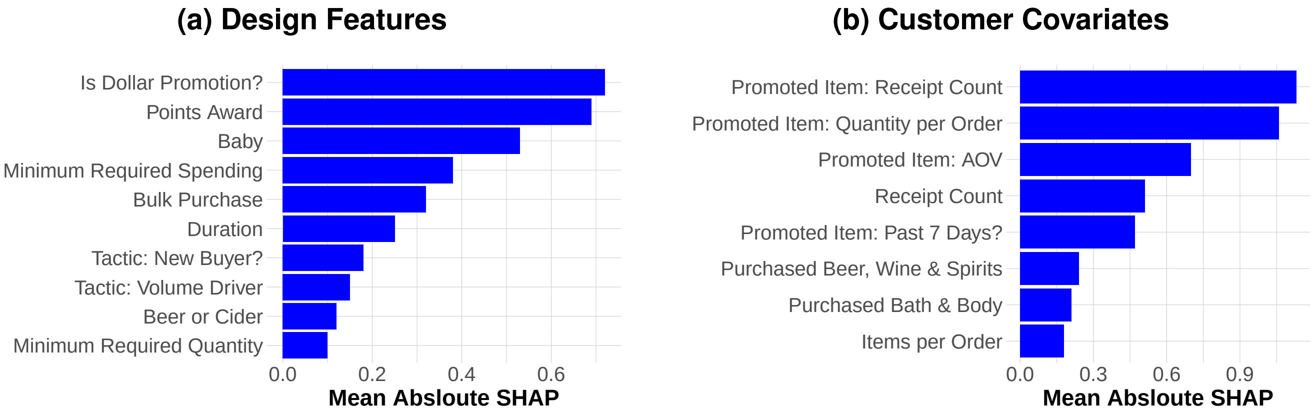
In addition to the proposed two-stage estimation framework, we further examine in Web Appendix F whether the idea of synergizing information across experiments can be extended to other treatment effect estimation approaches, such as T-learners (Künzel et al. 2019). Specifically, we construct two spending prediction models (one for the treatment group and one for the control group) using data pooled across all experiments, and compute the predicted treatment effect as the difference between the two predictions. We then compare the performance of this joint T-learner with that of conventional T-learners estimated separately for each experiment. Overall, we find that the joint T-learner also significantly outperforms the individual CATE models, regardless of the model classes used for spending models. Taken together, the findings convey a clear and promising message: synergizing past experiments creates substantial value. In practice, firms can implement our two-stage framework—or alternative methods for integrating experimental data—using standard off-the-shelf models and empirically evaluate which approach performs best in their specific context.

7 Designing Promotional Offers Using the CIF Model

So far, we have demonstrated that the proposed DRL-based CIF model effectively matches offers to customers with high treatment effects, thereby maximizing incremental profit. In this section, we extend the analysis to demonstrate how the model can also inform the design of new promotional offers.

Figure 4 reports the variable importance from the DRL-based CIF model trained on 362 experiments. Importance is quantified by the mean absolute SHAP values Lundberg and Lee (2017), which capture the average magnitude of each variable's contribution to shifting predictions away from the overall mean. Higher values indicate that a variable exerts a stronger influence on predicted treatment effects. Panel (a) shows that offer design features, such as whether the promotion required customers to spend enough on the promoted items, the reward points given to customers, and minimum spending requirements, are particularly influential in shaping treatment effects. Panel (b) highlights that customer purchase behavior with respect to the promoted item — such as prior receipt counts, quantity per order, and average order value — plays a critical role among covariates. Together, these insights indicate that the platform can enhance campaign effectiveness by calibrating offer structures (e.g., rewards and spending thresholds) and by prioritizing interventions that align with customers' past purchase intensity with respect to the promoted items in an offer.

Figure 4: Variable Importance Plot of Top 10 Design Features and Customer Covariates



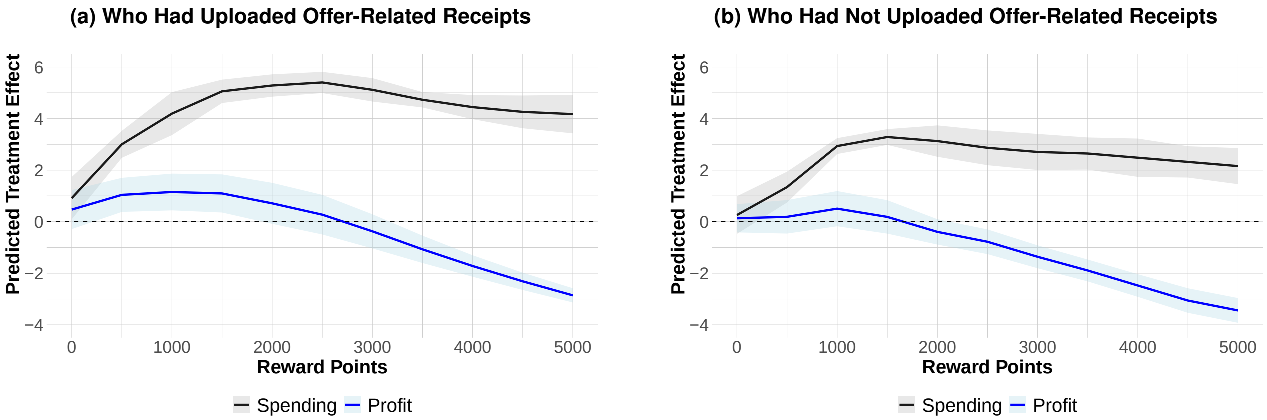
(a) Design Features

(b) Customer Covariates

Note. The mean absolute SHAP value measures how much each variable contributes to shifting model predictions away from the overall mean. The higher the value, the greater the variable's overall influence on the model's predictions.

Next, we demonstrate how the company can use the proposed model to better calibrate promotional offers for different customer types. For simplicity, we focus on two customer groups defined by whether they had uploaded a receipt related to the promoted item during the 90 days prior to the experiment, as this is the most influential covariate in the model's predictions (Figure 4). We examine two key design features of the offers: the number of reward points provided (see Figure 5) and the spending required on the promoted items to claim the reward (see Figure 6). For each of these features, we evaluate the predicted treatment effects in terms of both customer spending and profits.

Figure 5: Partial Dependence Plot of Reward Points for Two Customer Types

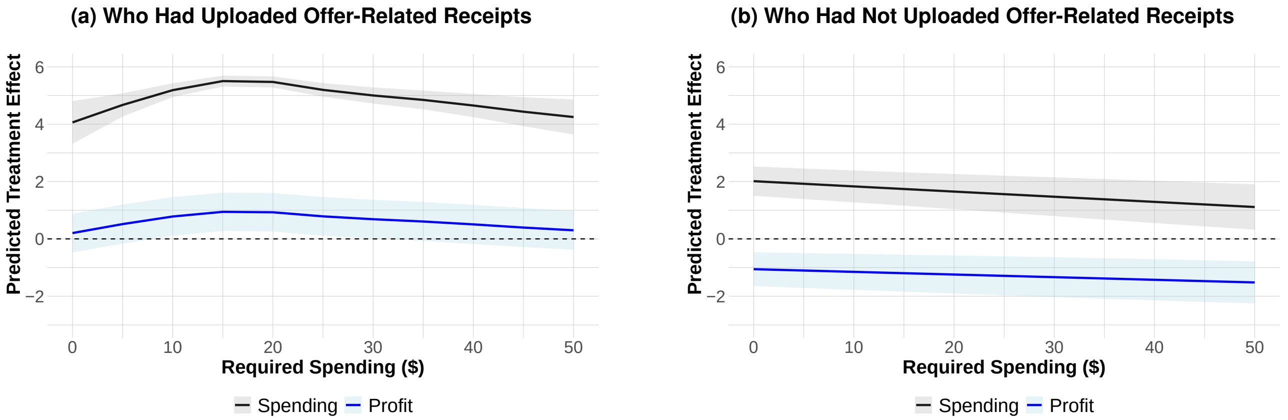


Note. Each line represents the predicted treatment effect from the model, averaged across all customers who either had (Panel (a)) or had not (Panel (b)) uploaded receipts related to the promoted item. For each panel, we vary the reward points while holding all other design features constant to isolate their marginal effect. The shaded areas represent two standard deviations across customers, reflecting the variability in predicted treatment effects.

Figure 5 presents partial dependence plots (PDPs) (Friedman 2001) of reward points for the two customer types. A PDP isolates the marginal effect of a focal variable on the model's predictions by (i) varying the value of that variable across all observations, and then (ii) averaging the resulting predictions, thereby illustrating how predictions change with the focal variable while holding all others constant. Panel (a) focuses on customers who had previously uploaded receipts for the promoted items. For this group, incremental spending rises steadily and peaks around 2,500 points, while profit peaks earlier at about 1,500 points before declining. The predicted profit effect becomes negative after 2,500 reward points, indicating that incremental spending may no longer justify the cost once the reward level is too high.

Panel (b) presents results for customers without prior receipts for the promoted items. Their incremental spending peaks around 1,500 points, while the profit-maximizing reward level is approximately 1,000 points, with profits turning negative beyond roughly 1,750 points. The magnitude of profit gains for this segment is also substantially smaller than that for prior buyers. These findings highlight the importance of segment-specific calibration: for existing buyers, offering around 1,500 points maximizes profitability, whereas for non-buyers, capping rewards at 1,000 points yields better returns.

Figure 6: Partial Dependence Plot of Minimum Required Spending for Two Customer Types


(a) Who Had Uploaded Offer-Related Receipts

(b) Who Had Not Uploaded Offer-Related Receipts

Note. Each line represents the predicted treatment effect from the model, averaged across all customers who either had (Panel (a)) or had not (Panel (b)) uploaded receipts related to the promoted item. For each panel, we vary the munimum required spending while holding all other design features constant to isolate their marginal effect. The shaded areas represent two standard deviations across customers, reflecting the variability in predicted treatment effects.

Figure 6 presents the PDP of the minimum required spending threshold for the two customer segments. Panel (a) shows that for customers who had uploaded offer-related receipts, the spending effect follows an inverse-U shape: both predicted spending and profit increase as the threshold rises, peaking around $20, and then decline at higher levels. This suggests that moderate thresholds can meaningfully increase both spending and profit among engaged customers, whereas overly high requirements reduce customers' willingness to pursue the offer and thereby diminish the treatment effect.

By contrast, Panel (b) shows that for customers who had not uploaded offer-related receipts, both predicted spending and profit decline almost linearly as the threshold increases. Raising the spending requirement appears to always deter participation, resulting in lower treatment effects. These results suggest that setting moderate spending thresholds is profitable for existing buyers, while low thresholds are preferable for less engaged customers to avoid discouraging participation.[11]

Beyond the analyses using standard model interpretation tools, Web Appendix G illustrates how the learned representations from the proposed model uncover insights for promotion de-

[11] Note that the treatment effect on profit in Panel (b) is negative because the average reward points across offers is 2,540 points, which aligns with the results shown in Figure 5 (b).

sign. The offer-level analyses show that representations are largely driven by key incentive features and that certain product categories appear close to one another in the representation space. This pattern suggests that when designing promotions, firms should focus on incentive structure and consider cross-category similarities when transferring or adapting successful promotion configurations. On the other hand, customer segmentation based on the learned representations distinguishes customers by their baseline propensity to purchase promoted products, implying that customized promotions should account not only for overall platform engagement but also for customers' own preference toward promoted items.

8 Generalizability: Evaluating Attribute Shifts and Concept Shifts

In previous sections, we presented the empirical performance of our model and demonstrated its applications in targeting and new offer design. We now evaluate the robustness of the CIF model when applying it to new offers and customers. Specifically, we examine the two types of data shifts discussed in Section 3.4: *attribute shift* and *concept shift*. For *attribute shift*, we examine how differences in design features of new offers or differences in customer covariate distributions relative to prior experiments affect the performance of the proposed CIF model. For *concept shift*, we test whether variation in customer responsiveness to an offer can be explained by the timing of the offer deployment.

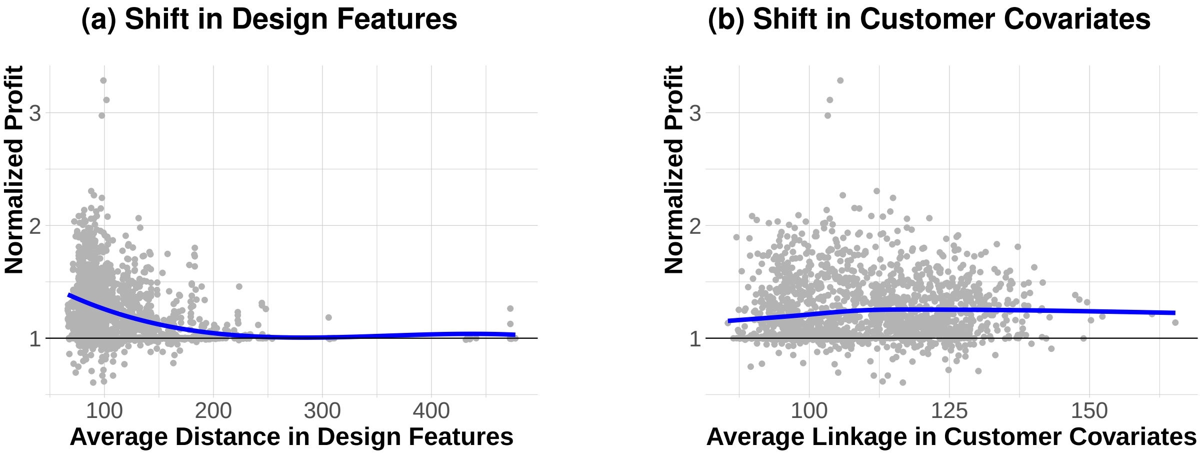
8.1 Evaluating Attribute Shifts

We first examine how attribute shift affects the performance of the proposed model. Building on the estimation approach in Section 6.2, we randomly hold out 181 offers for evaluation. For each holdout offer, we compute the distance between its design and customer features and those in the training set, and analyze how this distance relates to the model's performance.

Specifically, we quantify the difference in design features between a holdout offer h and the training offers $k = 1, \cdots, 181$ by computing the *average squared Euclidian distance* of their differences: $D_h^{\mathbf{Z}} = \frac{1}{181}\sum_{k=1}^{181} \|\mathbf{Z}_h - \mathbf{Z}_k\|_2^2$. To quantify differences in customer covariates, we randomly sample 100,000 observations from the training data and compare them with the customer covariates in each holdout experiment. We then compute the *average linkage*, defined as the mean pairwise distance between each holdout experiment and the 181 training experiments. Formally, let $\mathbf{X}_{k_i,i}$ denote the covariate vector of training observation i (with $i = 1, \cdots, 100000$) and $\mathbf{X}_{h,j}$ denote the covariate vector of customer j in the holdout experiment h. The average linkage distance is then defined as $D_h^{\mathbf{X}} = \frac{1}{100000 \times N_h} \sum_{i=1}^{100000} \sum_{j=1}^{N_h} \|\mathbf{X}_{k_i,i} - \mathbf{X}_{h,j}\|_2^2$, where N_h is the number of observations in the holdout experiment.

Next, we examine how the average distances in design features and customer covariates relate to the normalized profit metric defined in Section 5.3. We perform 20 train–holdout offer splits, yielding a total of $181 \times 20 = 3,620$ data points. Each data point corresponds to a holdout offer, characterized by its average distances from the training offers in both the design and customer spaces, as well as its associated normalized profit under the DRL-based CIF model.

Figure 7: Average Distance in Design Features, Customer Covariates, and Normalized Profit



Note. Each data point represents a holdout offer and its related metrics for a given training set in one split, characterized by its distances from the training offers in terms of design features and customer covariates, as well as its associated normalized profit. The blue line shows the kernel-smoothed estimate of the regression between average distance and the normalized profit.

Figure 7 illustrates how the average distance in design features (Panel (a)) and the average linkage in customer covariates (Panel (b)) relate to normalized profit. Panel (a) reveals a clear pattern: normalized profit decreases as the design features of new offers diverge further from those tested in past experiments. Nevertheless, even for offers with distances greater than 200 (representing 11% of all data points), normalized profit remains above 1, indicating that applying the proposed model still performs at least as well as the baseline strategy of launching an offer whenever its average incremental profit is positive. Panel (b), by contrast, shows that variations in customer covariate distributions between past experiments and new settings do not meaningfully affect profitability. This result suggests that the CIF model generalizes well across customer populations, even when covariate distributions shift.

Table 5 reports the results from regressing normalized profit on average distance in design features and average linkage in customer covariates, using the full set of 3,620 data points (181 offers and 20 splits). The results show a significantly negative coefficient for design-feature distance, while the coefficient on customer covariates is indistinguishable from zero. These findings reinforce that similarity in design features is a key determinant of generalization performance, whereas shifts in customer covariates are comparatively less critical in our context.

Table 5: Regression Evaluation of Attribute Shifts

	Dependent variable:
	Normalized Profit
Average Distance in Design Features	−0.002 (0.0001)
Average Linkage in Customer Covariates	0.0001 (0.00006)
Constant	1.430 (0.018)
Observations	3,620
Adjusted R^2	0.097
F Statistic	88.768 (df = 2; 3617)

8.2 Evaluating Concept Shift

Another important concern when applying models trained on historical experimental data to future marketing decisions is the possibility of *concept shift*: changes over time in the underlying relationship between offer design features, customer covariates, and customer responsiveness to treatment. When such shifts occur, models that once performed well may fail to generalize to future offers and customers.

To assess whether concept shift is a concern in our application, we develop an empirical procedure grounded in the double machine learning framework (e.g., Chernozhukov et al. 2018, Ellickson et al. 2023), which enables managers to detect such shifts using their own experimental data. Specifically, we test whether variation in treatment effects can be systematically attributed to the timing of the intervention, beyond what is explained by offer design features and customer covariates, for all 362 offers.

1. **Compute DR scores:** For each observation i in experiment k, compute the DR score $\tilde{\tau}_{k,i}$.
2. **Partial out variations driven by design features and covariates:** We remove variations $\tilde{\tau}_{k,i}$ that can be explained by explained by $\mathbf{Z}_k$ and $\mathbf{X}_{k,i}$. Specifically, we compute the residual:

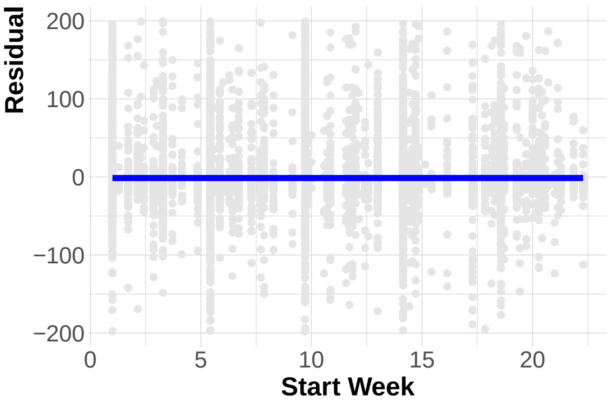
$$\tilde{\varepsilon}^{\tau}_{k,i} = \tilde{\tau}_{k,i} - \widehat{\tau}(\mathbf{Z}_k, \mathbf{X}_{k,i}),$$

   where $\widehat{\tau}(\mathbf{Z}_k, \mathbf{X}_{k,i})$ is the predicted treatment effect obtained using the DRL-based CIF model. This step isolates the portion of treatment effect variation that is not explained by observed offer features and customer covariates, which may be correlated with the offer start time.
3. **Test for temporal variation in treatment effects:** In the final step, we assess whether the residual variation in treatment effects, after controlling for $\mathbf{Z}_k$ and $\mathbf{X}_{k,i}$, can be systematically explained by the offer start time T_k. Specifically, we test for a statistically significant relationship between $\tilde{\varepsilon}^{\tau}_{k,i}$ and the offer start time T_k. A significant association would indicate that the treatment effect function varies over time, providing evidence of concept shift.

Applying these steps to our data, we conduct both a visual inspection and a regression-based test to evaluate whether offer start time explains additional variation in treatment effects. Figure 8 presents a scatter plot of offer start week (T_k) against the residual treatment effects ($\tilde{\varepsilon}^{\tau}_{k,i}$) for a random sample of 100,000 observations drawn from the full dataset of 30,701,442 observations. The absence of any discernible temporal pattern indicates that the residual treatment effects are unrelated to offer timing, suggesting that treatment effects remained stable over the sampling period and that concept shift is unlikely to be a concern in our context.

Beyond the visual evidence, we formally test for temporal variation in treatment effects by regressing the residual treatment effects, $\tilde{\varepsilon}^{\tau}_{k,i}$, on the offer start week (T_k) and its square (T_k^2), using the full sample of 30,701,442 observations. The estimation results in Table 6 suggest that

Figure 8: Relationship between Offer Start Week and Residual Treatment Effects



Note. Each point represents an observation randomly drawn from the full dataset of 30,701,442 observations. The x-axis indicates the offer's start week (T_k), normalized such that week 1 corresponds to the launch of the first offer. The y-axis represents the residual treatment effect ($\tilde{\varepsilon}^{\tau}_{k,i}$). The blue line shows the kernel-smoothed regression estimate of the realtionship between residual treatment effects and offer start time.

both coefficients are statistically indistinguishable from zero, and the model explains virtually none of the variance in residuals ($R^2 = 0$). Taken together with the visual analysis, the results provide no statistical evidence of a systematic relationship between treatment effects and offer start time.

Table 6: Regression Results of Residual Treatment Effects on Offer Start Week

	Dependent variable:
	$\tilde{\varepsilon}^{\tau}_{k,i}$
T_k	0.011 (0.029)
T_k^2	−0.001 (0.001)
Constant	0.029 (0.113)
Observations	30,701,442
Adjusted R^2	−0.000
F Statistic	0.891 (df = 2; 30701439)

9 Conclusion and Future Directions

We propose a two-stage causal machine learning framework that enables firms to personalize interventions by leveraging data from past randomized experiments. By modeling treatment effects as a joint function of intervention design features and customer covariates, the framework synergizes information across different experiments, overcoming the limitations of analyzing each experiment in isolation. This approach not only delivers substantial gains in targeting accuracy but also demonstrates strong generalizability to untested interventions.

Empirical evidence from 362 promotion experiments involving more than seven million customers demonstrates that our proposed framework predicts treatment effects more accurately than the conventional approach of estimating a separate model for each experiment. Moreover, when extrapolating to previously untested promotional offers, the framework outperforms traditional CATE models trained directly on experimental data from those same offers. Together, these results indicate that firms can transform their existing archive of experiments into a valuable resource for scalable personalization. Finally, we show that balancing flexibility with complexity is critical for model performance, and that in deep learning applications this balance can be effectively achieved through careful architecture design.

While our research provides a valuable framework for personalized marketing interventions, it also has several limitations that point to promising avenues for future work. A key practical challenge lies in managing both attribute shift and concept shift. Although our empirical evidence suggests that these shifts are not a major concern in our setting, they may arise in other contexts — for example, when a firm is in a rapid growth phase and the composition of its customer base changes substantially. To mitigate attribute shift, active learning techniques (e.g., Jesson et al. 2021, Chen et al. 2024) could be extended to guide new data collection toward offers with high uncertainty in their predicted CATEs or in their estimated profit-maximizing levels. To address concept shift, fine-tuning methods from transfer learning (e.g., Alshalali and Josyula 2018) could be employed to incrementally update the model using small-scale experiments conducted over time. Future research could integrate these strategies into a continuous experimentation framework that explicitly accounts for both attribute and concept shifts.

Second, although we employ standard post-hoc explanation techniques to demonstrate how managers can use the proposed model to inform promotional design, more granular customization could be achieved by directly personalizing intervention design. Prior research often frames this problem as one of mathematical optimization (e.g., Ansari and Mela 2003, Zhang and Krishnamurthi 2004). In our context, one could query the estimated CIF model to identify the set of design features that maximizes predicted profit for a given customer segment. Because the decision space is high-dimensional, this optimization could be implemented using approaches such as Bayesian optimization (Frazier 2018) or optimal transport (Zhang and Misra 2025). Future research could integrate these optimization procedures with our two-stage CIF framework to algorithmically generate profit-maximizing offer designs tailored to specific customer segments.

Third, while we demonstrate the benefits of synergizing experiments, extending this approach to other marketing domains could yield valuable insights. For instance, a retailer might integrate results from multiple in-store coupon experiments to inform decisions about which customers should receive incentives (Gabel et al. 2025); a hotel brand could synthesize evidence from various compliance-promotion campaigns (Daljord et al. 2023) to refine its loyalty program; and an advertising platform could leverage findings from prior ad campaign experiments (Rafieian 2023) to improve targeting for new campaigns. Evaluating the effectiveness of our framework across such diverse contexts would help clarify the conditions under which personalized marketing strategies benefit most from integrating historical experimentation.

Finally, our analysis focuses on experiments that share a common set of covariates and a single outcome of interest. In practice, however, experiments may differ in their intervention features, available covariates, or outcome variables. For example, one experiment might use demographic and purchase history data to estimate treatment effects on promotion redemption, while another might rely on browsing behavior data to predict incremental website visits from email campaigns. Jointly modeling such heterogeneous experiments could still be valuable but would require additional assumptions and methodological innovations. For cases with differing covariate sets, heterogeneous domain adaptation methods (e.g., Dai et al. 2008, Pan et al. 2010) could be extended to the second-stage model architecture to reconcile feature-space differences. For cases with distinct outcome variables, multi-outcome regression approaches (e.g., Richardson et al. 2015) could enable joint learning across experiments. Establishing the theoretical foundations and practical estimation strategies for these more complex settings represents an important avenue for future research.

References

Alshalali T, Josyula D (2018) Fine-tuning of pre-trained deep learning models with extreme learning machine. *International Conference on Computational Science and Computational Intelligence (CSCI)*, 469–473 (IEEE).

Anil C, Lucas J, Grosse R (2019) Sorting out lipschitz function approximation. *International conference on machine learning*, 291–301 (PMLR).

Ansari A, Mela CF (2003) E-customization. *Journal of Marketing Research* 40(2):131–145.

Ascarza E (2018) Retention futility: Targeting high-risk customers might be ineffective. *Journal of Marketing Research* 55(1):80–98.

Bach P, Chernozhukov V, Kurz MS, Spindler M (2022) Doubleml-an object-oriented implementation of double machine learning in python. *Journal of Machine Learning Research* 23(53):1–6.

Bartlett PL, Mendelson S (2002) Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3(Nov):463–482.

Bastani H (2021) Predicting with proxies: Transfer learning in high dimension. *Management Science* 67(5):2964–2984.

Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW (2010) A theory of learning from different domains. *Machine learning* 79:151–175.

Chen F, Liu X, Proserpio D, Troncoso I (2022) Product2vec: Leveraging representation learning to model consumer product choice in large assortments. *NYU Stern School of Business* .

Chen YW, Ascarza E, Netzer O (2024) Policy-aware experimentation: Strategic sampling for optimized targeting policies. *Columbia Business School Research Paper* (5044549).

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1):C1–C68.

Chollet F, et al. (2015) Keras. `https://keras.io`.

Cisse M, Bojanowski P, Grave E, Dauphin Y, Usunier N (2017) Parseval networks: Improving robustness to adversarial examples. *International conference on machine learning*, 854–863 (PMLR).

Crump RK, Hotz VJ, Imbens GW, Mitnik OA (2009) Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96(1):187–199.

Cybenko G (1989) Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* 2(4):303–314.

Dai W, Chen Y, Xue GR, Yang Q, Yu Y (2008) Translated learning: Transfer learning across different feature spaces. *Advances in neural information processing systems* 21.

Daljord Ø, Mela CF, Roos JM, Sprigg J, Yao S (2023) The design and targeting of compliance promotions. *Marketing Science* 42(5):866–891.

Damodaran A (2025) Operating and net margins by sector (us). `https://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/margin.html`, accessed January 2025.

Dew R, Ansari A, Toubia O (2022) Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science* 41(2):401–425.

Drucker H, Le Cun Y (1992) Improving generalization performance using double backpropagation. *IEEE transactions on neural networks* 3(6):991–997.

Ellickson PB, Kar W, Reeder III JC (2023) Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science* 0(0):1–22.

Ellickson PB, Kar W, Reeder III JC, Zeng G (2024) Using contextual embeddings to predict the effectiveness of novel heterogeneous treatments. *Available at SSRN 4845956* .

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–231, KDD'96 (AAAI Press).

Frazier PI (2018) A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811* .

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 1189–1232.

Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M (2011) Doubly robust estimation of causal effects. *American journal of epidemiology* 173(7):761–767.

Gabel S, Guhl D, Klapper D (2019) P2v-map: Mapping market structures for large retail assortments. *Journal of Marketing Research* 56(4):557–580.

Gabel S, Simester D, Timoshenko A (2025) In-store coupons: A large-scale field experiment .

Gabel S, Timoshenko A (2022) Product choice with large assortments: A scalable deep-learning model. *Management Science* 68(3):1808–1827.

Golowich N, Rakhlin A, Shamir O (2018) Size-independent sample complexity of neural networks. *International Conference on Learning Theory*, 297–299 (PMLR).

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of wasserstein gans. *Advances in neural information processing systems* 30.

Haushofer J, Niehaus P, Paramo C, Miguel E, Walker M (2025) Targeting impact versus deprivation. *American Economic Review* 115(6):1936–1974.

Hitsch GJ, Misra S, Zhang W (2023) Heterogeneous treatment effects and optimal targeting policy evaluation. *Available at SSRN 3111957* .

Hornik K (1991) Approximation capabilities of multilayer feedforward networks. *Neural Networks* 4(2):251–257.

Huang TW, Ascarza E (2024) Doing more with less: Overcoming ineffective long-term targeting using short-term signals. *Marketing Science* 0(0).

Imai K, Nakamura K (2024) Causal representation learning with generative artificial intelligence: Application to texts as treatments. *arXiv preprint arXiv:2410.00903* .

Imbens GW, Rubin DB (2015) *Causal inference in statistics, social, and biomedical sciences* (Cambridge university press).

Jesson A, Tigas P, van Amersfoort J, Kirsch A, Shalit U, Gal Y (2021) Causal-bald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data. *Advances in Neural Information Processing Systems* 34:30465–30478.

Kennedy EH (2023) Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics* 17(2):3008–3049.

Khan S, Saveski M, Ugander J (2023) Off-policy evaluation beyond overlap: partial identification through smoothness. *arXiv preprint arXiv:2305.11812* .

Kidger P, Lyons T (2020) Universal Approximation with Deep Narrow Networks. Abernethy J, Agarwal S, eds., *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, 2306–2327 (PMLR), URL `https://proceedings.mlr.press/v125/kidger20a.html`.

Kini V, Manjunatha A (2020) Revenue maximization using multitask learning for promotion recommendation. *2020 International Conference on Data Mining Workshops (ICDMW)*, 144–150 (IEEE).

Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116(10):4156–4165.

Ledoux M, Talagrand M (2013) *Probability in Banach Spaces: isoperimetry and processes* (Springer Science & Business Media).

Lemmens A, Roos J, Gabel S, Ascarza E, Bruno H, Gordon B, Israeli A, Feit EM, Mela C, Netzer O (2025) Personalization and targeting: How to experiment, learn & optimize. *International Journal of Research in Marketing* .

Liang S, Srikant R (2016) Why deep neural networks for function approximation? *arXiv preprint arXiv:1610.04161* .

Liberali G, Ferecatu A (2022) Morphing for consumer dynamics: Bandits meet hidden markov models. *Marketing Science* 41(4):769–794.

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30.

Ma L, Huang TW, Ascarza E, Israeli A (2025) Dynamic personalization with multiple customer signals: Multi-response state representation in reinforcement learning. *Available at SSRN* .

Maurer A, Pontil M, Romera-Paredes B (2016) The benefit of multitask representation learning. *Journal of Machine Learning Research* 17(81):1–32.

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Nethery RC, Mealli F, Dominici F (2019) Estimating population average causal effects in the presence of non-overlap: The effect of natural gas compressor station exposure on cancer mortality. *The annals of applied statistics* 13(2):1242.

Nie X, Wager S (2021) Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108(2):299–319.

Pan SJ, Ni X, Sun JT, Yang Q, Chen Z (2010) Cross-domain sentiment classification via spectral feature alignment. *Proceedings of the 19th international conference on World wide web*, 751–760.

Petersen ML, Porter KE, Gruber S, Wang Y, Van Der Laan MJ (2012) Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* 21(1):31–54.

Pinkus A (1999) Approximation theory of the mlp model in neural networks. *Acta numerica* 8:143–195.

Rafieian O (2023) A matrix completion solution to the problem of ignoring the ignorability assumption .

Richardson DB, Hamra GB, MacLehose RF, Cole SR, Chu H (2015) Hierarchical regression for analyses of multiple outcomes. *American journal of epidemiology* 182(5):459–467.

Semenova V, Chernozhukov V (2021) Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal* 24(2):264–289.

Simester D, Timoshenko A, Zoumpoulis SI (2020) Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Science* 66(6):2495–2522.

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* .

Timoshenko A, Ibragimov M, Simester D, Parker J, Schoar A (2020) Transferring information between marketing campaigns to improve targeting policies. Technical report, Working Paper.

Tripuraneni N, Jordan M, Jin C (2020) On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems* 33:7852–7862.

Van der Maaten L, Hinton G (2008) Visualizing data using t-sne. *Journal of Machine Learning Research* 9(11).

Virmaux A, Scaman K (2018) Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems* 31.

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge university press).

Wang Y, Lewis M, Cryder C, Sprigg J (2016) Enduring effects of goal achievement and failure within customer loyalty programs: A large-scale field experiment. *Marketing Science* 35(4):565–575.

Xu Z, Meisami A, Tewari A (2021) Decision making problems with funnel structure: a multi-task learning approach with application to email marketing campaigns. *International Conference on Artificial Intelligence and Statistics*, 127–135 (PMLR).

Yadlowsky S, Fleming S, Shah N, Brunskill E, Wager S (2025) Evaluating treatment prioritization rules via rank-weighted average treatment effects. *Journal of the American Statistical Association* 120(549):38–51.

Ye Z, Zhang Z, Zhang D, Zhang H, Zhang RP (2025) Deep learning based causal inference for large-scale combinatorial experiments: Theory and empirical evidence. *Management Science* .

Yoganarasimhan H (2020) Search personalization using machine learning. *Management Science* 66(3):1045–1070.

Yoganarasimhan H, Barzegary E, Pani A (2023) Design and evaluation of optimal free trials. *Management Science* 69(6):3220–3240.

Yoshida Y, Miyato T (2017) Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941* .

Zeiler MD (2012) Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* .

Zhang J, Krishnamurthi L (2004) Customizing promotions in online stores. *Marketing science* 23(4):561–578.

Zhang WW, Misra S (2025) Coarse personalization. *arXiv preprint arXiv:2204.05793* .

Zhang Y, Yang Q (2021) A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering* 34(12):5586–5609.

Zhao Z, Harinen T (2019) Uplift modeling for multiple treatments with cost optimization. *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 422–431 (IEEE).

Zhu AY, Mitra N, Roy J (2023) Addressing positivity violations in causal effect estimation using gaussian process priors. *Statistics in Medicine* 42(1):33–51.

Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q (2020) A comprehensive survey on transfer learning. *Proceedings of the IEEE* 109(1):43–76.

Zivich PN, Edwards JK, Lofgren ET, Cole SR, Shook-Sa BE, Lessler J (2024) Transportability without positivity: a synthesis of statistical and simulation modeling. *Epidemiology* 35(1):23–31.

Web Appendix

Web Appendix A Proofs of Identification and Unbiasedness of DR Score

Web Appendix A.1 Proof for Theorem 1

We begin by presenting the proof of identification of the CIF function under the five conditions stated in Assumption 1 to Assumption 5. For reference, recall Theorem 1.

Theorem 1 (CIF Identification) Under Assumption 1 to Assumption 5, for any design feature $\mathbf{Z} \in \mathcal{Z}$ that was tested in at least one experiment (i.e., there exists $k \in \{1, \cdots, K\}$ such that $\mathbf{Z}_k = \mathbf{Z}$) and any individual with covariates $\mathbf{X} \in \mathcal{X}$, the CIF defined in Equation (1) is pointwise identified such that $\tau(\mathbf{Z}_k, \mathbf{X}_{k,i}) = \mathbb{E}[Y_{k,i}|W_{k,i} = 1, \mathbf{X}_{k,i}] - \mathbb{E}[Y_{k,i}|W_{k,i} = 0, \mathbf{X}_{k,i}]$.

Proof. By the first three assumptions (overlap, unconfoundedness, and no interference), we have

$$\mathbb{E}[Y_{k,i}|W_{k,i} = 1, \mathbf{X}_{k,i}] = \mathbb{E}[Y_{k,i}(W_{k,i} = 1)|\mathbf{X}_{k,i}],$$
$$\mathbb{E}[Y_{k,i}|W_{k,i} = 0, \mathbf{X}_{k,i}] = \mathbb{E}[Y_{k,i}(W_{k,i} = 0)|\mathbf{X}_{k,i}]$$

for all the past experiments. Then, by the ignorability of unobserved design features and stability, we have

$$\mathbb{E}[Y_{k,i}(W_{k,i} = 1)|\mathbf{Z}_k, \mathbf{X}_{k,i}] = \mathbb{E}[Y_{k,i}(\mathbf{Z}_k)|\mathbf{X}_{k,i}],$$
$$\mathbb{E}[Y_{k,i}(W_{k,i} = 0)|\mathbf{Z}_k, \mathbf{X}_{k,i}] = \mathbb{E}[Y_{k,i}(\mathbf{0})|\mathbf{X}_{k,i}],$$

Then, by the stability assumption, we have

$$\mathbb{E}[Y_{k,i}(\mathbf{Z}_k)|\mathbf{X}_{k,i}] = \mathbb{E}[Y(\mathbf{Z}_k)|\mathbf{X}_{k,i}],$$
$$\mathbb{E}[Y_{k,i}(\mathbf{0})|\mathbf{X}_{k,i}] = \mathbb{E}[Y(\mathbf{0})|\mathbf{X}_{k,i}],$$

Combining the above results, we prove the identification results. ■

Web Appendix A.2 Proof for Corollary 1

Corollary 1 (Unbiased Score) Let $\mu_{k,w}(\mathbf{x}) = \mathbb{E}[Y_{k,i}|W_{k,i} = w, \mathbf{X}_{k,i} = \mathbf{x}]$ be the conditional expected outcome for offer k, and let $\pi_k(\mathbf{x}) = \mathbb{P}[W_{k,i} = 1|\mathbf{X}_{k,i} = \mathbf{x}]$ denote the propensity score of being treated in experiment k. Then, under Assumption 1 to Assumption 5, the DR score is an unbiased signal for the CIF defined in (1):

$$\tau(\mathbf{Z}_k, \mathbf{X}_{k,i}) = \mathbb{E}\left[\mu_{k,1}(\mathbf{X}_{k,i}) - \mu_{k,0}(\mathbf{X}_{k,i}) + \tfrac{W_{k,i}-\pi_k(\mathbf{X}_{k,i})}{\pi_k(\mathbf{X}_{k,i})[1-\pi_k(\mathbf{X}_{k,i})]}\left[Y_{k,i} - \mu_{k,W_{k,i}}(\mathbf{X}_{k,i})\right] \middle| \mathbf{Z}_k, \mathbf{X}_{k,i}\right].$$

Proof. From the proof of Theorem 1, we have

$$\mu_{k,1}(\mathbf{X}_{k,i}) = \mathbb{E}[Y_{k,i}(\mathbf{Z}_k)|\mathbf{X}_{k,i}] \quad \text{and} \quad \mu_{k,0}(\mathbf{X}_{k,i}) = \mathbb{E}[Y_{k,i}(\mathbf{0})|\mathbf{X}_{k,i}].$$

As a result, the target estimand can be written as:

$$\tau(\mathbf{Z}_k, \mathbf{X}_{k,i}) = \mu_{k,1}(\mathbf{X}_{k,i}) - \mu_{k,0}(\mathbf{X}_{k,i}). \tag{App-1}$$

Next, by definition of the propensity score $\pi_k(\mathbf{X}_{k,i})$, we have

$$\begin{aligned}\mathbb{E}\left[W_{k,i} \mid \mathbf{X}_{k,i}, \mathbf{Z}_k\right] &= \mathbb{P}\left[W_{k,i} = 1 \mid \mathbf{X}_{k,i}\right] \cdot 1 + \mathbb{P}\left[W_{k,i} = 0 \mid \mathbf{X}_{k,i}\right] \cdot 0 \\ &= \mathbb{P}\left[W_{k,i} = 1 \mid \mathbf{X}_{k,i}\right] \\ &= \pi_k(\mathbf{X}_{k,i}).\end{aligned}$$

These two results imply that

$$\begin{aligned}&\mathbb{E}\left[\frac{W_{k,i} - \pi_k(\mathbf{X}_{k,i})}{\pi_k(\mathbf{X}_{k,i})[1-\pi_k(\mathbf{X}_{k,i})]}\left[Y_{k,i} - \mu_{k,W_{k,i}}(\mathbf{X}_{k,i})\right] \middle| \mathbf{X}_{k,i}, \mathbf{Z}_k\right] \\ &\quad = \mathbb{E}\left[\frac{W_{k,i} - \pi_k(\mathbf{X}_{k,i})}{\pi_k(\mathbf{X}_{k,i})[1-\pi_k(\mathbf{X}_{k,i})]} \middle| \mathbf{X}_{k,i}, \mathbf{Z}_k\right] \cdot \mathbb{E}\left[Y_{k,i} - \mu_{k,W_{k,i}}(\mathbf{X}_{k,i}) \middle| \mathbf{X}_{k,i}, \mathbf{Z}_k\right] \\ &\quad = 0.\end{aligned} \tag{App-2}$$

The first equation holds by the unconfoundedness assumption (such that $W_{k,i}$ is independent of $Y_{k,i}$ given $\mathbb{X}_{k,i}$, and the second equation hold by the above results. Finally, combining (App-1) and (App-2) proves Corollary 1. ■

Note that Equation (App-2) also demonstrates the double robustness property: if either $\pi_k(\mathbf{X}_{k,i})$ or $\mu_{k,W_{k,i}}(\mathbf{X}_{k,i})$ is unbiased for its true counterpart, the estimator remains unbiased.

This implies that in randomized controlled experiments, where the true propensity score is known, the DR score always provides an unbiased signal for the true CIF.

Web Appendix B Theoretical Guarantees

In this appendix, we provide statistical guarantees for the proposed two-stage CIF estimation framework. Specifically, we build on existing theory in multitask learning — particularly Tripuraneni et al. (2020) — to show that as the number of experiments increases, the model achieves mean-squared consistent estimation of the treatment effects with high probability.

Web Appendix B.1 Data Generating Procedure

We begin by formalizing the CIF prediction problem. Specifically, we assume that the true CIF can be decomposed into two components: (i) a representation function that transforms raw intervention features and customer covariates into a lower-dimensional representation relevant for treatment effects, and (ii) a mapping function that links this representation to the corresponding treatment effect. Formally, this decomposition can be expressed as:

Assumption App-1 (CIF Function) *Assume that the true CIF function defined in Equation* (1) *can be written as*

$$\tau(\mathbf{Z}, \mathbf{X}) = f^{\star} \circ \mathbf{h}^{\star}(\mathbf{Z}, \mathbf{X}),$$

where $\mathbf{h}^{\star} : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}^{K}$ *denotes the representation function and* $f^{\star} : \mathbb{R}^{K} \to [-D, D], D < \infty$ *is the mapping function. The symbol* $\circ$ *denotes function composition; that is,* $f^{\star} \circ \mathbf{h}^{\star}(\mathbf{Z}, \mathbf{X}) \equiv f^{\star}(\mathbf{h}^{\star}(\mathbf{Z}, \mathbf{X}))$.

Note that this assumption does not require the existence of low-dimensional representations that fully capture treatment-relevant information. If all information in the original intervention features and customer covariates is relevant for determining treatment effects, then the representation function simply reduces to the identity mapping, i.e., $\mathbf{h}^{\star}(\mathbf{Z}_k, \mathbf{X}_{k,i}) = (\mathbf{Z}_k, \mathbf{X}_{k,i})$.

Next, we introduce additional assumptions concerning the sampling process underlying the experimental data. Beyond Assumption 1 to Assumption 5, we impose the following assumption on the firm's experimentation protocol:

Assumption App-2 (Experiment Protocol) *We assume the company follows the protocol below when conducting experiments:*

1. *Each offer with design features $\mathbf{Z} \in \mathcal{Z}$ has a non-zero probability of being selected for testing in an experiment.*
2. *For simplicity and without loss of generality, we assume that each past experiment involves the same number of observations, denoted by $N > 0$.*

Note that Assumption App-2.1 requires only that the firm's experimentation protocol allow for exploration across the entire design feature space. We do not assume that interventions are sampled uniformly from the decision space; firms may adaptively select interventions based on prior knowledge or past outcomes, so long as the protocol ensures sufficient coverage of the design space. Assumption App-2.2 is introduced solely to simplify notation.

Web Appendix B.2 Assumptions on Model Classes

The main objective of our two-stage CIF estimation framework is learning to predict the DR score $\tilde{\tau}_{k,i}$ through the minimization of a loss function ℓ, i.e.,

$$(\widehat{f}, \widehat{\mathbf{h}}) = \underset{f \in \mathcal{F},\ \mathbf{h} \in \mathcal{H}}{\arg\min} \ \frac{1}{KN} \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k^E} \ell\left(f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i}), \tilde{\tau}_{k,i}\right). \tag{App-3}$$

Here, $\mathcal{H}$ and $\mathcal{F}$ denote the classes of representation and prediction models employed in the second-stage estimation (e.g., deep neural networks). Note that, the second-stage model does not require constructing an explicit representation function $\widehat{\mathbf{h}}$. For instance, when a standard machine learning model is used to predict the DR score directly from the raw inputs, the representation function can be interpreted as the identity mapping.

In our application, we focus on the squared loss, defined as $\ell(a, b) = (a-b)^2$. In addition, we impose the following standard regularity conditions on the CATE model and the loss function.

Assumption App-3 (Assumption 1 in Tripuraneni et al. (2020))

1. *Every function within $\mathcal{F}$ is L-Lipschitz w.r.t. the norm $\|\cdot\|_2$, i.e., $\|f(\mathbf{a}) - f(\mathbf{a}')\|_2 \leq L\|\mathbf{a} - \mathbf{a}'\|_2$ for all $f \in \mathcal{F}$ and $\mathbf{a}, \mathbf{a}' \in dom(f)$.*

2. *The loss function ℓ is nonnegative and B-bounded, i.e., $0 \leq \ell(y', y) \leq B$ for any $y, y' \in \mathcal{R}$.*

The first condition requires that the class of prediction models used in the second-stage estimation be Lipschitz continuous. This property can be ensured by imposing suitable structural constraints on the chosen model. For neural networks, several strategies have been proposed in the literature, including using Lipschitz-continuous activation functions such as sigmoid, hyperbolic tangent, or ReLU (Szegedy et al. 2013); applying regularization techniques (e.g., Drucker and Le Cun 1992, Gulrajani et al. 2017); and imposing constraints on network weights (e.g., Cisse et al. 2017, Yoshida and Miyato 2017) or gradients (e.g., Virmaux and Scaman 2018, Anil et al. 2019) during training.

The second condition requires that the loss function be bounded. This requirement is generally not restrictive, as the prediction target, the DR score, and the true CIF function are finite whenever the outcome of interest is finite. For most model classes that represent predictions as weighted averages of outcomes, such as tree-based methods, nearest-neighbor estimators, or kernel regressions (Huang and Ascarza 2024), the boundedness condition is automatically satisfied whenever the outcome variable is finite. If concerns arise about unbounded predictions in neural network models, this can be addressed by employing bounded activation functions (e.g., sigmoid or tanh) or by directly constraining network weights (e.g., Cisse et al. 2017, Yoshida and Miyato 2017).

Web Appendix B.3 Prediction Error Bound for Two-stage CIF Estimation

We now provide the error-bound for our two-stage framework. We first define the expected prediction error as the average mean squared error between the predictions of a model and the true CIF function:

$$\text{Error}\,(f, \mathbf{h}) = \mathbb{E}\left[\frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N}\ell\left(f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i}), f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i})\right)\right], \tag{App-4}$$

where the expectation is taken over both the experimental protocol for past experiments (i.e., the distribution of $\mathbf{Z}_k$) and the data-generating process within each experiment (i.e., the distribution of $(\mathbf{X}_{k,i}, W_{k,i}, Y_{k,i})$).

In standard supervised learning settings where the prediction target is directly observed, such as the supervised multi-task learning problem in (Tripuraneni et al. 2020), it can be shown that the gap between the empirical prediction error and the expected prediction error converges to zero with high probability as the sample size grows. Consequently, minimizing the empirical prediction error is asymptotically equivalent to minimizing the true expected error. In contrast, in the context of treatment effect prediction, the true treatment effect is not directly observed, and thus we can only minimize the empirical loss with respect to the DR score rather than the empirical discrepancy between predictions and the true treatment effect. This motivates the need to formally establish the equivalance between the expected loss computed using the DR score and the true expected prediction error with respect to the underlying treatment effect.We can then adapt the proof techniques from Tripuraneni et al. (2020) to establish a bound on the gap between the empirical loss evaluated with respect to the DR score and its corresponding expected loss.

The following lemma establishes the connection between the expected loss with respect to the DR score and the loss with respect to the true CIF. Specifically, if the DR score $\tilde{\tau}_{k,i}$ is an unbiased proxy for the true treatment effect, then the difference between the expected loss of the predicted CIF relative to the DR score and the expected loss of the true CIF relative to the DR score is exactly the true expected prediction error of the predicted CIF relative to the true CIF.

Lemma App-1 *For a given representation function h and a prediction model f, the difference in mean squared loss between $f \circ \mathbf{h}$ and the DR scores $\tilde{\tau}_{k,i}$, compared to that of the oracle $f^\star \circ \mathbf{h}^\star$, is defined as:*

$$\mathcal{L}(f, \mathbf{h}, f^\star, \mathbf{h}^\star) = \mathbb{E}\left[\frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N} \ell\left(f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i}), \tilde{\tau}_{k,i}\right) - \ell\left(f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i}), \tilde{\tau}_{k,i}\right)\right].$$

Then, this difference is equivalent to the expected prediction error in Equation (App-4)*, i.e.,*

$$l(f, \mathbf{h}, f^\star, \mathbf{h}^\star) = \mathit{Error}\,(f, \mathbf{h})\,.$$

Proof. We first note that

$$
\begin{aligned}
&\mathbb{E}\left[\ell\left(f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i}), \tilde{\tau}_{k,i}\right) | \mathbf{Z}_k, \mathbf{X}_{k,i}\right] \\
&\quad = \left[f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i})\right]^2 - 2f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i})\mathbb{E}\left[\tilde{\tau}_{k,i}|\mathbf{Z}_k, \mathbf{X}_{k,i}\right] + \mathbb{E}\left[\tilde{\tau}_{k,i}^2|\mathbf{Z}_k, \mathbf{X}_{k,i}\right] \\
&\quad = \left[f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i})\right]^2 - 2f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i})\mathbb{E}\left[\tilde{\tau}_{k,i}|\mathbf{Z}_k, \mathbf{X}_{k,i}\right] + \mathbb{E}^2\left[\tilde{\tau}_{k,i}|\mathbf{Z}_k, \mathbf{X}_{k,i}\right] + \text{Var}\left[\tilde{\tau}_{k,i}|\mathbf{Z}_k, \mathbf{X}_{k,i}\right] \\
&\quad = \left[f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i})\right]^2 - 2f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i}) \cdot f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i}) + \left[f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i})\right]^2 + \text{Var}\left[\tilde{\tau}_{k,i}|\mathbf{Z}_k, \mathbf{X}_{k,i}\right].
\end{aligned}
$$

for given f and $\mathbf{h}$ since $\tilde{\tau}_{k,i}$ is unbiased. Similarly, we have

$$
\begin{aligned}
&\mathbb{E}\left[\ell\left(f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i}), \tilde{\tau}_{k,i}\right) | \mathbf{Z}_k, \mathbf{X}_{k,i}\right] \\
&\quad = \left[f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i})\right]^2 - 2f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i})\mathbb{E}\left[\tilde{\tau}_{k,i}|\mathbf{Z}_k, \mathbf{X}_{k,i}\right] + \mathbb{E}^2\left[\tilde{\tau}_{k,i}|\mathbf{Z}_k, \mathbf{X}_{k,i}\right] + \text{Var}\left[\tilde{\tau}_{k,i}|\mathbf{Z}_k, \mathbf{X}_{k,i}\right] \\
&\quad = \left[f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i})\right]^2 - 2\left[f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i})\right]^2 + \left[f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i})\right]^2 + \text{Var}\left[\tilde{\tau}_{k,i}|\mathbf{Z}_k, \mathbf{X}_{k,i}\right] \\
&\quad = \text{Var}\left[\tilde{\tau}_{k,i}|\mathbf{Z}_k, \mathbf{X}_{k,i}\right].
\end{aligned}
$$

Combining these two observations together, we have

$$
\begin{aligned}
&\mathbb{E}\left[\ell\left(f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i}), \tilde{\tau}_{k,i}\right) - \ell\left(f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i}), \tilde{\tau}_{k,i}\right) | \mathbf{Z}_k, \mathbf{X}_{k,i}\right] \\
&\quad = \left[f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i})\right]^2 - 2f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i}) \cdot f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i}) + \left[f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i})\right]^2 \\
&\quad = \ell\left(f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i}), f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i})\right).
\end{aligned}
$$

By taking expectation of the above result over the distribution of $\mathbf{Z}_k$ and $\mathbf{X}_{k,i}$, we prove the lemma. ■

Next, we apply the proof technique of Tripuraneni et al. (2020) to bound the gap between the empirical loss with respect to the DR score and the corresponding expected loss. Before presenting the formal theorem, we review a key measure of model complexity from statistical learning theory, namely Gaussian complexity (Bartlett and Mendelson 2002, Maurer et al. 2016, Tripuraneni et al. 2020). Gaussian complexity provides a standard tool for characterizing the richness of a function class — that is, its ability to fit diverse patterns in the data, including those driven by noise.

Definition App-1 (Gaussian Complexity) *For a generic function class $\mathcal{Q}$, comprising functions $\mathbf{q}(\cdot) = (q_1(\cdot), \cdots, q_B(\cdot)) : \mathbb{R}^A \to \mathbb{R}^B$, and given N observed data points, $\mathcal{D} = \{\mathbf{d}_1, \cdots, \mathbf{d}_N\}$, the* empirical Gaussian complexity *of the function class is defined as*

$$\widehat{\mathfrak{G}}_{\mathcal{D}}(\mathcal{Q}) = \mathbb{E}_{\varepsilon_{n,b}}\left[\sup_{\mathbf{q}\in\mathcal{Q}} \frac{1}{N}\sum_{n=1}^{N}\sum_{b=1}^{B} \varepsilon_{n,b}\, q_b(\mathbf{d}_n)\right], \quad \varepsilon_{n,b} \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

Following this, the worst-case Gaussian complexity is defined as $\overline{\mathfrak{G}}(\mathcal{Q}) = \arg\max_{\mathcal{D}} \widehat{\mathfrak{G}}_{\mathcal{D}}(\mathcal{Q})$, and the population Gaussian complexity is defined as $\mathfrak{G}(\mathcal{Q}) = \mathbb{E}_{\mathcal{D}}\left[\widehat{\mathfrak{G}}_{\mathcal{D}}(\mathcal{Q})\right]$.

Intuitively, Gaussian complexity measures the extent to which a function class can fit pure white noise. If the class can closely approximate randomly generated Gaussian noise, it indicates high flexibility and thus greater risk of overfitting. In this sense, Gaussian complexity captures the trade-off between expressive power and generalization: higher values reflect richer function classes but also signal a greater potential to overfit.

We now formally present an upper bound — adapted from Theorem 1 in Tripuraneni et al. (2020) — for the expected prediction error of our two-stage estimation framework:

Theorem App-1 (Upper Bound for Prediction Error) *Suppose that the following conditions hold: (i) the regularity conditions in Assumption App-3; (ii) the data is collected under an experimental protocol satisfying Assumption App-2; and (iii) there exist $f^\star \in \mathcal{F}$ and $\mathbf{h}^\star \in \mathcal{H}$,*[1] *then, with probability at least $1-\delta$, the prediction error of the model $(\widehat{f}, \widehat{\mathbf{h}})$ obtained from the empirical risk minimization problem in* (App-3) *is bounded as follows:*

$$\text{Error}\left(\widehat{f}, \widehat{\mathbf{h}}\right) \leq \underbrace{\frac{C_1}{(KN)^2} + \log(KN)\left[C_2\mathfrak{G}(\mathcal{H}) + C_3\overline{\mathfrak{G}}(\mathcal{F})\right]}_{\textit{Overfitting Potential for the Model Classes}} + C_4\sqrt{\frac{\log(2/\delta)}{KN}}, \qquad \text{(App-5)}$$

where C_1, C_2,, C_3, and C_4 are some constants.

[1] While this assumption may initially appear strong, it is reasonable in the context of deep neural networks. Prior work has established their universal approximation capability (e.g., Cybenko 1989, Hornik 1991, Pinkus 1999, Liang and Srikant 2016, Kidger and Lyons 2020), demonstrating their effectiveness in approximating a broad range of function classes.

Proof. First, define the empirical excess loss for $f, \mathbf{h}$ is defined as:

$$\widehat{\mathcal{L}}(f, \mathbf{h}, f^\star, \mathbf{h}^\star) = \frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N} \ell\left(f \circ \mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i}), \tilde{\tau}_{k,i}\right) - \mathbb{E}\left[\frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N} \ell\left(f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i}), \tilde{\tau}_{k,i}\right)\right].$$

Let $\widehat{f}$ and $\widehat{\mathbf{h}}$ denote the representation and prediction models obtained by solving the empirical loss minimization problem in Equation (App-3). Then, we have:

$$\begin{aligned}
\mathcal{L}(\widehat{f}, \widehat{\mathbf{h}}, f^\star, \mathbf{h}^\star) &= \mathcal{L}(\widehat{f}, \widehat{\mathbf{h}}, f^\star, \mathbf{h}^\star) - \underbrace{\mathcal{L}(f^\star, \mathbf{h}^\star, f^\star, \mathbf{h}^\star)}_{=0} \\
&= \underbrace{\mathcal{L}(\widehat{f}, \widehat{\mathbf{h}}, f^\star, \mathbf{h}^\star) - \widehat{\mathcal{L}}(\widehat{f}, \widehat{\mathbf{h}}, f^\star, \mathbf{h}^\star)}_{\equiv a} \\
&\quad + \underbrace{\widehat{\mathcal{L}}(\widehat{f}, \widehat{\mathbf{h}}, f^\star, \mathbf{h}^\star) - \widehat{\mathcal{L}}(f^\star, \mathbf{h}^\star, f^\star, \mathbf{h}^\star)}_{\equiv b} \\
&\quad + \underbrace{\widehat{\mathcal{L}}(f^\star, \mathbf{h}^\star, f^\star, \mathbf{h}^\star) - \mathcal{L}(f^\star, \mathbf{h}^\star, f^\star, \mathbf{h}^\star)}_{\equiv c} \\
&\leq a + c.
\end{aligned}$$

The inequality holds because $b \leq 0$, as $\widehat{f}$ and $\widehat{\mathbf{h}}$ are chosen to minimize $\widehat{\mathcal{L}}(\widehat{f}, \widehat{\mathbf{h}}, f^\star, \mathbf{h}^\star)$. Note that in both a and c, the common term $\mathbb{E}\left[\frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N} \ell\left(f^\star \circ \mathbf{h}^\star(\mathbf{Z}_k, \mathbf{X}_{k,i}), \tilde{\tau}_{k,i}\right)\right]$ cancels out, leaving only the difference between the empirical and expected loss with respect to the DR score.

This result bounds the expected prediction error $l(\widehat{f}, \widehat{\mathbf{h}}, f^\star, \mathbf{h}^\star)$ by two terms that are directly related to the empirical loss (terms a and c in the above inequality), for which we can derive high-probability error bounds using Gaussian complexity.

Next, we bound terms a and c using the standard generalization bounds based on Rademacher complexity, we obtain the following bounds for a and c with probability at least $1 - \delta$:

$$a, c \leq \sup_{f \in \mathcal{F}, \mathbf{h} \in \mathcal{H}} \left|\mathcal{L}(f, h, f^\star, \mathbf{h}^\star) - \widehat{\mathcal{L}}(f, h, f^\star, \mathbf{h}^\star)\right| \leq 2\mathfrak{R}_{\mathbf{Z},\mathbf{X}}(\ell(\mathcal{F}(\mathcal{H}))) + 2B\sqrt{\frac{\log(2/\delta)}{KN}},$$

where $\mathcal{F}(\mathcal{H}) = \{(f, \mathbf{h}) : f \in \mathcal{F} \text{ and } \mathbf{h} \in \mathcal{H}\}$, $\mathbf{Z}, \mathbf{X}$ denotes the empirical data of design features and customer covariates, and $\mathfrak{R}_{\mathbf{Z},\mathbf{X}}(\ell(\mathcal{F}(\mathcal{H})))$ is the Rademacher complexity of $\ell(\mathcal{F}(\mathcal{H}))$. The first inequality is a direct consequence of the definition of $\sup$. The second inequality leverages the well-established probabilistic upper bound derived using the Rademacher complexity (Theorem 8 in Bartlett and Mendelson (2002) and Theorem 4.10 in Wainwright (2019)).

Next, using Inequality (11) from the Appendix of Tripuraneni et al. (2020), along with the fact that Rademacher complexity is bounded by Gaussian complexity — specifically,

$$\mathfrak{R}_{\mathbf{Z},\mathbf{X}}(\mathcal{F}(\mathcal{H})) \leq \sqrt{\frac{\pi}{2}} \mathfrak{G}_{\mathbf{Z},\mathbf{X}}(\mathcal{F}(\mathcal{H}))$$

(see Lemma 4 in Bartlett and Mendelson (2002) and p.97 in Ledoux and Talagrand (2013)) — we can bound the Rademacher complexity $\ell(\mathcal{F}(\mathcal{H}))$. as follows:

$$\mathfrak{R}(\ell(\mathcal{F}(\mathcal{H}))) \leq 2L\mathfrak{R}(\mathcal{F}(\mathcal{H})) \leq 4L\mathfrak{G}(\mathcal{F}(\mathcal{H})).$$

Combining the above results together, we can establish the upper bound for the expected prediction error:

$$\begin{aligned} l(\widehat{f}, \widehat{\mathbf{h}}, f^{\star}, \mathbf{h}^{\star}) &\leq a + c \leq 4\mathfrak{R}(\ell(\mathcal{F}(\mathcal{H}))) + 4B\sqrt{\frac{\log(2/\delta)}{KN}} \\ &\leq 16L\mathfrak{G}(\mathcal{F}(\mathcal{H})) + 8B\sqrt{\frac{\log(2/\delta)}{KN}}. \end{aligned} \tag{App-6}$$

Next, we follow Tripuraneni et al. (2020) to further decompose the Gaussian complexity of the composite model class $\mathcal{F}(\mathcal{H})$ into a (weighted) sum of the Gaussian complexities of the representation class $\mathcal{H}$ and the prediction class $\mathcal{F}$. To accomplish this, we draw on the chain rule presented in Theorem 7 of Tripuraneni et al. (2020). However, Tripuraneni et al. (2020) considers a setting in which a separate prediction model is trained for each task, leading to a complexity bound of the form $\mathfrak{G}_N(\mathcal{F}^{\bigotimes K}(\mathcal{H}))$, where each task uses N samples independently. In contrast, our setting involves a shared predictive model trained on a total of KN samples across all tasks, yielding the quantity $\mathfrak{G}(\mathcal{F}(\mathcal{H}))$. As a result, we modify Theorem 7 in Tripuraneni et al. (2020) accordingly to enable the decomposition of $\mathfrak{G}(\mathcal{F}(\mathcal{H}))$ under this unified modeling framework.

Claim. *(Modification of Theorem 7 in Tripuraneni et al. (2020))*

$$\mathfrak{G}(\mathcal{F}(\mathcal{H})) \le \frac{4D}{(KN)^2} + 64\left[L\,\mathcal{G}(\mathcal{H}) + \overline{\mathcal{G}}(\mathcal{F})\right]\log\left(KN\right). \tag{App-7}$$

Proof. Note that the proof is largely the same as Theorem 7 in Tripuraneni et al. (2020), with the only difference from bounding the coverging number of the composite model class $\mathcal{F}(\mathcal{H})$. To see this, first, by definition, the Gaussian complexity of the composite function class $\mathcal{F}(\mathcal{H})$ is given by

$$\mathfrak{G}(\mathcal{F}(\mathcal{H})) = \mathbb{E}\left[\frac{1}{\sqrt{KN}}\sup_{f(\mathbf{h})\in\mathcal{F}(\mathcal{H})} Z_{f(\mathbf{h})}\right],$$

where $Z_{f(\mathbf{h})} = \frac{1}{\sqrt{KN}}\sum_{k=1}^{K}\sum_{i=1}^{N} g_{k,i} f(\mathbf{h}(\mathbf{Z}_k, \mathbf{X}_{k,i}))$ and $g_{k,i} \sim \mathcal{N}(0,1)$ are i.i.d. standard normal random variables.

Following the approach on page 18 of Tripuraneni et al. (2020), we can derive a Dudley entropy integral upper bound on this Gaussian complexity:

$$\mathbb{E}\left[\sup_{f(\mathbf{h})\in\mathcal{F}(\mathcal{H})} Z_{f(\mathbf{h})}\right] \le 4\mathbb{E}\left[\sup_{d_2(f(\mathbf{h}),f'(\mathbf{h}'))<\kappa} Z_{f(\mathbf{h})} - Z_{f'(\mathbf{h}')}\right] + 32\int_{\kappa}^{D}\sqrt{\log N(u,\mathcal{F}(\mathcal{H}),d_2)}du,$$

where $d_2(f(\mathbf{h}), f'(\mathbf{h}')) = \sqrt{\frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N}\left[f(\mathbf{h}(\mathbf{Z}_k,\mathbf{X}_{k,i}) - f'(\mathbf{h}'(\mathbf{Z}_k,\mathbf{X}_{k,i})\right]^2}$ is the empirical ℓ_2 norm between two functions, and $N(u, \mathcal{F}(\mathcal{H}), d_2)$ denotes the u-covering number of the composite class $\mathcal{F}(\mathcal{H})$ with respect to the metric d_2 (see Definition 5.1 in Wainwright (2019)).

As shown on page 18 of Tripuraneni et al. (2020), the first term is bounded by $\sqrt{KN}\kappa$. Therefore, our objective is to adapt the derivation from that proof to obtain a bound on the covering number of $\mathcal{F}(\mathcal{H})$ as an additive combination of the covering numbers of $(\mathcal{H}, d_2)$ and $(\mathcal{F}, d_2)$. Specifically, by applying the final inequality on page 18 of Tripuraneni et al. (2020) with $t = 1$ — corresponding to our setting with a single unified prediction model — we have:

$$\log N\left(\epsilon_1\cdot L + \varepsilon_2, \mathcal{F}(\mathcal{H}), d_2\right) \le \log N\left(\epsilon_1, \mathcal{H}, d_2\right) + \max_{(\boldsymbol{\zeta},\boldsymbol{\chi})\in\text{Range}(\mathcal{H})}\log N_2\left(\epsilon_2, \mathcal{F}, d_2\right),$$

where L is the Lipschitz constant of the model class $\mathcal{F}$.

Following the same reasoning as on page 19 of Tripuraneni et al. (2020), we can apply the Sudakov minoration theorem (Theorem 5.30 in Wainwright (2019)), which provides an upper bound on each covering number in terms of the corresponding Gaussian complexity:

$$\log N\left(\epsilon_1, \mathcal{H}, d_2\right) \leq 4\left(\frac{\sqrt{KN}\,\widehat{\mathcal{G}}(\mathcal{H})}{\epsilon_1}\right)^2, \quad \log N_2\left(\epsilon_2, \mathcal{F}, d_2\right) \leq 4\left(\frac{\sqrt{KN}\,\widehat{\mathcal{G}}(\mathcal{F})}{\epsilon_2}\right)^2.$$

Then, following the derivation on page 19 of Tripuraneni et al. (2020) by taking $\epsilon_1 = \frac{\epsilon}{2L}$, $\epsilon_2 = \frac{\epsilon}{2}$, and $\kappa = \frac{D}{(KN)^2}$,we can bound the integrated covering number as follows:

$$\int_{\kappa}^{D} \sqrt{\log N(u, \mathcal{F}(\mathcal{H}), d_2)}du \leq 2\sqrt{KN}\left[L\,\widehat{\mathcal{G}}(\mathcal{H}) + \max_{(\boldsymbol{\zeta},\boldsymbol{\chi})\in\text{Range}(\mathcal{H})} \widehat{\mathcal{G}}(\mathcal{F})\right]\ \log\left(KN\right).$$

Then, substituting every bounds into Inequality (App-7) and by the definition of the Gaussian complexity, we have

$$\begin{aligned}\mathfrak{G}(\mathcal{F}(\mathcal{H})) &\leq \frac{1}{\sqrt{KN}}\left(4\sqrt{KN}\cdot\frac{D}{(KN)^2} + 64\sqrt{KN}\left[L\,\widehat{\mathcal{G}}(\mathcal{H}) + \max_{(\boldsymbol{\zeta},\boldsymbol{\chi})\in\text{Range}(\mathcal{H})} \widehat{\mathcal{G}}(\mathcal{F})\right]\log\left(KN\right)\right)\\ &= \frac{4D}{(KN)^2} + 64\left[L\,\widehat{\mathcal{G}}(\mathcal{H}) + \max_{(\boldsymbol{\zeta},\boldsymbol{\chi})\in\text{Range}(\mathcal{H})} \widehat{\mathcal{G}}(\mathcal{F})\right]\log\left(KN\right),\end{aligned}$$

where Range$(\mathcal{H}) = \bigcup_{h\in\mathcal{H}} \text{Range}(h(\mathbf{Z},\mathbf{X}))$ denotes the set of all possible representations produced by the representation function class $\mathcal{H}$, given all the observed design features and customer covariates in the experiment data. Finally, taking the expectation on both sides of the above inequality yields the result stated in the claim. ■

Finally, combining Inequality (App-6) and Inequality (App-7), we can prove the theorem:

$$\text{Error}\left(\widehat{f}, \widehat{\mathbf{h}}\right) \leq \underbrace{\frac{C_1}{(KN)^2} + \log(KN)\left[C_2\mathfrak{G}(\mathcal{H}) + C_3\overline{\mathfrak{G}}(\mathcal{F})\right]}_{\text{Overfitting Potential for the Model Classes}} + C_4\sqrt{\frac{\log(2/\delta)}{KN}},$$

with probability at least $(1-\delta)$. ■

Theorem App-1 shows that the prediction error bound depends on the Gaussian complexity of (i) the representation function class ($\mathcal{H}$) and (ii) the prediction function class ($\mathcal{F}$). To ensure that the estimated CIF function is asymptotically correct with high probability, it is therefore necessary that the Gaussian complexity of both classes decreases sufficiently fast. This result also yields an important design implication: prediction models must strike a balance between expressiveness and complexity. On the one hand, the model class should be sufficiently flexible to approximate the true function well; on the other hand, its Gaussian complexity must be controlled to achieve tighter prediction error bounds.

While our proof builds on the framework and techniques of Tripuraneni et al. (2020), we introduce two key modifications. First, their framework assumes a standard supervised learning setting with directly observed prediction targets, whereas our setting relies on a noisy proxy for the true treatment effect (the DR score). Second, whereas Tripuraneni et al. (2020) analyze a multi-task learning environment in which each task has its own ground-truth function and separate model, we instead study a single, unified CATE function that generalizes across all offers. This structural distinction requires adjustments to the Gaussian complexity analysis. Among these modifications, the first is more consequential, as it establishes the connection between treatment effect prediction and the broader body of theoretical results on Rademacher and Gaussian complexity for bounding generalization error. The second modification is more technical in nature.

Web Appendix B.4 Prediction Error Bound for DNN-based CIF Models

Since our second-stage model primarily relies on deep neural networks (DNNs), we derive the prediction error bound using existing Gaussian complexity analyses. Following the approach of Tripuraneni et al. (2020), we establish a prediction error bound for the class of deep learning models represented by depth-d vector-valued neural networks as the second-stage estimator. Formally, a neural network is specified as:

$$\mathbf{NN}(\mathbf{x}) = \mathbf{W}_d \, \sigma_d \left(\mathbf{W}_{d-1} \, \sigma_{d-1} \left(\cdots \mathbf{W}_1 \, \sigma_1 \left(\mathbf{x} \right) \right) \right),$$

where $\sigma_1, \cdots, \sigma_d$ represent 1-Lipschitz activation functions that are zero at the origin, such as the linear and ReLU functions. Here, we define $M(j)$ as the maximum possible value of the

Frobenius norm for the matrix $\mathbf{W}_j$, that is, the square root of the sum of the squares of all elements in $\mathbf{W}_j$.

Corollary App-2 (Prediction Error Bound of DNN-based CIF Estimator) *Consider the case when $\mathcal{H}$ is the class of d_1-layer neural networks with 1-Lipschitz activation functions and w-dimensional outputs, and $\mathcal{F}$ denote a class of neural networks with d_2 layers, also employing 1-Lipschitz activation functions. Furthermore, it is assumed that $M(j) < \infty$ for all parameter matrices. Then, with probability at least $1-\delta$, the prediction error of the IRL model $(\widehat{f}, \widehat{\mathbf{h}})$ derived from* (App-3) *can be bounded as follows:*

$$\text{Error}\left(\widehat{f}, \widehat{\mathbf{h}}\right) \leq \frac{C_1}{(KN)^2} + C_2\sqrt{\frac{\log(2/\delta)}{KN}} + \\ \log(KN)\left[C_3\frac{\sqrt{\log(KN)}w\prod_{j=1}^{d_1} M(j)}{\sqrt{KN}} + C_4(w)\frac{\sqrt{\log(KN)}\prod_{j=1}^{d_2} M(j)}{\sqrt{KN}}\right],$$

where C_1, C_2, C_3 and $C_4(w)$ are some constants. Note that $C_4(w) = \mathcal{O}(\sqrt{w})$.

Proof. By Theorem 2 in Golowich et al. (2018), we can bound the sample Rademacher complexity of deep neural networks $\mathcal{NN}$ with w-dimensional output and depth d as follows:

$$\widehat{\mathfrak{R}}(\mathcal{NN}) \leq \frac{2w\sqrt{d+1+\log(r)}\prod_{j=1}^{j} M(j)}{\sqrt{KN}}, \tag{App-8}$$

where $M(j)$ is the maximum possible value of the Frobenius norm for the weight matrix in the j-th layer, and r is the dimension of the input variables.

Using the fact that $\widehat{\mathfrak{G}}(\mathcal{NN}) \leq 2\sqrt{\log(KN)}\ \widehat{\mathfrak{R}}(\mathcal{NN})$ (page 97 in Ledoux and Talagrand (2013)) together with (App-8) and taking expectation on the left side, we have

$$\mathfrak{G}(\mathcal{NN}) \leq \frac{4\sqrt{\log(KN)}w\sqrt{d+1+\log(r)}\prod_{j=1}^{j} M(j)}{\sqrt{KN}}.$$

As a result, for the function classes $\mathcal{H}$ and $\mathcal{F}$ specified in the corollary, we have the following bounds for their Gaussian complexities:

$$\mathfrak{G}_{\mathbf{Z},\mathbf{X}}(\mathcal{H}) \leq A\frac{\sqrt{\log(KN)}w\prod_{j=1}^{j} M(j)}{\sqrt{KN}}, \overline{\mathfrak{G}}(\mathcal{F}) \leq B\sqrt{d+1+\log(w)}\frac{\sqrt{\log(KN)}\prod_{j=1}^{j} M(j)}{\sqrt{KN}}.$$

Finally, by applying the above bounds of Gaussian complexities on Theorem App-1, we can prove the corollary. ■

Corollary App-2 yields several key insights into the use of DNNs as the second-stage model. First, it provides a theoretical guarantee that the upper bound of the prediction error for a DNN-based model converges to zero as the number of experiments grows. Second, it demonstrates the value of dimensionality reduction in representation learning: the prediction error bound tightens as the representation width w decreases, provided that a network of dimension w remains sufficiently expressive to recover the true representation function $\mathbf{h}^{\star}$.

Third, the corollary demonstrates the advantages of employing separate network structures for design and covariate representations. By construction, such architectures impose structural zeros in the parameter matrices $\mathbf{W}_j$ —specifically, those parameters linking offer representations to customer representations—thereby reducing the maximum possible Frobenius norm of $\mathbf{W}_j$ relative to a fully connected network of comparable width and depth. This reduction yields a tighter prediction error bound compared to fully connected architectures that allow unrestricted interactions between design features and customer covariates. A similar logic applies when comparing shared-parameter prediction networks with offer-specific prediction networks, where the former typically attain more favorable error bounds under analogous conditions.

Web Appendix C Characterizing Generalization Error for New Data

In this appendix, we extend the discussion in Section 3.4 by theoretically characterizing the generalization error that arises when models trained on past experiments are applied to new settings, following the theoretical framework proposed by Ben-David et al. (2010). Let the data-generating process for design features and customer covariates in past experiments be denoted by $\mathcal{D}_{\text{past}}$, with the corresponding true CIF represented by τ_{past}. The goal is to predict treatment effects for a new set of offers and customers, generated from a different process $\mathcal{D}_{\text{new}}$, with the true CIF denoted by τ_{new}. To analyze this problem, we draw on classical results from domain adaptation theory (Ben-David et al. 2010) to identify the fundamental factors that determine the generalization performance of any CIF model estimated on past experimental data when applied to new business settings.

Theorem App-2 (Generalization Error) *Let $\widehat{\tau}$ be a CIF model trained on past experiments by the empirical risk minimization problem:*

$$\min_{\widehat{\tau}} \; \frac{1}{\sum_{k=1}^{K} |\#\mathcal{C}_k^E|} \; \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k^E} \ell\left(\widehat{\tau}(\mathbf{Z}_k, \mathbf{X}_{k,i}), \tilde{\tau}_{k,i}\right),$$

where $\tilde{\tau}_{k,i}$ is an unbiased estimate of $\tau_{\text{past}}(\mathbf{Z}_k, \mathbf{X}_{k,i})$, ℓ is the squared loss function, and $|\#\mathcal{C}_k^E|$ denotes the number of observations in $\mathcal{C}_k^E$. Also, assume that $\widehat{\tau}$, τ_{past}, and τ_{new} are bounded. Then, the expected loss for the new offers can be bounded as follows:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] \leq \quad & \underbrace{\mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right]}_{\textit{Prediction Error for Past Experiments}} \quad + \; C \underbrace{\delta(\mathcal{D}_{\text{past}}, \mathcal{D}_{\text{new}})}_{\textit{Attribute Shift}} \\ & + \underbrace{\min\left\{\mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\tau_{\text{past}}, \tau_{\text{new}}\right)\right], \mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\tau_{\text{past}}, \tau_{\text{new}}\right)\right]\right\}}_{\textit{Concept Shift}},\end{aligned}$$

where C is a constant, $\delta(\mathcal{D}_{\text{past}}, \mathcal{D}_{\text{new}})$ is the total variation distance between the two data generating distributions, i.e., $\delta(\mathcal{D}_{\text{past}}, \mathcal{D}_{\text{new}}) = \sup_{B \in \mathcal{B}} \left|\mathbb{P}_{\mathcal{D}_{\text{past}}}(B) - \mathbb{P}_{\mathcal{D}_{\text{new}}}(B)\right|$, and $\mathcal{B}$ denotes the collection of all Borel-measurable sets applicable to $\mathcal{D}_{\text{past}}$ and $\mathcal{D}_{\text{new}}$. For simplicity, $(\mathbf{Z}_k, \mathbf{X}_{k,i})$ are omitted from each function's notation.

Proof. To prove the theorem, we first note that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] = {} & \mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] + \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right] - \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right] \\ & \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] - \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] \\ \leq {} & \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right] + \underbrace{\left|\mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] - \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right]\right|}_{=a} + \\ & \underbrace{\left|\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{new}}\right)\right] - \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau}, \tau_{\text{past}}\right)\right]\right|}_{=b}.\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau},\tau_{\text{new}}\right)\right] = \mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau},\tau_{\text{new}}\right)\right] & \\
+\ \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau},\tau_{\text{past}}\right)\right] - \ \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau},\tau_{\text{past}}\right)\right] & \\
+\ \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau},\tau_{\text{new}}\right)\right] - \ \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau},\tau_{\text{new}}\right)\right] & \\
\leq \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau},\tau_{\text{past}}\right)\right] & \\
+\ \underbrace{\left|\mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau},\tau_{\text{new}}\right)\right] - \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau},\tau_{\text{past}}\right)\right]\right|}_{=a} & \\
+\ \underbrace{\left|\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau},\tau_{\text{new}}\right)\right] - \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau},\tau_{\text{new}}\right)\right]\right|}_{=b}. &
\end{aligned}
$$

For the first term, by the reverse triangle inequality— $|d(x,y) - d(x,z)| \leq d(y,z)$ for any valid distance metric d—we have $a \leq \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\tau_{\text{past}},\tau_{\text{new}}\right)\right]$.

For the second term, we have

$$
b \leq \int \left|\rho_{\mathcal{D}_{\text{new}}}(\mathbf{z},\mathbf{x}) - \rho_{\mathcal{D}_{\text{past}}}(\mathbf{z},\mathbf{x})\right| \ell\left(\widehat{\tau},\tau_{\text{new}}\right) \leq C\delta(\mathcal{D}_{\text{past}},\mathcal{D}_{\text{new}}),
$$

where $\rho_{\mathcal{D}_{\text{new}}}$ and $\rho_{\mathcal{D}_{\text{past}}}$ are the density functions of the two data generating processes. Note that the term $\ell\left(\widehat{\tau},\tau_{\text{new}}\right)$ can be bounded by C as $\widehat{\tau},\tau_{\text{new}}$ are boubded by assumption. Therefore, we have

$$
\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau},\tau_{\text{new}}\right)\right] \leq \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau},\tau_{\text{past}}\right)\right] + \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\tau_{\text{past}},\tau_{\text{new}}\right)\right] + C\delta(\mathcal{D}_{\text{past}},\mathcal{D}_{\text{new}}).
$$

Similarly, if we add and subtract $\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau},\tau_{\text{new}}\right)\right]$ instead of $\mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau},\tau_{\text{new}}\right)\right]$, we have

$$
\mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\widehat{\tau},\tau_{\text{new}}\right)\right] \leq \mathbb{E}_{\mathcal{D}_{\text{past}}}\left[\ell\left(\widehat{\tau},\tau_{\text{past}}\right)\right] + \mathbb{E}_{\mathcal{D}_{\text{new}}}\left[\ell\left(\tau_{\text{past}},\tau_{\text{new}}\right)\right] + C\delta(\mathcal{D}_{\text{past}},\mathcal{D}_{\text{new}}).
$$

Combining these results together proves the theorem. ■

Theorem App-2 highlights three key factors that determine the upper bound of generalization error when extending to new settings: (i) the in-distribution generalization error of the model on past experiments, assuming identical offers, customer profiles, and consistent customer behaviors; (ii) distributional differences in design features and customer covariates between past and new offers (i.e., *attribute shifts*); and (iii) discrepancies between the true CIF functions in past and new settings (i.e., *concept shifts*).

Theorem App-2 has several implications for practice. First, even if a model achieves strong performance on past experimental data, its predictive accuracy on new offers may deteriorate if the distribution of design features or customer covariates differs substantially. Second, while attribute shifts can often be mitigated through careful experiment design, concept shifts represent a more fundamental challenge: if customer responsiveness to promotions changes in ways not captured by past experiments, prediction error cannot be fully eliminated without collecting new experimental data. Therefore, before deploying models to new settings, companies should evaluate the extent of such shifts, such as the diagnostic analysis we conduct in Section 8, to assess the reliability of extrapolating past models to future offers.

Web Appendix D Randomization and Interference Checks

Web Appendix D.1 Randomization Check

We perform covariate balance checks across 362 experiments by calculating the standardized mean difference (SMD), which is the absolute difference in mean covariate values between the treatment and control groups, normalized by the pooled within-group standard deviation. The distribution of SMDs across the experiments is detailed in Table App-1. Notably, out of 14,480 (experiment, covariate) pairs, only 0.3% exhibit SMDs greater than 0.2, and 1.4% show SMDs exceeding 0.1. These results suggest that there are no meaningful differences in covariate means between the treatment and control groups across experiments.

We further test for distributional differences in each covariate between the treatment and control groups using the Kolmogorov–Smirnov test, conducted separately for each experiment. The distribution of the resulting p-values is summarized in Table App-2. Across 14,480 (experiment, covariate) pairs, only 1.8% yield p-values below 0.01 and 4.9% below 0.05, indicating no systematic evidence of imbalance in covariate distributions between treatment and control groups. These results support the validity of the randomization procedure across experiments.

Table App-1: Standardized Mean Differences of Customer Covariates

Variable	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
receipt_count	0.00	0.02	-0.13	-0.01	0.01	0.06
aov	0.00	0.01	-0.06	-0.01	0.01	0.04
item_per_order	0.00	0.01	-0.07	0.00	0.01	0.04
has_past_7	0.00	0.01	-0.06	0.00	0.01	0.06
has_past_30	0.00	0.01	-0.05	-0.01	0.01	0.08
has_past_60	0.00	0.01	-0.04	0.00	0.01	0.07
offer_receipt_count	0.03	0.06	-0.32	0.00	0.04	0.33
offer_aov	0.03	0.06	-0.18	0.00	0.04	0.33
offer_item_per_order	0.03	0.07	-0.18	0.00	0.05	0.42
offer_has_past_7	0.02	0.04	-0.13	0.00	0.03	0.15
offer_has_past_30	0.03	0.06	-0.15	0.00	0.04	0.49
offer_has_past_60	0.04	0.08	-0.16	0.00	0.05	0.56
7-ELEVEN	0.00	0.01	-0.07	-0.01	0.01	0.05
AMAZON	0.00	0.02	-0.07	-0.01	0.01	0.06
COSTCO	0.00	0.01	-0.05	-0.01	0.00	0.03
CVS	0.00	0.02	-0.10	-0.01	0.01	0.05
KROGER	0.00	0.01	-0.04	-0.01	0.01	0.06
MCDONALD'S	0.00	0.01	-0.06	-0.01	0.01	0.07
TARGET	0.00	0.01	-0.05	-0.01	0.01	0.07
THE HOME DEPOT	0.00	0.02	-0.08	-0.01	0.01	0.05
WALGREENS	0.00	0.01	-0.05	-0.01	0.01	0.06
WALMART	0.00	0.01	-0.05	-0.01	0.01	0.06
Appetizers & Sides	0.00	0.01	-0.08	-0.01	0.01	0.05
Bakery & Bread	0.00	0.01	-0.05	-0.01	0.01	0.04
Bath & Body	0.00	0.01	-0.05	-0.01	0.01	0.05
Beer, Wine & Spirits	0.00	0.01	-0.07	-0.01	0.01	0.04
Beverages	0.00	0.01	-0.05	-0.01	0.01	0.05
Candy	0.00	0.01	-0.07	-0.01	0.01	0.04
Cat	0.00	0.02	-0.06	0.00	0.01	0.05
Dairy	0.00	0.01	-0.06	-0.01	0.01	0.04
Dog	0.00	0.02	-0.08	-0.01	0.01	0.05
Hair Care	0.00	0.01	-0.06	0.00	0.01	0.05
Household Paper & Plastic	0.00	0.01	-0.05	-0.01	0.01	0.04
Produce	0.00	0.01	-0.05	-0.01	0.01	0.04
is_referral	0.00	0.01	-0.05	-0.01	0.01	0.08
age	0.00	0.02	-0.06	-0.01	0.01	0.05
tenure	0.00	0.02	-0.08	-0.01	0.01	0.10
is_male	0.00	0.02	-0.05	-0.01	0.01	0.05
is_female	0.00	0.02	-0.05	-0.01	0.01	0.06
is_non_binary?	0.00	0.02	-0.06	-0.01	0.01	0.06

Table App-2: P-values of the Kolmogorov–Smirnov Test for Distribution Difference

Variable	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
receipt_count	0.62	0.32	0.03	0.34	0.93	1.00
aov	0.57	0.32	0.02	0.30	0.86	1.00
item_per_order	0.57	0.31	0.00	0.33	0.87	1.00
has_past_7	0.93	0.18	0.00	0.98	1.00	1.00
has_past_30	0.97	0.10	0.07	1.00	1.00	1.00
has_past_60	0.99	0.07	0.14	1.00	1.00	1.00
offer_receipt_count	0.71	0.43	0.03	0.16	1.00	1.00
offer_aov	0.67	0.43	0.01	0.06	1.00	1.00
offer_item_per_order	0.71	0.43	0.02	0.18	1.00	1.00
offer_has_past_7	0.80	0.37	0.03	0.98	1.00	1.00
offer_has_past_30	0.75	0.41	0.02	0.52	1.00	1.00
offer_has_past_60	0.73	0.43	0.02	0.24	1.00	1.00
7-ELEVEN	0.97	0.11	0.13	1.00	1.00	1.00
AMAZON	0.82	0.26	0.01	0.73	1.00	1.00
COSTCO	0.85	0.23	0.01	0.77	1.00	1.00
CVS	0.82	0.26	0.04	0.72	1.00	1.00
KROGER	0.91	0.18	0.01	0.94	1.00	1.00
MCDONALD’S	0.77	0.28	0.02	0.61	1.00	1.00
TARGET	0.69	0.33	0.00	0.40	0.99	1.00
THE HOME DEPOT	0.88	0.20	0.03	0.86	1.00	1.00
WALGREENS	0.80	0.26	0.00	0.69	1.00	1.00
WALMART	0.61	0.33	0.00	0.30	0.95	1.00
Appetizers & Sides	0.80	0.27	0.00	0.65	1.00	1.00
Bakery & Bread	0.67	0.32	0.00	0.41	0.98	1.00
Bath & Body	0.76	0.30	0.01	0.57	1.00	1.00
Beer, Wine & Spirits	0.74	0.30	0.00	0.51	1.00	1.00
Beverages	0.63	0.32	0.00	0.35	0.92	1.00
Candy	0.67	0.32	0.00	0.42	0.98	1.00
Cat	0.86	0.23	0.04	0.83	1.00	1.00
Dairy	0.63	0.32	0.00	0.35	0.96	1.00
Dog	0.80	0.25	0.01	0.66	1.00	1.00
Hair Care	0.75	0.30	0.00	0.54	1.00	1.00
Household Paper & Plastic	0.69	0.32	0.01	0.44	0.99	1.00
Produce	0.65	0.32	0.00	0.38	0.97	1.00
is_referral	0.90	0.22	0.03	0.95	1.00	1.00
age	0.59	0.31	0.00	0.33	0.89	1.00
tenure	0.47	0.31	0.00	0.17	0.74	1.00
is_male	0.96	0.12	0.16	1.00	1.00	1.00
is_female	0.93	0.16	0.02	0.97	1.00	1.00
is_non_binary?	1.00	0.02	0.76	1.00	1.00	1.00

Web Appendix D.2 Testing the No Interference Assumption

This appendix provides empirical evidence that interference from multiple offers is minimal. Because the number of concurrent offers a customer receives is correlated with purchasing behavior (likely due to specific eligibility criteria set by the platform), we apply the double machine learning framework (Chernozhukov et al. 2018) to control for this selection effect. Specifically, we test whether the number of additional offers a customer receives causally influences the estimated treatment effects of the focal offer.

For every offer, we first calculate the DR score $\tilde{\tau}_{k,i}$ using the approach described in Web Appendix E.1. We then apply the double machine learning approach to estimate the following partially linear models:

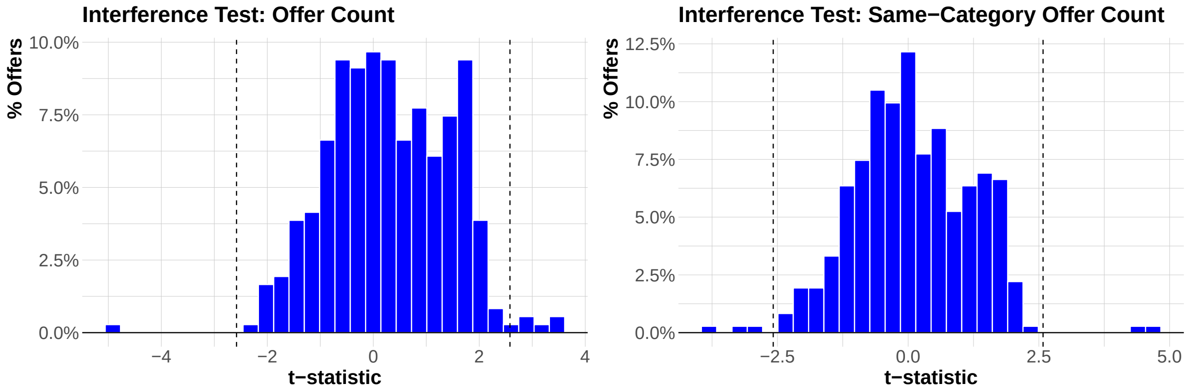
$$\tilde{\tau}_{k,i} = \beta_k^{\text{Total}} N_{k,i}^{\text{Total}} + m^{\text{Total}}(\mathbf{X}_{k,i}) + \varepsilon_{k,i}^{\text{Total}}, \qquad N_{k,i}^{\text{Total}} = g^{\text{Total}}(\mathbf{X}_{k,i}) + \eta_{k,i}^{\text{Total}},$$

$$\tilde{\tau}_{k,i} = \beta_k^{\text{Category}} N_{k,i}^{\text{Category}} + m^{\text{Category}}(\mathbf{X}_{k,i}) + \varepsilon_{k,i}^{\text{Category}}, \qquad N_{k,i}^{\text{Category}} = g^{\text{Category}}(\mathbf{X}_{k,i}) + \eta_{k,i}^{\text{Category}},$$

where $N_{k,i}^{\text{Total}}$ denotes the total number of active offers available to customer i when offer k was active, and $N_{k,i}^{\text{Category}}$ denotes the number of active offers promoting the same category available to customer i at that time. Here, we use XGBoost with default settings as implemented in the `DoubleML` package (Bach et al. 2022) with ten-fold cross-fitting to estimate the control functions m and g. In the absence of interference among different offers, we would expect that $\beta_k^{\text{Total}} = 0$ and $\beta_k^{\text{Category}} = 0$ for most offers.

Figure App-1 shows the distribution of t-statistics for the estimated β_k^{Total} and $\beta_k^{\text{Category}}$ across 362 offers. Notably, (i) only 1.6% of offers have β_k^{Total} exceeding the critical value at the 1% significance level, and 5.2% exceed the threshold at the 5% level; and (ii) only 1.4% of offers have $\beta_k^{\text{Category}}$ exceeding the 1% threshold, with 4.7% exceeding the 5% threshold. These low rejection rates suggest that there is no evidence of interference in our empirical context.

Figure App-1: t-statistics for Interference Test across 362 Offers



Note. The dashed line indicates the critical value for rejecting the null hypothesis that the coefficient is zero at the 1% significance level.

Web Appendix E Model Implementation

Web Appendix E.1 Construction of DR Score

To construct the DR score for model calibration in Section 6, we use observations from the training set to estimate the conditional mean outcome models ($\widehat{\mu}_{k,1}^{[-i]}$ and $\widehat{\mu}_{k,0}^{[-i]}$) using five-fold cross-fitting. Because treatment assignment is random, we set the propensity score for each offer equal to the proportion of observations in the training set that received the offer. We then substitute these estimated nuisance components into Equation 1 to obtain the final DR score.

To estimate the nuisance conditional mean outcome models for each offer, we consider three model classes: elastic net regression, random forests, and XGBoost. For each class, we conduct five-fold cross-validation over a predefined grid of hyperparameters:

1. ELASTIC NET: Implemented using the `glmnet` package in R. The two main hyperparameters are α, which controls the mixing between L1 and L2 penalties, and λ, the regularization strength. We select $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ and $\lambda \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$.

2. RANDOM FORESTS: Implemented using the `ranger package` in R. We tune the number of trees and the maximum depth of each tree. The number of trees is selected from $\{100, 250, 500, 750, 1000\}$, and maximum depth from $\{2, 5, 10, 15, 20\}$.

3. XGBOOST: Implemented using the `xgboost` package in R. We tune the maximum depth of a tree from $\{2, 5, 10, 15, 20\}$, the learning rate $\eta \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, and the number of boosting rounds from $\{20, 40, 60, 80, 100\}$.

Table App-3 reports the mean squared error (MSE) of the best-performing model within each model class. Overall, both random forest and XGBoost outperform the elastic net, with random forest and XGBoost exhibiting comparable levels of accuracy. For all results in this paper, we adopt the XGBoost model with hyperparameters (depth = 5, $\eta = 0.2$, number of iterations = 40), as it achieves the lowest MSE.

Table App-3: Mean Squared Error of the Best-Performing Model in Each Class

Model Class	Treatment	Control
ELASTIC NET ($\alpha = 0.4$, $\lambda = 0.8$)	8651.5	8601.9
RANDOM FOREST (Depth = 10, Number of trees = 250)	8550.1	8543.5
XGBOOST (Depth = 5, $\eta = 0.2$, number of iterations = 40)	8544.2	8536.9

Web Appendix E.2 Model Specification Deep Learning Models

As described in Section 5.1.2, we implement two three-layer fully connected neural networks for representation learning and one three-layer fully connected neural network for outcome prediction. Each hidden layer contains 10 nodes. When the representation dimensions are set to $K_1 = K_2 = 10$, the overall architecture comprises 1,501 trainable parameters.

For the other two joint models in Section 5.2 (i.e., the standard fully connected neural networks architecture with and without design features), we employ an six-layer fully-connected neural network. The first hidden layer consists of 20 nodes, while the subsequent layers each have 10 nodes. This configuration results in a total of 1,441 parameters for the standard DNN with design features and 881 parameters for the standard DNN without design features.

For the joint models in Section 5.2 — that is, the standard fully connected neural network architectures with and without design features—we employ a six-layer fully connected neural network. The first hidden layer contains 20 nodes, and each subsequent layer contains 10 nodes. This configuration yields a total of 1,441 parameters for the standard DNN with design features and 881 parameters for the version without design features. For the individual-modeling approach, we apply the same neural network architecture as the standard DNN without design features to each offer separately.

Across all neural network models, we use ReLU activation functions for the hidden units and a linear activation function for the output layer. For calibrating the parameterss, we employ mini-batch stochastic gradient descent with a batch size of 30 across all architectures. The learning rate is adaptively adjusted using the Ada-delta algorithm (Zeiler 2012), as implemented in Keras (Chollet et al. 2015). Each network is trained for 10 epochs, which we selected based on the observation that validation performance stabilizes within this training horizon.

Web Appendix F Robustness Check

Web Appendix F.1 Additional Benchmark Models

In this appendix, we examine alternative model classes for CIF estimation and compare their performance with that of our proposed two-stage model with the DRL architecture.

Web Appendix F.1.1 Model Classes

We consider the following alternative model classes as benchmarks.

Alternative Model Classes for the Second Stage Models. We first evaluate alternative model classes within our proposed CIF estimation framework, using them as the second-stage model to predict the DR scores constructed in the first stage. For hyperparameter tuning, we perform five-fold cross-validation to identify the hyper-parameters that minimize the MSE in approximating the first-stage DR scores.

1. ELASTIC NET: Implemented using the `h2o` package in R. The two key hyperparameters are α, which determines the mixing ratio between L1 and L2 penalties, and λ, which controls the overall regularization strength. We consider $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ and $\lambda \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. The optimal values are $\alpha = 0.4$ and $\lambda = 0.6$.

2. XGBOOST: Implemented using the `xgboost` package in R. We tune the maximum tree depth from $\{2, 5, 10, 15, 20\}$, the learning rate $\eta \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, and the number of boosting rounds from $\{20, 40, 60, 80, 100\}$. The optimal configuration is a tree depth of 5, learning rate $\eta = 0.2$, and 40 boosting iterations.

Joint T-learner Models. We also adapt the popular T-learner approach for CATE estimation (Künzel et al. 2019) to our context. Specifically, we pool all experiments and jointly estimate two conditional mean outcome models as functions of design features and customer covariates, one for observations in the treatment group and the other for those in the control group.

1. STANDARD DNN: We implement a standard DNN with the same architecture as described in Section 6.

2. ELASTIC NET: Implemented using the `h2o` package in R. The two key hyperparameters are α, which determines the mixing ratio between L1 and L2 penalties, and λ, which controls the overall regularization strength. We consider $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ and $\lambda \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$. The optimal hyperparameters are $\alpha = 0.2$ and $\lambda = 0.6$ for the treatment group model, and $\alpha = 0.4$ and $\lambda = 0.6$ for the control group model.

3. XGBOOST: Implemented using the `xgboost` package in R. We tune the maximum tree depth from $\{2, 5, 10, 15, 20\}$, the learning rate $\eta \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, and the number of boosting rounds from $\{20, 40, 60, 80, 100\}$. The optimal configuration is a tree depth of 5, learning rate $\eta = 0.4$, and 40 boosting iterations for the treatment group model, and a tree depth of 5, learning rate $\eta = 0.2$, and 60 boosting iterations for the control group model.

Individual CATE Models. We also evaluate alternative individual CATE models for each experiment using both the DR-Learner and the T-Learner approaches. For the DR-Learner, we implement both ELASTIC NET and XGBOOST using the same hyperparameter tuning grids described above. The optimal hyperparameters for ELASTIC NET are $\alpha = 0.6$ and $\lambda = 0.8$, while the optimal configuration for XGBOOST is a tree depth of 5, learning rate $\eta = 0.2$, and 40 boosting iterations.

For the T-learner, we consider three model classes: STANDARD DNN (without design features), XGBOOST, and ELASTIC NET. For the STANDARD DNN, we use the same model architecture described in Section 6. For ELASTIC NET, the optimal hyperparameters are $\alpha = 0.4$ and $\lambda = 0.8$ for the treatment group model, and $\alpha = 0.6$ and $\lambda = 0.6$ for the control group model. For XGBOOST, the optimal configuration is a tree depth of 5, learning rate $\eta = 0.2$, and 40 boosting iterations for the treatment group model, and a tree depth of 5, learning rate $\eta = 0.4$, and 40 boosting iterations for the control group model.

Web Appendix F.1.2 Results

Targeting for Tested Offers. We first replicate the analysis in Section 6.1 for the models described above. Table App-4 reports the mean AUTOC and normalized profit across 20 training–holdout splits for different treatment effect prediction models. First, we find that the proposed model (Row 1) consistently outperforms all alternatives. Second, and importantly, most joint models (Row 1 to Row 8) except for the Elastic Net outperform their counterparts that model each experiment separately (Row 9 to Row 14). Third, the joint T-learner models (Row 6 to Row 8) also achieve higher treatment prioritization and profitability compared to individually estimated models (Row 12 to Row 14).

Targeting for Untested Offers. Next, we replicate the analysis in Section 6.2 for the models described above. Table App-5 reports the mean AUTOC and normalized profit across 20 training–holdout splits. First, we find that the proposed model (Row 1) consistently outperforms all alternatives. Second, all joint models (Row 1 to Row 8) outperform their counterparts that model each experiment separately (Row 9 to Row 14). Third, the joint T-learner models (Row 6 to Row 8) also achieve higher treatment prioritization and profitability compared to individually estimated models (Row 12 to Row 14).

Together, these results demonstrate that the benefits of leveraging and integrating insights from past experiments are robust across different modeling frameworks.

Table App-4: Targeting Performance on Holdout Observations: Different Treatment Effect Models

Model	AUTOC	% Outperf.	Profit	% Outperf.
Joint Modeling: Two-Stage Estimation Using DR Scores				
Joint: Proposed DRL	0.47 (0.12)	—	1.37 (0.10)	—
Joint: Standard DNN	0.36 (0.11)	85%	1.16 (0.19)	90%
Joint: Standard DNN w/o Design Features	0.37 (0.09)	100%	1.33 (0.13)	85%
Joint: XGBoost	0.41 (0.33)	60%	1.13 (0.10)	100%
Joint: Elastic Net	0.15 (0.10)	90%	0.88 (0.10)	100%
Joint Modeling: T-learner				
Joint T-learner: Standard DNN	0.29 (0.11)	100%	1.25 (0.10)	75%
Joint T-learner: XGBoost	0.37 (0.17)	70%	1.23 (0.21)	75%
Joint T-learner: Elastic Net	0.19 (0.09)	100%	1.31 (0.09)	85%
Individual Modeling: DR-learner				
Individual: Standard DNN	0.11 (0.07)	100%	0.96 (0.07)	100%
Individual: XGBoost	0.18 (0.07)	100%	1.05 (0.12)	100%
Individual: Elastic Net	0.16 (0.17)	80	1.12 (0.09) %	100%
Individual Modeling: T-learner				
Individual T-learner: Standard DNN	0.15 (0.09)	100%	1.04 (0.06)	100%
Individual T-learner: XGBoost	0.14 (0.05)	100%	1.04 (0.14)	100%
Individual T-learner: Elastic Net	0.13 (0.12)	100%	1.09 (0.08)	100%

Note: We report the mean AUTOC and normalized profit across 20 train–holdout splits. Standard deviations are shown in parentheses. The third and fifth columns ("% Outperf.") report the percentage of splits in which the proposed model (first row) outperformed each benchmark in AUTOC and normalized profit, respectively.

Table App-5: Targeting Performance on Holdout Offers: Different Treatment Effect Models

Model	AUTOC	% Outperf.	Profit	% Outperf.
Joint Modeling: Two-Stage Estimation Using DR Scores				
Joint: Proposed DRL	0.47 (0.14)	—	1.19 (0.09)	—
Joint: Standard DNN	0.22 (0.13)	100%	0.97 (0.06)	100%
Joint: Standard DNN w/o Design Features	0.37 (0.11)	75%	1.11 (0.09)	90%
Joint: XGBoost	0.41 (0.24)	65%	1.09 (0.09)	100%
Joint: Elastic Net	0.19 (0.27)	100%	0.91 (0.04)	100%
Joint Modeling: T-learner				
Joint T-learner: Standard DNN	0.30 (0.10)	90%	1.08 (0.1)	100%
Joint T-learner: XGBoost	0.37 (0.10)	80%	1.14 (0.12)	85%
Joint T-learner: Elastic Net	0.17 (0.17)	100%	1.14 (0.09)	90%
Individual Modeling: DR-learner				
Individual: Standard DNN	0.02 (0.08)	100%	0.85 (0.03)	100%
Individual: XGBoost	0.11 (0.04)	100%	0.75 (0.04)	100%
Individual: Elastic Net	0.06 (0.12)	100%	0.82 (0.03)	100%
Individual Modeling: T-learner				
Individual T-learner: Standard DNN	0.04 (0.13)	100%	0.80 (0.03)	100%
Individual T-learner: XGBoost	0.00 (0.06)	100%	0.74 (0.04)	100%
Individual T-learner: Elastic Net	-0.01 (0.08)	100%	0.84 (0.03)	100%

Note: We report the mean AUTOC and normalized profit across 20 train–holdout splits. Standard deviations are shown in parentheses. The third and fifth columns ("% Outperf.") report the percentage of splits in which the proposed model (first row) outperformed each benchmark in AUTOC and normalized profit, respectively.

Web Appendix F.2 Sizes of Neural Network Models

In this appendix, we evaluate the robustness of our results with respect to depth and width of different neural network models.

Web Appendix F.2.1 Depth of Neural Network Models

We test a shallow architecture by removing one hidden layer from both the representation and prediction networks described in Section Web Appendix E.2. We also test a deep architecture by adding one additional hidden layer to each of the two networks. Each hidden layer contains ten nodes. The same architectural modifications are applied to the other deep learning models discussed in Section Web Appendix F.1.

Targeting for Tested Offers. We replicate the analysis in Section 6.1 using DNN-based models with varying neural network depths. Table App-6 reports the AUTOC and normalized profit for DNN-based models. Column 4 shows the proportion of each metric relative to that of the proposed DRL model with the same network depth (i.e., the first row of each subtable). Overall, the results remain consistent across different network depths, with the baseline architecture slightly outperforming both deeper and shallower variants. This finding suggests that the proposed model's performance is robust to different network depth.

Targeting for Untested Offers. Next, we extend the analysis in Section 6.2 to DNN-based models with varying network depths. Table App-7 reports the AUTOC and normalized profit for these models. Column 4 presents each metric as a proportion relative to the proposed DRL model with the same depth (i.e., the first row of each subtable). Overall, the results are consistent across different network depths, with the baseline architecture slightly outperforming both deeper and shallower variants. This finding indicates that the proposed model's performance is robust to variations in network depth.

Table App-6: Targeting Performance on Holdout Observations: Different DNN Depths

Model	AUTOC	% Outperf.	Profit	% Outperf.
Shallow **Architecture**				
Joint: Proposed DRL	0.46 (0.07)	—	1.28 (0.09)	—
Joint: Standard DNN	0.34 (0.09)	100%	1.14 (0.13)	100%
Joint: Standard DNN w/o Design Features	0.34 (0.10)	85%	1.20 (0.08)	90%
Joint T-learner: Standard DNN	0.29 (0.07)	100%	1.24 (0.08)	90%
Individual: Standard DNN	0.14 (0.14)	100%	0.87 (0.05)	100%
Individual T-learner: Standard DNN	0.33 (0.17)	80%	0.90 (0.06)	100%
Baseline **Architecture (Main Analysis in Section 6.1)**				
Joint: Proposed DRL	0.47 (0.12)	—	1.37 (0.10)	—
Joint: Standard DNN	0.36 (0.11)	85%	1.16 (0.19)	90%
Joint: Standard DNN w/o Design Features	0.37 (0.09)	100%	1.33 (0.13)	85%
Joint T-learner: Standard DNN	0.29 (0.11)	100%	1.25 (0.11)	75%
Individual: Standard DNN	0.11 (0.07)	100%	0.96 (0.07)	100%
Individual T-learner: Standard DNN	0.15 (0.09)	100%	1.04 (0.06)	100%
Deep **Architecture**				
Joint: Proposed DRL	0.46 (0.08)	—	1.24 (0.08)	—
Joint: Standard DNN	0.39 (0.12)	90%	1.02 (0.14)	100%
Joint: Standard DNN w/o Design Features	0.41 (0.05)	85%	1.22 (0.08)	80%
Joint T-learner: Standard DNN	0.35 (0.06)	100%	1.24 (0.10)	70%
Individual: Standard DNN	0.15 (0.12)	100%	0.88 (0.05)	100%
Individual T-learner: Standard DNN	0.28 (0.10)	100%	1.08 (0.06)	100%

Note: We report the mean AUTOC and normalized profit across 20 train–holdout splits. Standard deviations are shown in parentheses, and the fourth column presents the percentage of splits in which the proposed DRL architecture (first row) outperformed each benchmark in terms of AUTOC and normalized profit, respectively.

Table App-7: Targeting Performance on Holdout Observations: Different DNN Depths

Model	AUTOC	% Outperf.	Profit	% Outperf.
***Shallow* Architecture**				
Joint: Proposed DRL	0.42 (0.07)	—	1.18 (0.09)	—
Joint: Standard DNN	0.34 (0.09)	75%	1.04 (0.13)	100%
Joint: Standard DNN w/o Design Features	0.34 (0.10)	70%	1.10 (0.08)	85%
Joint T-learner: Standard DNN	0.29 (0.07)	100%	1.14 (0.08)	90%
Individual: Standard DNN	0.14 (0.14)	100%	0.87 (0.05)	100%
Individual T-learner: Standard DNN	0.23 (0.17)	85%	0.90 (0.06)	100%
***Baseline* Architecture (Main Analysis in Section 6.2)**				
Joint: Proposed DRL	0.47 (0.14)	—	1.19 (0.09)	—
Joint: Standard DNN	0.22 (0.13)	100%	0.97 (0.06)	100%
Joint: Standard DNN w/o Design Features	0.37 (0.11)	75%	1.11 (0.09)	90%
Joint T-learner: Standard DNN	0.30 (0.10)	90%	1.08 (0.1)	100%
Individual: Standard DNN	0.02 (0.08)	100%	0.85 (0.03)	100%
Individual T-learner: Standard DNN	0.04 (0.13)	100%	0.80 (0.03)	100%
***Deep* Architecture**				
Joint: Proposed DRL	0.45 (0.09)	—	1.18 (0.05)	—
Joint: Standard DNN	0.34 (0.17)	75%	1.06 (0.13)	90%
Joint: Standard DNN w/o Design Features	0.41 (0.08)	85%	1.16 (0.06)	75%
Joint T-learner: Standard DNN	0.36 (0.07)	80%	1.09 (0.08)	90%
Individual: Standard DNN	-0.01 (0.1)	100%	0.83 (0.04)	100%
Individual T-learner: Standard DNN	-0.03 (0.07)	100%	0.81 (0.04)	100%

Note: We report the mean AUTOC and normalized profit across 20 train–holdout splits. Standard deviations are shown in parentheses. The third and fifth columns (“% Outperf.”) report the percentage of splits in which the proposed model (first row) outperformed each benchmark in AUTOC and normalized profit, respectively.

Web Appendix F.2.2 Representation Dimensionality of Neural Network Models

In this appendix, we evaluate the robustness of the model's targeting performance with respect to variations in both representation dimensionality. To assess robustness to representation dimensionality, we vary the number of dimensions for both the intervention and covariate representations, setting them to 5, 10, 20, 30, and 40, respectively. For the other deep learning models, we adjust the width of the middle hidden layer to approximate these same representation dimensionalities, ensuring comparability with the proposed architecture.

Targeting for Tested Offers. We replicate the analysis in Section 6.1 using DNN-based models with varying representation dimensions. Table App-8 reports the AUTOC and normalized profit for these models across different network widths. Column 4 shows each metric as a proportion relative to the proposed DRL model with the same representation dimension (i.e., the first row of each subtable). Overall, the results are consistent across variations in representation dimensionality, indicating that the proposed model's performance is robust to architectural changes.

Targeting for Untested Offers. We reproduce the analysis from Section 6.2 using DNN-based models with different representation dimensionalities. Table App-9 presents the AUTOC and normalized profit for these models across varying network widths. Column 4 reports each metric as a ratio relative to the proposed DRL model with the corresponding representation dimension (i.e., the first row in each subtable). Overall, the results remain stable across changes in representation size, suggesting that the proposed model's performance is resilient to reasonable architectural variations.

Table App-8: Targeting Performance on Holdout Observations: Varying Dimensionality

Model	AUTOC	% Outperf.	Profit	% Outperf.
Dimension = 5				
Joint: Proposed DRL	0.45 (0.10)	—	1.21 (0.06)	—
Joint: Standard DNN	0.34 (0.18)	75%	1.04 (0.16)	100%
Joint: Standard DNN w/o Design Features	0.33 (0.06)	90%	1.13 (0.08)	80%
Joint T-learner: Standard DNN	0.25 (0.09)	90%	1.13 (0.07)	80%
Individual: Standard DNN	0.07 (0.15)	100%	0.88 (0.05)	100%
Individual T-learner: Standard DNN	0.21 (0.14)	100%	0.93 (0.05)	100%
Dimension = 10 (Main Analysis in Section 6.1)				
Joint: Proposed DRL	0.47 (0.12)	—	1.37 (0.10)	—
Joint: Standard DNN	0.36 (0.11)	85%	1.16 (0.19)	90%
Joint: Standard DNN w/o Design Features	0.37 (0.09)	100%	1.33 (0.13)	85%
Joint T-learner: Standard DNN	0.29 (0.11)	100%	1.25 (0.10)	75%
Individual: Standard DNN	0.11 (0.07)	100%	0.96 (0.07)	100%
Individual T-learner: Standard DNN	0.15 (0.09)	100%	1.04 (0.06)	100%
Dimension = 20				
Joint: Proposed DRL	0.53 (0.32)	—	1.28 (0.05)	—
Joint: Standard DNN	0.41 (0.09)	75%	1.19 (0.14)	90%
Joint: Standard DNN w/o Design Features	0.28 (0.09)	90%	1.23 (0.06)	90%
Joint T-learner: Standard DNN	0.33 (0.07)	90%	1.21 (0.04)	80%
Individual: Standard DNN	0.02 (0.12)	100%	0.94 (0.08)	100%
Individual T-learner: Standard DNN	0.21 (0.14)	100%	0.93 (0.05)	100%
Dimension = 30				
Joint: Proposed DRL	0.44 (0.15)	—	1.32 (0.04)	—
Joint: Standard DNN	0.38 (0.20)	90%	1.11 (0.15)	90%
Joint: Standard DNN w/o Design Features	0.37 (0.05)	90%	1.19 (0.08)	100%
Joint T-learner: Standard DNN	0.34 (0.09)	100%	1.32 (0.05)	70%
Individual: Standard DNN	0.22 (0.08)	100%	0.93 (0.07)	100%
Individual T-learner: Standard DNN	0.38 (0.14)	80%	1.02 (0.04)	100%
Dimension = 40				
Joint: Proposed DRL	0.47 (0.26)	—	1.31 (0.04)	—
Joint: Standard DNN	0.34 (0.20)	90%	1.08 (0.15)	100%
Joint: Standard DNN w/o Design Features	0.38 (0.06)	85%	1.23 (0.09)	100%
Joint T-learner: Standard DNN	0.38 (0.07)	100%	1.29 (0.05)	80%
Individual: Standard DNN	0.20 (0.19)	100%	0.95 (0.06)	100%
Individual T-learner: Standard DNN	0.36 (0.16)	90%	1.02 (0.05)	100%

Note: We report the mean AUTOC and normalized profit across 20 train–holdout splits. Standard deviations are shown in parentheses. The third and fifth columns ("% Outperf.") report the percentage of splits in which the proposed model (first row) outperformed each benchmark in AUTOC and normalized profit, respectively.

Table App-9: Targeting Performance on Holdout Offers: Varying Dimensionality

Model	**AUTOC**	**% Outperf.**	**Profit**	**% Outperf.**
Dimension = 5				
Joint: Proposed DRL	0.48 (0.11)	—	1.20 (0.08)	—
Joint: Standard DNN	0.28 (0.16)	90%	1.06 (0.10)	100%
Joint: Standard DNN w/o Design Features	0.42 (0.09)	85%	1.01 (0.42)	100%
Joint T-learner: Standard DNN	0.37 (0.10)	75%	1.08 (0.06)	90%
Individual: Standard DNN	0.01 (0.10)	100%	0.89 (0.04)	100%
Individual T-learner: Standard DNN	-0.02 (0.06)	100%	0.87 (0.04)	100%
Dimension = 10 (Main Analysis in Section 6.2)				
Joint: Proposed DRL	0.47 (0.14)	—	1.19 (0.09)	—
Joint: Standard DNN	0.22 (0.13)	100%	0.97 (0.06)	100%
Joint: Standard DNN w/o Design Features	0.37 (0.11)	75%	1.11 (0.09)	90%
Joint T-learner: Standard DNN	0.30 (0.10)	90%	1.08 (0.1)	100%
Individual: Standard DNN	0.02 (0.08)	100%	0.85 (0.03)	100%
Individual T-learner: Standard DNN	0.04 (0.13)	100%	0.80 (0.03)	100%
Dimension = 20				
Joint: Proposed DRL	0.47 (0.19)	—	1.21 (0.11)	—
Joint: Standard DNN	0.26 (0.22)	90%	1.01 (0.13)	100%
Joint: Standard DNN w/o Design Features	0.38 (0.14)	85%	1.14 (0.12)	100%
Joint T-learner: Standard DNN	0.34 (0.15)	100%	1.12 (0.10)	100%
Individual: Standard DNN	0.00 (0.06)	100%	0.86 (0.04)	100%
Individual T-learner: Standard DNN	0.05 (0.08)	100%	0.84 (0.04)	100%
Dimension = 30				
Joint: Proposed DRL	0.48 (0.12)	—	1.18 (0.10)	—
Joint: Standard DNN	0.37 (0.20)	90%	1.08 (0.14)	90%
Joint: Standard DNN w/o Design Features	0.34 (0.14)	100%	1.11 (0.09)	90%
Joint T-learner: Standard DNN	0.38 (0.18)	80%	1.09 (0.10)	100%
Individual: Standard DNN	-0.01 (0.09)	100%	0.86 (0.04)	100%
Individual T-learner: Standard DNN	0.01 (0.09)	100%	0.83 (0.03)	100%
Dimension = 40				
Joint: Proposed DRL	0.45 (0.31)	—	1.18 (0.12)	—
Joint: Standard DNN	0.20 (0.18)	90%	1.02 (0.12)	100%
Joint: Standard DNN w/o Design Features	0.32 (0.16)	85%	1.12 (0.12)	90%
Joint T-learner: Standard DNN	0.30 (0.23)	100%	1.13 (0.14)	90%
Individual: Standard DNN	-0.01 (0.07)	100%	0.87 (0.02)	100%
Individual T-learner: Standard DNN	0.01 (0.08)	100%	0.83 (0.02)	100%

Note: We report the mean AUTOC and normalized profit across 20 train–holdout splits. Standard deviations are shown in parentheses. The third and fifth columns ("% Outperf.") report the percentage of splits in which the proposed model (first row) outperformed each benchmark in AUTOC and normalized profit, respectively.

Web Appendix G Analysis of Offer and Covariate Representations

One key feature of the proposed architecture is that it generates two separate representations: one for the intervention design features ($\boldsymbol{\zeta}_k$) and another for the customer covariates ($\boldsymbol{\chi}_{k,i}$). Importantly, these representations capture information directly relevant to treatment effect heterogeneity, providing insights more actionable for managers than those derived from analyzing raw design features or customer covariates alone. In this section, we illustrate these benefits by conducting segmentation analyses based on the learned representations in our empirical application and comparing them with analyses using the original variables, highlighting how the model uncovers distinct and more informative patterns.

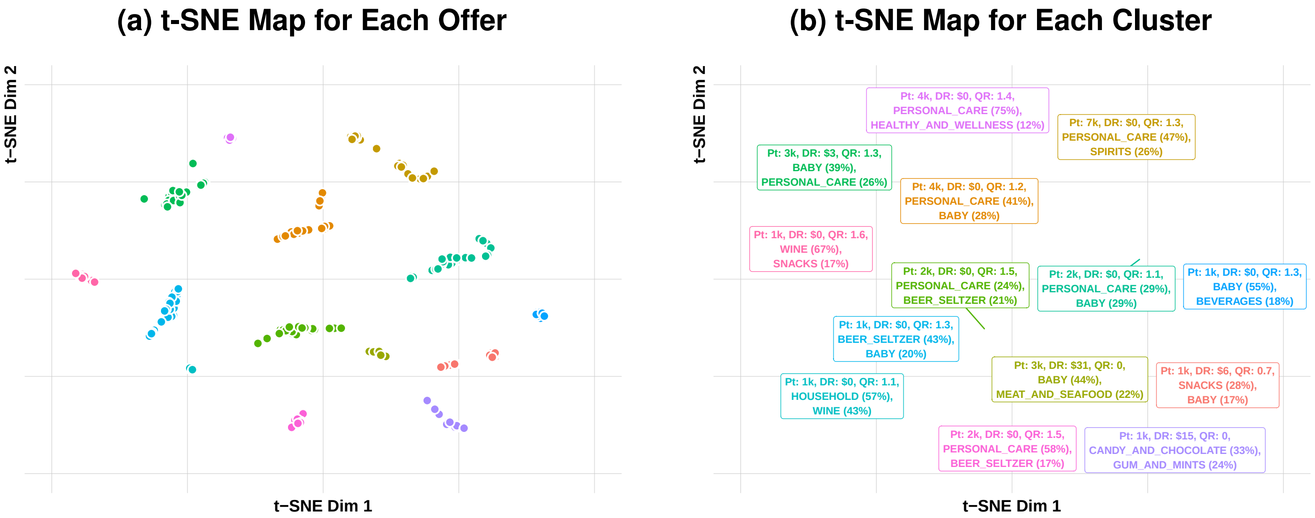
Web Appendix G.1 Offer Design

To illustrate the interpretability of the learned design representations, we visualize the two-dimensional projection of the offer embeddings. Specifically, we extract the representation vectors for the 362 existing offers (denoted as $\boldsymbol{\zeta}_k$) from the proposed DRL-based CIF model. We then apply t-distributed stochastic neighbor embedding (t-SNE; Van der Maaten and Hinton 2008) to project these high-dimensional representations into a two-dimensional space and use the DBSCAN algorithm (Ester et al. 1996) to identify clusters of offers with similar incrementality representations. The resulting t-SNE visualization is presented in Figure App-2 (a), where each point corresponds to a distinct offer design, and the spatial proximity between points reflects similarity in their learned representations. In addition, Figure App-2 (b) summarizes the profiles of each identified cluster based on the raw design features ($\mathbf{Z}_k$) of the offers it contains. For continuous variables, we report the cluster means, and for product categories, we present the most frequently occurring labels within each cluster.

Overall, we find that the representation clusters are primarily shaped by incentive features such as reward points, required spending, and required quantities, which jointly determine the attractiveness and difficulty of an offer. For example, clusters in the upper half of the map correspond to promotions with higher reward point values (e.g., 4,000–7,000 points) and requiring minimum purchase quantities, whereas those in the lower half contain offers with lower to moderate reward levels (around 3,000 points) and requiring minimum dollar spending. This pattern suggests that the company should consider the balance between incentive magnitude and offer difficulty when designing new promotions, ensuring that the reward structure remains appealing while maintaining profitability.

In addition, the t-SNE map shows that offers from different product categories are often located close to one another, reflecting cross-category similarity in their learned incrementality representations. In particular, Candy and Chocolate offers are positioned near Gum and Mints within the purple cluster in the lower-right corner, and this cluster lies adjacent to Snacks (orange cluster). This spatial proximity suggests that promotions in these categories generate similar incremental responses. Such patterns provide insight into which product categories may exhibit greater cross-category transferability, where effective promotion strategies in one category (e.g., Candy and Chocolate) could be applied successfully to others with comparable incentive dynamics, such as Snacks or Gum and Mints.

Figure App-2: t-SNE Maps for Offer Incrementality Representations (ζ_k)



(a) t-SNE Map for Each Offer

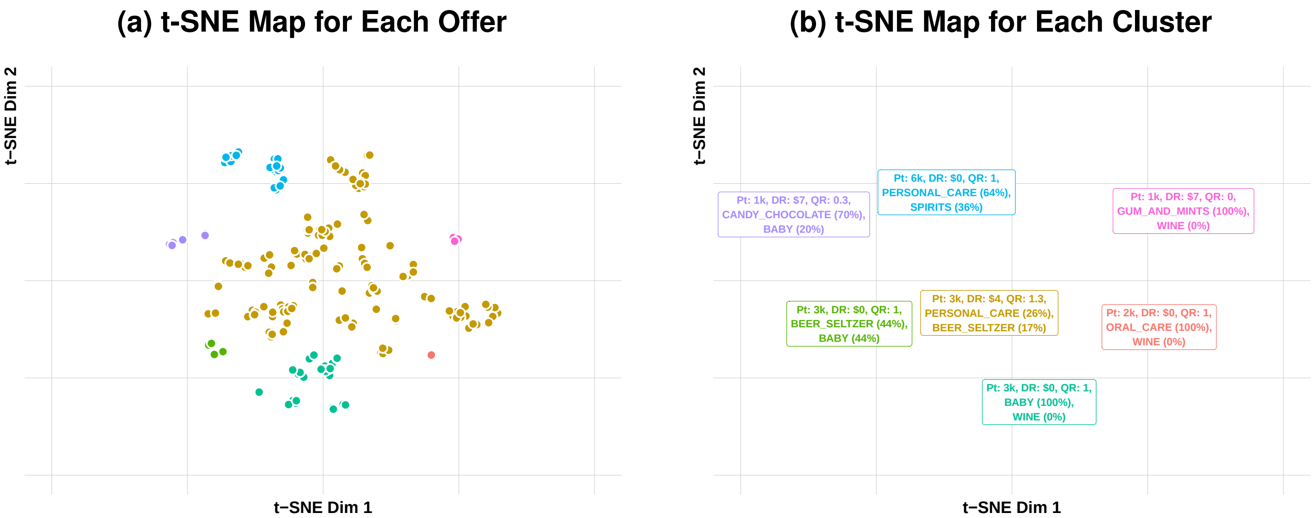
(b) t-SNE Map for Each Cluster

Note. Each point on the map corresponds to a promotional offer, with colors indicating the clusters determined by the DBSCAN algorithm. Labels with colored backgrounds show the average reward points (Pts), required dollar amounts (DR, set as zero for offers requiring certain product quantities), required quantities (QR, set as zero for offers requiring minimum dollar spending), and the top two promoted categories for each cluster with their incidence in parentheses.

Next, we compare the offer incrementality representations to the raw design features to demonstrate that the learned embeddings capture treatment-effect–relevant information beyond simple design similarity. To do so, we replicate the previous t-SNE and clustering exercise but instead use the standardized raw design features (e.g., reward points, required spending, required quantity) as inputs. These features primarily reflect observable design configurations, rather than the underlying incremental impact of the offers. Figure App-3 (a) presents the t-SNE projection of offers based on raw design features, and Figure App-3 (b) summarizes the corresponding cluster features.

Overall, compared to the t-SNE map of the incrementality representations, we find that the t-SNE map based on raw design features exhibits no clear or coherent cluster structure. The resulting clusters are primarily driven by product category rather than by the mechanical incentive structure of the offers. In contrast to the learned representation space, these raw-feature clusters show limited overlap across categories and fail to reveal cross-category relationships. This comparison underscores that the proposed model learns representations driven by treatment-effect similarity, rather than by differences in observed design features across offers.

Figure App-3: t-SNE Maps for Offer Design Features



(a) t-SNE Map for Each Offer

(b) t-SNE Map for Each Cluster

Note. Each point on the map corresponds to a promotional offer, with colors indicating the clusters determined by the DBSCAN algorithm. Labels with colored backgrounds show the average reward points (Pts), required dollar amounts (DR, set as zero for offers requiring certain product quantities), required quantities (QR, set as zero for offers requiring minimum dollar spending), and the top two promoted categories for each cluster with their incidence in parentheses.

Web Appendix G.2 Customer Covariates

We conduct a parallel analysis comparing clusters derived from the learned representations of customer covariates ($\chi_{k,i}$) with those obtained directly from the standardized raw covariates. Specifically, we apply K-means clustering and select three distinct customer segments for illustrative purposes. The analysis is performed using both (a) the incrementality representations derived from the proposed model and (b) the rescaled raw covariate values from a random subsample of 100,000 observations across 362 experiments. Once the segments are identified, we profile each group by averaging customers' covariates to characterize their behavioral patterns. We report both general purchase behavior, which captures all receipts uploaded within the 90 days prior to the offer and the corresponding average order value (AOV), and offer-related behavior, defined as the number of receipts and AOV specifically associated with purchases that include products from the focal offer categories. Table App-10 presents both sets of results, with segments sorted by the average total number of receipts (first column).

Table App-10: Summary of Different Customer Segmentation Approaches)

(a) Segments Based on Incrementality Representations

General		Offer-related		Proportion
Receipts	AOV	Receipts	AOV	
10.3	$21.2	0.19	$1.9	44%
73.6	$59.4	0.42	$3.1	46%
204.7	$101.2	0.74	$5.2	10%

(b) Segments Based on Covariate Values

General		Offer-related		Proportion
Receipts	AOV	Receipts	AOV	
1.0	$4.3	0.00	$0.0	23%
71.0	$61.6	0.06	$0.4	58%
84.6	$54.9	1.82	$ 14.0	19%

Upon initial examination, both approaches appear to segment customers primarily by their overall engagement level with the platform, as reflected in the differences in the number of receipts shown in the first column of Tables App-10 (a) and App-10 (b). However, segmentation based on the incrementality representations reveals more granular distinctions in offer-related behavior. Specifically, while both approaches identify a large segment of moderately engaged customers, the representation-based segmentation distinguishes between those who frequently upload receipts and are more likely to purchase promoted items (Row 2 in Table App-10 (a), receipts = 0.42, offer-related AOV = $3.1) and those with no prior purchase of promoted items (Row 1 in Table App-10 (a), receipts = 0.19, offer-related AOV = $1.9). In contrast, segmentation based on raw covariate values produces clusters that mainly capture general purchasing volume (e.g., high vs. low activity) but fail to differentiate customers by their baseline propensity to purchase promoted items, as both the least and moderately engaged segments (Row 1 and 2 in App-10 (b)) contain customers with no recorded purchases of promoted items.

This divergence has important managerial implications for targeting decisions. For instance, even if the firm aims to deliver interventions to "medium-engagement" customers, the two segmentation approaches would identify significantly different target groups. Under covariate-based segmentation, the firm might target customers who are moderately active overall but have demonstrated little interest in promoted products. In contrast, representation-based segmentation would target customers who exhibit medium engagement but with prior purchases on promoted items — those most likely to generate incremental gains. Thus, segmentation grounded in learned representations enables more precise and causally informed targeting, leading to more effective allocation of marketing resources.