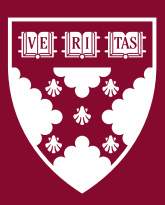
Working Paper 24-076

Incrementality Representation Learning: Synergizing Past Experiments for Intervention Personalization

Ta-Wei Huang
Eva Ascarza
Ayelet Israeli



Harvard
Business
School

Incrementality Representation Learning: Synergizing Past Experiments for Intervention Personalization

Ta-Wei Huang
Harvard Business School

Eva Ascarza
Harvard Business School

Ayelet Israeli
Harvard Business School

Working Paper 24-076

Funding for this research was provided in part by Harvard Business School.

Incrementality Representation Learning: Synergizing Past Experiments for Intervention Personalization*

Ta-Wei Huang Eva Ascarza Ayelet Israeli

June 10, 2024

Abstract

This paper introduces *Incrementality Representation Learning* (IRL), a novel multi-task representation learning framework that predicts heterogeneous causal effects of marketing interventions. By leveraging past experiments, IRL efficiently designs and targets personalized interventions, eliminating the need for extensive testing of numerous potential interventions. To ensure generalization to untested interventions and customers, the IRL model extracts low-dimensional representations of intervention features and customer covariates that are predictive of treatment effects and are generalizable across previously tested interventions. Unlike traditional multi-task learning methods that build separate models for each intervention, IRL uses a unified prediction model across past experiments to enhance generalizability.

We empirically validate our framework in the context of promotional campaigns for consumer packaged goods. By synergizing data from 274 previously conducted experiments, our IRL model not only improves the targeting accuracy of tested interventions but also significantly outperforms existing methods in targeting *untested* interventions and customer segments, overcoming the generalization challenge in high-dimensional decision spaces and the cold-start problem associated with designing new interventions. Furthermore, we develop a decision framework to identify key design features and customer segments for tailoring interventions. Using our model interpretation tool, we demonstrate how companies can customize promotions to enhance profitability across different customer segments.

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

*Ta-Wei Huang is a doctoral candidate at Harvard Business School (email: thuang@hbs.edu). Eva Ascarza is the Jakurski Family Associate 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 data preparation. They also appreciate the valuable feedback from Jeremy Yang, Liangzong Ma, Zhongming Jiang, faculty and students in the HBS Marketing Unit, and members of the Customer Intelligence Lab. Additionally, the authors thank participants of the 2024 American Causal Inference Conference and 2024 Interactive Marketing Research Conference for their helpful comments.

1 Introduction

Businesses are increasingly adopting personalization strategies to enhance their interactions with customers. *Targeting* and *tailoring* have emerged as critical pillars for crafting marketing interventions that enhance customer value. Targeting involves identifying specific customers most likely to respond favorably to interventions and exclusively targeting them. Tailoring further complements these efforts by modifying various design features of interventions to suit different customers, thereby maximizing their effectiveness.

At the core of effective targeting and tailoring is a firm's ability to accurately predict the *incremental impact* of various marketing interventions on individual customers. This precision is typically achieved through *conditional average treatment effect* (CATE) modeling, which estimates the incremental impact of specific interventions on customers based on their observed characteristics. This process traditionally consists of three stages. Initially, a randomized controlled experiment is conducted where a variety of predefined marketing interventions are assigned to a subgroup of customers. Subsequently, a CATE model is constructed for each intervention to predict its incremental impact using pre-treatment customer covariates. Finally, for a new set of customers, the intervention predicted to have the highest CATE is selected for each individual. This method has proven effective across various marketing domains, including free trial design (Yoganarasimhan 2020), product experience (Ye et al. 2023), communication (Ellickson et al. 2023), and promotion (Simester et al. 2020, Zhang and Misra 2022, Daljord et al. 2023). Indeed, many companies have adopted this strategy, conducting experiments for every intervention of interest and developing CATE models to optimize targeting for each specific intervention.

However, this approach becomes impractical when interventions have many customizable design features. For example, optimizing ten features, each with three levels, would require testing $3^{10} = 59,049$ unique interventions in a full factorial design. In addition, each intervention would need a large customer base for accurate CATE estimation, mak-

ing such experiments unfeasible or expensive for many organizations. This research introduces a novel approach that enables companies to target and tailor new interventions by *leveraging past experiments with readily available data*, instead of testing numerous interventions simultaneously (e.g., Ellickson et al. 2024). By pooling knowledge from past experiments, firms can reduce the need for extensive testing and cut costs while effectively targeting and tailoring their marketing interventions. Furthermore, synthesizing information across multiple experiments improves CATE prediction accuracy and targeting effectiveness, providing a significant advantage over analyzing each intervention in isolation, which often fails to detect treatment effect heterogeneity for personalization.

Leveraging past experiments to target and tailor new interventions presents several challenges. First, past experiments usually cover a limited range of interventions, complicating the generalization of model predictions to new, untested interventions. Besides, these experiments are typically tested on specific target populations, limiting the model's ability to predict impact across broader customer segments. Second, the complex interplay between high-dimensional design features and customer covariates makes it challenging to predict and interpret heterogeneous treatment effects across interventions, complicating decision-making when targeting and tailoring new interventions.

To address these challenges, we propose a novel causal machine learning framework called *Incrementality Representation Learning* (IRL). By leveraging past experiments, this framework serves as a foundation model (Bommasani et al. 2021) for treatment effect prediction by integrating intervention design features and customer covariates from various experiments. This approach addresses the *cold-start problem* — evaluating new interventions without prior data. Additionally, our model condenses high-dimensional design features and customer covariates into *low-dimensional representations* that are highly predictive of treatment effects *across interventions*. This enhances the model's generalization to new interventions and broadens its applicability to previously unconsidered customer segments, even with small sample sizes and limited design space coverage.

The IRL methodology consists of two stages. First, we generate an unbiased proxy of the CATE for each individual in each experiment using the cross-fitted doubly robust score (Kennedy 2023). Next, we develop a specialized *deep multi-task representation learning* model. This model has two components: separate encoding networks for intervention design features and customer characteristics, transforming high-dimensional variables into low-dimensional latent representations that capture generalizable information on treatment effect heterogeneity; and a unified prediction network that uses these representations to predict the proxy CATE across all past experiments. This approach allows firms to predict treatment effects for interventions that differ from those previously tested but share similar characteristics, enabling the design and targeting of personalized interventions without extensive factorial experiments. Both the encoding and prediction networks are meticulously calibrated to enhance the accuracy of proxy treatment effects, ensuring their generalizability across various interventions and customer segments. We provide theoretical guarantees for the IRL framework and illustrate the accuracy benefits of the proposed deep learning architecture through theoretical analysis.

We assess the effectiveness of the proposed IRL model in the context of promotional campaigns for consumer packaged goods (CPG). By analyzing data from 274 experiments involving 6,884,833 customers on a leading customer engagement platform, we show that the IRL model significantly outperforms traditional approaches, which typically generate separate models for each intervention. Our model surpasses existing benchmarks in accurately predicting treatment effects and optimizing targeting for both tested promotional offers on *new customers* within the same target population and *untested offers* across different target segments. Notably, the IRL model even outperforms state-of-the-art CATE models trained on *actual* data collected from experiments that tested these new interventions and customer segments (data purposefully excluded when training the IRL model). The results highlight the IRL model's capacity to effectively predict treatment effects by synergizing past experiments and generalizing successfully to new, untested settings.

In addition to enhancing targeting efficiency, we provide a practical framework for more effective intervention tailoring. We demonstrate how managers can identify critical design features and customer covariates through cluster analyses of learned incrementality representations, offering deeper insights into treatment effect heterogeneity compared to traditional analyses. This approach narrows the decision space by focusing on the most important aspects for tailoring interventions. We then introduce the *Segment Incrementality Plot* (SIP), which graphically illustrates how modifying key design feature affect treatment effects across customer segments. Combining the advantages of partial dependence (Friedman 2001) and individual conditional expectation plots (Goldstein et al. 2015), the SIP provides clear, segment-specific insights into promotion responsiveness and effectiveness. Our empirical analysis shows that optimal promotion design varies significantly across customer segments based on engagement levels, highlighting opportunities to boost profitability through tailored promotions

Our research makes three key contributions. Methodologically, we introduce a novel causal machine learning framework that combines deep multi-task representation learning with causal inference to predict treatment effects across different experiments, addressing cold-start and generalization issues for untested interventions or customer segments. We provide theoretical guarantees for the model and outline practical considerations for applying it to new interventions or customers. Substantively, we demonstrate that leveraging data from past experiments can significantly improve both targeting performance and tailoring capabilities, highlighting the potential to capitalize on underutilized past data. Managerially, we streamline the design of personalized marketing interventions with a decision-support tool that uses insights from past experiments, enhancing business outcomes by tailoring interventions to diverse customer segments.

The remainder of the paper is organized as follows: Section 2 provides a review of the literature. Section 3 introduces the IRL modeling framework and its theoretical foundations. Section 4 describes the empirical context and the data used in our study. Section 5

validates the proposed IRL method for treatment effect prediction and targeting for existing promotional offers, and compares its performance against common modeling alternatives. Section 6 demonstrates the superior capability of the proposed IRL in targeting new interventions and customers, as well as tailoring interventions across different customer profiles. Section 7 concludes with a discussion of the findings and outlines directions for future research.

2 Related Literature

Our paper contributes to multiple streams of literature. First, it expands the field of marketing customization, which focuses on predicting purchase or click probabilities and identifying optimal offers based on behaviors without interventions (e.g., Ansari and Mela 2003, Arora and Henderson 2007, Yoganarasimhan 2020, Gabel and Timoshenko 2022, Liberali and Ferecatu 2022). This research primarily aims to match relevant offers with customers rather than creating incremental impact. Recent studies have examined the causal effects of customizable components (Wang et al. 2016, Ellickson et al. 2023, Daljord et al. 2023), but they often focus on simplified decision spaces and estimate the average treatment effect (ATE) across basic customer segments using linear models. Our research introduces a flexible machine learning framework to estimate heterogeneous causal effects across a wide range of interventions and customers, incorporating high-dimensional design features and customer covariates. In addition, unlike previous research that treats intervention design as a mathematical optimization problem (e.g., Ansari and Mela 2003, Zhang and Krishnamurthi 2004, Zhang and Wedel 2009, Golrezaei et al. 2014), our model interpretation framework provides tools for managers to explore complex interactions between design features and customer covariates, enhancing their ability to understand the impact of specific design features on treatment effects.

Second, our research contributes to the evolving field of heterogeneous treatment effect estimation and targeting. While much of the methodological research (e.g., Wager

and Athey 2018, Nie and Wager 2021, Semenova and Chernozhukov 2021, Kennedy 2023) focuses on single experiments, our study expands these efforts to multiple interventions. Recent literature has explored treatment effect estimation for multiple treatments (Chintagunta et al. 2023, Daljord et al. 2023, Ellickson et al. 2023, Ye et al. 2023) within relatively small intervention and customer spaces. Ellickson et al. (2024) demonstrates the value of incorporating contextual embeddings from large language models to predict treatment effects of untested but similar email campaigns, using leave-one-email-out validation. Our paper extends this literature by (i) proposing a new model that handles complex, high-dimensional design features and customer covariates, improving generalization for untested interventions and customer segments, (ii) demonstrating the power of synergizing past experiments to predict treatment effects for significantly different interventions (new product categories) and customer segments (new demographics for interventions of interest), and (iii) characterize the practical limitations in predicting treatment effects for novel treatments using past experiments.

Third, our work relates to the literature on the transportability of causal effects (Simester et al. 2020, Degtiar and Rose 2023). This issue arises when the target population extends beyond the experimental population, making it challenging to identify causal effects due to the non-overlap of populations. Strategies to address this include (i) trimming the target population to match the experimental data coverage (e.g., Crump et al. 2009, Petersen et al. 2012), (ii) integrating behaviors for non-treated segments from other studies with existing experimental data (e.g., Gordon et al. 2023, Zivich et al. 2024), and (iii) extrapolating using smoothness assumptions on CATE functions (Nethery et al. 2019, Khan et al. 2023, Rafieian 2023, Zhu et al. 2023, Ellickson et al. 2024). Our methodology combines the latter two strategies, exploiting the smoothness of the CATE function to predict treatment effects for previously untested interventions or customers by synergizing past experiments. We also provide a theoretical characterization of the transportability problem using domain adaptation theories (Ben-David et al. 2010).

Fourth, our research adds to the literature on multi-task and transfer learning (e.g., Zhuang et al. 2020, Zhang and Yang 2021), which concentrates on leveraging knowledge from diverse machine learning problems (source tasks) to enhance performance in specific focal tasks (target tasks). While there is growing interest in applying these methods to address business decision challenges, our contributions are distinct. Unlike prior studies that primarily focus on non-experimental settings to predict outcomes (Kini and Manjunatha 2020, Bastani 2021, Xu et al. 2021), our framework enables firms to synergize multiple randomized controlled experiments to predict causal effects. Furthermore, previous research constructs models individually for each of the past experiments and integrates these models with the actual data from the focal experiment (Timoshenko et al. 2020); our approach demonstrates the advantage of modeling all past experiments simultaneously, even *without conducting the focal experiment*. This benefit stems from our model's ability to directly share information across previous experiments, enabling companies to design new interventions without conducting an actual experiment.

Finally, this paper relates to the growing body of work on representation learning within the context of marketing. Dew et al. (2022) introduce a multimodal representation learning framework for integrating diverse data types related to logos, while Gabel et al. (2019) and Chen et al. (2022) propose extend the Word2Vec model (Mikolov et al. 2013) to generate latent product representations based on co-purchasing patterns. Gabel and Timoshenko (2022) and Chen et al. (2022) extend these product representations into customer choice models for marketing mix modeling. Unlike previous work, our framework prioritizes distilling critical information related to treatment effect heterogeneity instead of merely summarizing the original raw data (e.g., logo images or product co-purchase matrices) into dense vectors. Therefore, we utilize *supervised representation learning*, enabling our algorithm to extract the most generalizable information and exclude idiosyncratic details.

3 Model

In this section, we first specify the necessary conditions for utilizing past experiments to identify the treatment effect of a marketing intervention on a customer, taking into account both the design features of the intervention and the customer's covariates. We then discuss the potential challenges of generalizing past experiments to new settings and propose the IRL method as a solution. This approach enables companies to effectively tackle two crucial questions in designing personalized interventions: "whom to target" (the *targeting* problem) and "what to offer" (the *tailoring* problem) for new interventions

In line with our empirical application, we use "offers" to refer to interventions and "customers" as the unit of analysis. However, the applicability of our framework extends beyond marketing scenarios. It is relevant in any context where researchers or decision-makers have administered various treatments or stimuli to diverse units or individuals and aim to either enhance the efficacy of existing treatments or develop new ones.

3.1 Problem Setup

We consider a scenario where a company has already executed a series of experiments on a variety of offers. In each of these experiments, a distinct offer $o \in \{1, \cdots, O\}$, characterized by its design features $\mathbf{Z}_o \in \mathcal{Z} \subseteq \mathbb{R}^D$), was tested on a representative sample ($\mathcal{C}_o^E$) before being deployed to the target population ($\mathcal{C}_o$). In the experimental sample, customers were randomly assigned to either a treatment group that received the offer ($W_{o,i} = 1$) or a control group that did not ($W_{o,i} = 0$). For each participant i in $\mathcal{C}_o^E$, the company collects pre-treatment covariates $\mathbf{X}_{o,i} \in \mathcal{X} \subseteq \mathbb{R}^P$. These covariates include both static information, such as demographics, and offer-dependent variables, like previous purchase behaviors. We define $Y_{o,i}$ as the target outcome that the company aims to influence through offer o, such as the total spending on items promoted by the offer.

3.1.1 Causal Estimand

When the objective is making targeting decisions for individual offers, the target causal estimand is typically the CATE for *each specific offer*. This is formally defined as follows:

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

where $Y_{o,i}(W_{o,i} = 1)$ and $Y_{o,i}(W_{o,i} = 0)$ denote the potential outcomes for customer i if they were to receive or not receive offer o, respectively. In this case, the company typically estimates O distinct CATE models based on experimental data and formulates separate targeting strategies for each offer.

While this estimand may be effective for targeting decisions on previously tested offers, estimating CATEs individually does not allow companies to effectively target new offers due to the absence of experimental data for these offers. A naive approach to address this issue might assume a uniform CATE function across all offers and construct a CATE model using data from all experiments. However, this approach does not account for how different design features impact incremental effects across various customer segments. Consequently, this model falls short in its ability to tailor offers to specific customer groups.

To effectively target and tailor new interventions, we propose using the CATE conditioned on both the design features and customer covariates as the target estimand for personalization. This estimand is defined as follows:

$$\tau(\mathbf{Z}_o, \mathbf{X}_{o,i}) \equiv \mathbb{E}_{\mathcal{C}_o}[Y_{o,i}(\mathbf{Z}_o) - Y_{o,i}(\mathbf{0})|\mathbf{Z}_o, \mathbf{X}_{o,i}], \tag{1}$$

where $Y_{o,i}(\mathbf{Z}_o)$ is the potential outcome for observation i should they receive an offer with design features $\mathbf{Z}_o$, and $Y_{o,i}(\mathbf{0})$ is the potential outcome in the absence of offer o.

By integrating data from multiple experiments, we can estimate (1) by leveraging variations in design features and customer covariates. This approach offers several advantages over estimating CATE models individually for each offer. First, analyzing design

features across various offers helps develop new marketing strategies tailored to maximize incremental gains. Second, using data from diverse offers and customer pools reveals treatment effects on previously ineligible individuals, enhancing the generalizability of predictions to new customer segments. Third, it improves targeting accuracy by considering interactions between offer designs and customer demographics. Finally, pooling data across experiments increases the sample size, improving the precision of treatment effect estimates.

3.1.2 Identification Assumptions

The ideal approach for estimating the quantity described in (1) would involve conducting an experiment with a full factorial design, which randomly assigns all possible combinations of design features across a large customer base. However, this approach poses significant operational challenges, particularly when it requires the random allocation of many potential treatments simultaneously — approximately one billion possible offers in our empirical application — making it impractical for most scenarios. Therefore, our goal is to estimate (1) using readily available data, specifically by leveraging experimental data from offers that have already been tested.

Since this approach does not involve direct manipulation of each offer's design features, it is essential to ensure that past experiments meet specific conditions to identify the target estimand in (1) using data from multiple single-offer experiments. Below, we outline the assumptions required to ensure identification based on past experiments.

Assumption 1 *The following identification assumptions are made:*

1. ***(Overlap)*** *In each experiment $o \in 1, \cdots, O$, the assignment of treatment is subject to random variation, that is,* $0 < \mathbb{P}(W_{o,i} = 1 | \mathbf{X}_{o,i}) < 1, \quad \forall\, \mathbf{X}_{o,i} \in \mathcal{X}_o, \;\; o = 1, \cdots, O.$

2. ***(Unconfoundedness)*** *In each experiment $o \in 1, \cdots, O$, the treatment assignment is free from unobserved confounders, that is,* $\{Y_{o,i}(\mathbf{Z}_o), Y_{o,i}(\mathbf{0})\} \perp W_{o,i} \mid \mathbf{X}_i, \mathbf{Z}_o, \quad \forall i \in \mathcal{C}_o^E, \; \forall o = 1, \cdots, O.$

3. ***(No Interference)*** *The potential outcomes for customer i in response to offer o are independent of the presence of other offers received by customer i and offers received by other customers i', meaning that $\{Y_{o,i}(\mathbf{Z}_o), Y_{o,i}(\mathbf{0})\} \perp \{W_{o',i}\}_{o' \neq o}$ and $\{Y_{o,i}(\mathbf{Z}_o), Y_{o,i}(\mathbf{0})\} \perp \{W_{o',i'}\}_{i' \neq i,\ o'=1,\cdots,O}$.*

4. ***(Stability)*** *The experimental sample (including both the treatment and control groups) and the future target population are comparable, that is, $\mathbb{E}_{\mathcal{C}_o^E}[Y_{o,i}(\mathbf{Z}_o)|\mathbf{X}_i] = \mathbb{E}_{\mathcal{C}_o}[Y_{o,i}(\mathbf{Z}_o)|\mathbf{X}_i]$ and $\mathbb{E}_{\mathcal{C}_o^E}[Y_{o,i}(\mathbf{0})|\mathbf{X}_i] = \mathbb{E}_{\mathcal{C}_o}[Y_{o,i}(\mathbf{0})|\mathbf{X}_i]$ for any $o \in \{1, \cdots, O\}$,*

The first and second assumptions are the standard overlap and unconfoundedness assumptions in the literature (Imbens and Rubin 2015). The third assumption ensures that the behavior of one customer does not influence another, and that individual offers do not interfere with each other. In our empirical setting, the likelihood of offer interference is minimal because: (i) customers did not receive offers for identical items simultaneously; (ii) the outcome variable solely captures purchases of items promoted by the focal offer; and (iii) both treatment and control groups had an equal likelihood of receiving other offers, due to the company's implementation of complete randomization. The fourth assumption ensures that the future targeting pool for an offer will behave in the same manner as the customers included in the experiment.

Under these assumptions, the customer response function $\mathbb{E}[Y_{o,i}(\mathbf{Z})|\mathbf{X}_{o,i}]$ is identified at $\mathbf{Z} = \mathbf{Z}_o$ and at $\mathbf{Z} = \mathbf{0}$ for customers within $\mathcal{C}_o$. We can express this result as:

$$\begin{aligned} \mathbb{E}_{i \in \mathcal{C}_o}[Y_{o,i}(\mathbf{Z}_o)|\mathbf{X}_{o,i}] &= \mathbb{E}_{i \in \mathcal{C}_o^E}[Y_{o,i}|W_{o,i} = 1, \mathbf{X}_{o,i}], \quad \forall\, \mathbf{X}_{o,i} \in \mathcal{X}_o,\ \ o = 1, \cdots, O, \\ \mathbb{E}_{i \in \mathcal{C}_o}[Y_{o,i}(\mathbf{0})|\mathbf{X}_{o,i}] &= \mathbb{E}_{i \in \mathcal{C}_o^E}[Y_{o,i}|W_{o,i} = 0, \mathbf{X}_{o,i}], \quad \forall\, \mathbf{X}_{o,i} \in \mathcal{X}_o,\ \ o = 1, \cdots, O. \end{aligned} \tag{2}$$

This formulation implies that one can aggregate data from prior experiments and employ well-established CATE models to estimate the treatment effect given specific design features and customer covariates.

3.1.3 Interpolation and Extrapolation Challenge

It is crucial to recognize that Equation (2) provides a point-wise identification result, which only guarantees the identification of the CATE for those combinations of design

features and customer covariates that have been previously observed. As such, predicting (1) for *untested* offers and customer segments — key to our research objectives — requires some degree of interpolation or extrapolation from the CATEs of previously tested offers.[1] Given the high-dimensional nature of design features and customer covariates, it is inevitable that some combinations of offer design and customer segments will remain unobserved, despite numerous past experiments (Balestriero et al. 2021).

Since the most widely-used CATE models, such as the R-learner (Nie and Wager 2021) and Causal Forest (Wager and Athey 2018), are primarily designed for scenarios within a single intervention and do not support the interpolation and extrapolation of CATEs from past experiments to untested interventions, there is a pressing need to develop a model that can effectively generalize CATE predictions to new combinations of designs and customers. Such a model should not only predict the CATE function for untested offers and customers using design features and covariates but also possess robust generalization ability by learning the common structure across different offers while discarding idiosyncratic, offer-specific patterns.

3.1.4 Solution Concept

To tackle this challenge, we introduce the IRL framework, which merges causal machine learning (Semenova and Chernozhukov 2021, Nie and Wager 2021, Kennedy 2023) with multi-task representation learning (Baxter 2000, Ben-David and Schuller 2003, Maurer et al. 2016, Watkins et al. 2024). This approach aims to accomplish two primary objectives: firstly, to derive low-dimensional representations from design features and customer covariates that summarize generalizable information regarding treatment effect heterogeneity across experiments; and secondly, to learn the common relationship between these representations and the actual treatment effects across different offers using a unified prediction model.

[1] This challenge is common in many causal inference settings, particularly when dealing with a finite experimental sample and high-dimensional customer covariates that influence treatment effect heterogeneity (Petersen et al. 2012, Hill and Su 2013).

The proposed model architecture provides an accurate foundation model for treatment effect predictions for three reasons. First, extracting low-dimensional common representations across all promotional offers can often mitigate the risks associated with extrapolation in the high-dimensional spaces of design features and customer covariates (Balestriero et al. 2021, Degtiar and Rose 2023). Typically, modeling in such high-dimensional decision spaces demands a significantly larger sample size to achieve adequate coverage. However, if there is an underlying low-dimensional latent space of input variables that effectively captures the essential information of CATE, achieving sufficient coverage in this reduced space becomes much more feasible (Bárány and Füredi 1988, Bonnasse-Gahot 2022).

Second, by training a single prediction model to minimize prediction errors across all experiments, we ensure that IRL discards offer-specific idiosyncrasies in the relationships between the learned representations and the true CATE. This approach enhances generalizability to untested offers and customer segments by identifying and leveraging overarching patterns of treatment effect heterogeneity across experiments.

Third, the proposed IRL model enhances the accuracy of CATE predictions by increasing statistical power through the pooling of information across experiments. Leveraging knowledge from multiple prediction tasks consistently outperforms models trained on individual tasks alone, as supported by both theoretical (Baxter 2000, Ben-David and Schuller 2003, Maurer et al. 2016) and empirical research (Timoshenko et al. 2020, Vafaeikia et al. 2020, Zhang and Yang 2021). In real-world marketing applications, estimating heterogeneous treatment effects for a single experiment can be challenging due to small treatment effects being overshadowed by significant noise in the outcome variable (Huang and Ascarza 2024). By pooling information from multiple experiments, IRL effectively increases the effective sample size and, thus, the statistical power, enhancing prediction accuracy compared to models that treat each experiment independently.

Next, we detail the proposed IRL model, present its theoretical foundations, and discuss practical considerations for its application to real-world data.

3.2 Incrementality Representation Learning

Our proposed IRL framework consists of two stages. First, we use state-of-the-art techniques to construct an unbiased proxy for the true CATE. Following this, we develop a deep supervised representation learning model tasked with predicting this proxy CATE. The model employs design features and customer covariates to generate low-dimensional representations that capture generalizable insights about treatment effect heterogeneity. These representations are then used to predict the proxy CATE.

3.2.1 Derive Unbiased Proxy for CATEs

The first step creates an unbiased proxy for the CATE corresponding to each customer and offer. This proxy will then act as the prediction target for our deep multi-task representation learning algorithm. The following theorem demonstrates how to derive an unbiased proxy for the CATE defined in (2) using the doubly robust score (Kennedy 2023).

Theorem 1 (Unbiased Score for CATE) *Let $\mu_{o,w}(\mathbf{X}_{o,i}) = \mathbb{E}_{\mathcal{C}_o^E}[Y_{o,i}|W_{o,i} = w, \mathbf{X}_{o,i}]$ represent the conditional expected outcome for offer o, and let $\pi_o(\mathbf{X}_{o,i}) = \mathbb{P}_{\mathcal{C}_o^E}[W_{o,i} = 1|\mathbf{X}_{o,i}]$ denote the propensity score of being treated in experiment o. Then, under Assumption 1, we have the following identification result for the doubly robust score:*

$$\tau(\mathbf{Z}_o, \mathbf{X}_{o,i}) = \mathbb{E}_{\mathcal{C}^E}\left[\mu_{o,1}(\mathbf{X}_{o,i}) - \mu_{o,0}(\mathbf{X}_{o,i}) + \frac{W_{o,i} - \pi_o(\mathbf{X}_{o,i})}{\pi_o(\mathbf{X}_{o,i})[1 - \pi_o(\mathbf{X}_{o,i})]}\left[Y_{o,i} - \mu_{o,W_{o,i}}(\mathbf{X}_{o,i})\right] \middle| \mathbf{X}_{o,i}, \mathbf{Z}_o\right]$$

for all $\mathbf{X}_{o,i} \in \mathcal{X}_o,\ o = 1, \cdots, O$. Here, $\mathcal{C}^E = \bigcup_{o=1}^{O} \mathcal{C}_o^E$ denotes the collection of all observations from past experiments.

The proof of Theorem 1 is provided in Web Appendix B.2. Theorem 1 demonstrates that to predict the CATE defined in (1), one can first calculate the doubly robust score for each

offer, then combine these scores across offers, and finally build a machine learning model using the design features and covariates across offers to predict the score.

In our empirical application, we utilize cross-fitting (Nie and Wager 2021) to construct the doubly robust score:

$$\widetilde{\tau}_{o,i} \equiv \left[\widehat{\mu}_{o,1}^{[-i]}(\mathbf{X}_{o,i}) - \widehat{\mu}_{o,0}^{[-i]}(\mathbf{X}_{o,i})\right] + \frac{W_{o,i} - \widehat{\pi}_{o}^{[-i]}(\mathbf{X}_{o,i})}{\widehat{\pi}_{o}^{[-i]}(\mathbf{X}_{o,i})[1 - \widehat{\pi}_{o}^{[-i]}(\mathbf{X}_{o,i})]} \left[Y_{o,i} - \widehat{\mu}_{i,W_{o,i}}^{[-i]}(\mathbf{X}_{o,i})\right], \quad (3)$$

where $\widehat{\mu}_{o,w}^{[-i]}(\mathbf{X}_{o,i})$ represents a (machine learning) model that predicts the expected outcome $Y_{o,i}$ for experiment o given the treatment assignment $W_{o,i} = w$ and the covariates $\mathbf{X}_{o,i}$, and $\widehat{\pi}_{o}^{[-i]}(\mathbf{X}_{o,i})$ denotes the propensity score model for experiment o, estimating the probability of receiving the treatment based on the covariates. The superscript $[-i]$ indicates that both the expected outcome and the propensity score predictions for observation i are based on models trained on data excluding that of observation i.

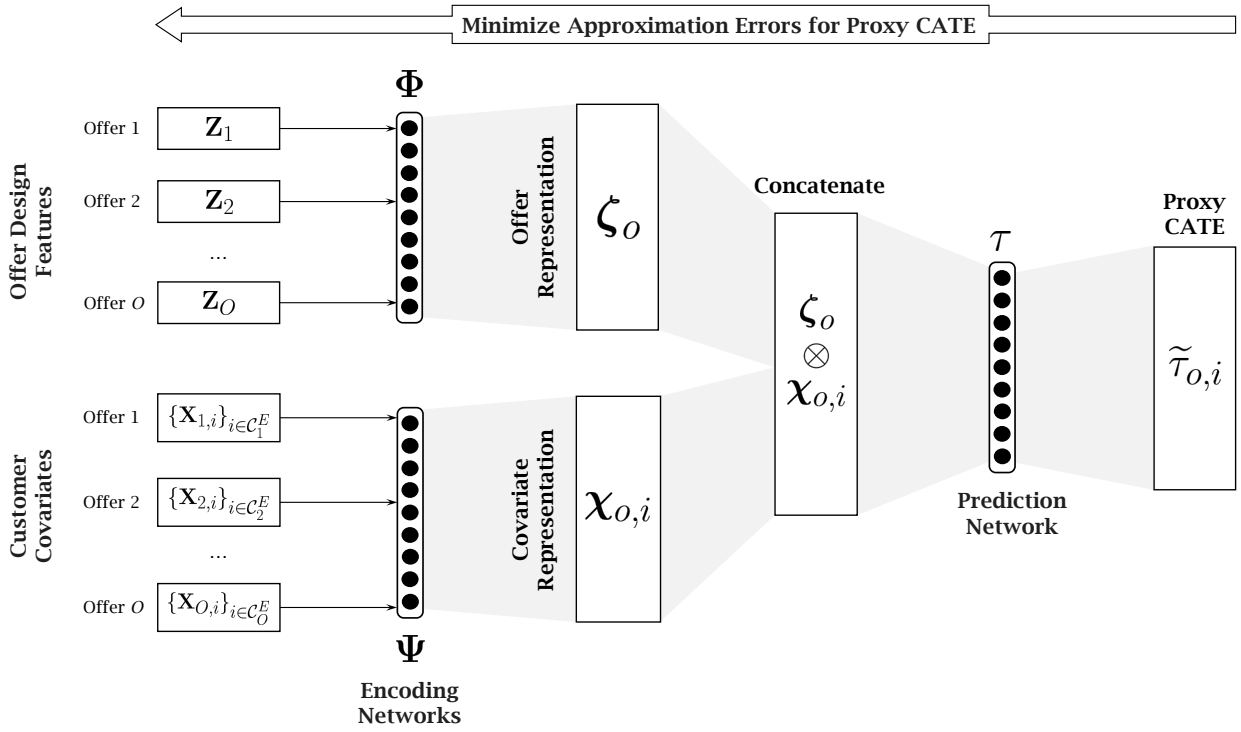
3.2.2 Deep Learning Framework for IRL

Next, we develop a deep multi-task representation learning model for CATE prediction. This model utilizes two distinct *encoding networks*: $\mathbf{\Phi} : \mathbb{R}^D \to \mathbb{R}^{K_1}$ for design features, where $K_1 < D$, and $\mathbf{\Psi} : \mathbb{R}^P \to \mathbb{R}^{K_2}$ for customer covariates, where $K_2 < P$. These networks convert the high-dimensional design features and customer covariates into low-dimensional representations: $\boldsymbol{\zeta}_o = \mathbf{\Phi}(\mathbf{Z}_o)$ for design features and $\boldsymbol{\chi}_{o,i} = \mathbf{\Psi}(\mathbf{X}_{o,i})$ for customer covariates. The primary function of these encoding networks is to extract the most relevant and generalizable information for predicting treatment effects across different offers. Then, the model implements a prediction network, $\tau : \mathbb{R}^{K_1+K_2} \to \mathbb{R}$, which utilizes these condensed representations to predict the doubly robust score (as detailed in Section 3.2.1) for each observation within an experiment.

Model Architecture. Figure 1 illustrates the proposed model architecture for IRL. Here, the model first transforms the high-dimensional design features $\{\mathbf{Z}_o : o = 1, \cdots, O\}$ and customer covariates $\{\mathbf{X}_{o,i} : \ i \in \mathcal{C}_o^E, \ o = 1, \cdots, O\}$ into lower-dimensional representations ($\boldsymbol{\zeta}_o$ for design features and $\boldsymbol{\chi}_{o,i}$ for customer covariates) through two separate deep

neural networks ($\mathbf{\Phi}$ for design features and $\mathbf{\Psi}$ for customer covariates). Next, these representations are concatenated into extended vectors and inputted into another deep neural network (τ) to predict the proxy CATE ($\widetilde{\tau}_{o,i}$).

Figure 1: Proposed Model Architecture for Incrementality Representation Learning



One key design choice in our proposed architecture is the *separate construction* of representations for design features and customer covariates, instead of combining them into a single shared representation. This approach provides two critical benefits. First, it reduces the risk of overfitting and enhances generalization. By isolating the representations, this method prevents idiosyncratic interactions between design features and customer covariates that are not generalizable across offers during the representation learning process. The theoretical benefits of this approach are detailed in Web Appendix A.3.

Second, separating design features and customer covariates into distinct representations enhances the model's interpretability. As detailed in Section 6.2, clustering design feature representations identifies distinct offer groups with similar treatment effect patterns and highlights influential design features. Similarly, clustering customer covariate

representations detects customer segments with similar responses, improving targeting strategies. In contrast, a combined representation might obscure critical insights into how specific aspects of the offer or customer characteristics influence treatment effects, thus complicating the analysis for managers.

Another key design aspect of the proposed model architecture is the decision to *share all parameters* across the prediction network for different experiments, rather than creating offer-specific prediction networks for each individual offer — a common approach in multi-task learning. This decision is driven by two considerations. First, the model is designed to predict treatment effects for *untested* offers, for which no experimental data exists. Offer-specific networks would not be feasible in this scenario, as they would be unable to predict the impact of previously untested offers. Second, from an accuracy perspective, the CATE signal $\widetilde{\tau}_{o,i}$ often exhibits significant variance, primarily due to inverse propensity score weighting and unobserved customer heterogeneity (Huang and Ascarza 2024). Constructing an accurate prediction network for each offer is challenging due to the high level of noise and the relatively small sample sizes typical of individual experiments. Conversely, a unified prediction network that consolidates all experiments can enhance prediction stability by leveraging a larger training dataset.

Model Calibration. We calibrate the three neural networks — $\mathbf{\Phi}$, $\mathbf{\Psi}$, and τ — simultaneously, ensuring that $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are optimized to capture the most salient representations for treatment effect prediction. Specifically, we first define the architecture for each network. Following this, we compile data from all experiments and calibrate $\mathbf{\Phi}, \mathbf{\Psi}, \tau$ by minimizing the squared error between the model's predictions and the doubly robust scores:

$$\underset{\tau, \mathbf{\Phi}, \mathbf{\Psi}}{\arg\min} \sum_{o=1,\cdots,O} \sum_{i \in \mathcal{C}_o^E} w_o \left[\tau(\mathbf{\Phi}(\mathbf{Z}_o), \mathbf{\Psi}(\mathbf{X}_{o,i})) - \widetilde{\tau}_{o,i}\right]^2, \tag{4}$$

where w_o denotes the relative importance weight assigned to each offer, which can be manually specified or treated as tuning parameters. For simplicity, we pick $w_o = 1$ for all offers in our empirical application.

3.2.3 Theoretical Guarantees

In Web Appendix A, we provide theoretical guarantees for the proposed IRL framework and model architecture. First, we establish upper bounds on the prediction error for the true CATE for any IRL model, drawing on the theoretical results from Tripuraneni et al. (2020). We detail how the error bound in our context is determined by three key factors: (i) the number of offers previously tested, (ii) the sample size of each past experiment, and (iii) the complexity of the representation and prediction models. Next, we illustrate how our specialized deep learning architecture can improve these theoretical upper bounds compared to traditional deep learning models. Specifically, we demonstrate the benefits of dimension reduction in deep representation learning, the advantages of employing distinct network structures for modeling offer and customer representations, and the benefit of learning shared parameters in the prediction network. This theoretical foundation supports the enhanced accuracy of our approach over conventional deep learning architectures in our empirical applications.

3.3 Practical Considerations

The IRL framework is a powerful tool for leveraging insights from past experiments to improve targeting and personalization for both existing and new offers. However, its effectiveness for untested offers and customers hinges on the similarity between the data-generating processes of the past experimental data and the new dataset. When managers apply models trained on past experimental data, there are two considerations that require particular attention. First, if past experiments lack diversity in offer design or customer covariates, a CATE model may not generalize well to new offers or customer groups significantly different from those in previous scenarios. Second, if customer responses to

offers evolve over time or untested customers have different sensitivities (i.e., the stability assumption in Assumption 1 is violated), reliance solely on historical data may cause the CATE model to generate inaccurate predictions for new data.

To formalize this, we denote the data-generating process for $\mathbf{Z}_o, \mathbf{X}_{o,i}$ in past experiments as $\mathcal{D}_{\text{past}}$, with the true CATE for these offers and customers represented by τ_{past}. The managers aim to predict the treatment effects for a new set of offers and customers, characterized by a different data-generating process $\mathcal{D}_{\text{new}}$ for their design features and customer covariates, with their true CATE function expressed as τ_{new}. Building on classical domain adaptation theory (Ben-David et al. 2010), we identify three critical factors that influence the generalization performance of any CATE model trained on past experiment data, as outlined in the following theorem.

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

$$\min_{\widehat{\tau}} \quad \frac{1}{\sum_{o=1}^{O} |\#\mathcal{C}_o^E|} \sum_{o=1}^{O} \sum_{i \in \mathcal{C}_o^E} \ell\left(\widehat{\tau}(\mathbf{Z}_o, \mathbf{X}_{o,i}), \widetilde{\tau}_{o,i}\right),$$

where $\widetilde{\tau}_{o,i}$ is an unbiased estimate of $\tau_{\text{past}}(\mathbf{Z}_o, \mathbf{X}_{o,i})$, ℓ is the squared loss function, and $|\#\mathcal{C}_o^E|$ denotes the number of observations in $\mathcal{C}_o^E$. 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:

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

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, $(\mathbf{Z}_o, \mathbf{X}_{o,i})$ are omitted from each function's notation.

The proof of Theorem 2 is provided in Web Appendix B.5. Theorem 2 identifies three critical factors influencing the upper bound of generalization error: (i) the accuracy of the model's predictions for existing offers, (ii) the differences in the distribution of design features and customer covariates between past and new offers (i.e., *attribute shifts*), and (iii) the disparity in the true CATE functions between past and new offers (i.e., *concept shifts*). The first factor represents prediction error when new data's distribution and true CATE function match those of past data. The second factor addresses challenges when new offers have design features and customer covariates that differ significantly from past experiments. For instance, if past experiments lacked offers for baby products but the new set includes them, this can hinder generalization. The third factor highlights potential errors when the responses of untested customers differ significantly from those previously tested, such as household versus non-household customer sensitivities to promotions.

In our empirical analysis, we demonstrate that the proposed IRL model outperforms existing methods in generalizing to previously untested offers, specifically in product categories that had never been tested before, and to customer segments that were previously ineligible for certain offers (see Section 6.1). However, the model's ability to generalize to untested offers or customer segments that are substantially different from those previously tested remains an empirical question. If companies are seriously concerned about these issues, they could conduct small-scale experiments with the new offers and customers, then apply the proposed IRL model to this new experimental data. Alternatively, they could perform model fine-tuning to enhance the performance of the IRL model trained on past experimental data.

4 Empirical Context and Data

4.1 Empirical Context

The dataset is provided by a leading consumer-engagement platform in North America. Customers on this platform earn reward points by scanning their shopping receipts,

which they can then redeem for gift cards or products. The platform generates revenue by providing customer relationship management services to CPG companies and retailers, who pay commissions for these services. A key feature of the platform is its assistance in creating *special promotional offers* for these partners, such as "buy two bottles of wine from Brand W and get 2,000 points." These offers are prominently displayed on the platform's discovery page, allowing users to earn additional points by submitting receipts that meet each offer's requirements.

The primary goal of the platform is to boost the incremental sales of promoted products. With numerous partners aiming to expand their customer base and increase sales through branded promotional campaigns, the company seeks to deliver more targeted offers on the discovery page to reduce the volume of irrelevant offers. Furthermore, since the platform's value proposition for partners centers on demonstrating a sales lift, it is crucial to strategically allocate limited promotional budgets to customers most likely to generate the highest incremental sales. Consequently, the success of the platform depends on precisely tailoring and targeting optimal promotional offers to the audiences most likely to make additional purchases in response to these promotions.

4.2 Data Description

To demonstrate sales lift, the platform routinely conducts randomized controlled experiments for its special promotional offers. For each offer, about 10% of eligible customers are randomly chosen to form a control group that does not receive the offer. We analyzed 274 such experiments, each involving different promotions conducted between October 2022 and April 2023. Each of these promotions was tested on 10,000 to 200,000 customers.[2]

4.2.1 Offers

Each offer is characterized by several key design features: (i) the promoted product and its category, (ii) point-earning criteria, like minimum purchases or quantities, (iii) points

[2]We focus exclusively on the 76% of offers that exhibit non-negative ATEs. Offers with negative effects are excluded as they may be influenced by substantial noise or issues with execution. However, our findings remain consistent even if we use all the offers, albeit with higher standard error in targeting performance. See Web Appendix F.2 for more details.

awarded for fulfilling these criteria, and (iv) extra details like redemption limits or promotional tactics. Overall, our dataset comprises fourteen design features, including six numerical and eight discrete variables, expanding to 46 variables following dummy encoding. A summary of these features is detailed in Table 1.

Table 1: Summary of Offer Design Features.

Discrete Design Features

Variable	Unique Value	Top Counts
Is the offer a multi-transaction offer?	2	False: 162, True: 112
Is the offer stackable?	2	False: 33, True: 241
Criteria: Purchase from selected items	2	False: 242, True: 32
Criteria: Meet minimum quantity	2	False: 34, True: 240
Criteria: Available for club members only	2	False: 260, True: 14
Criteria: Actions for reward claim	2	Quantity: 240, Spending: 34
Promoted product category	19	Beer: 57, Personal Care: 55, Baby: 52, Wine: 31
Promotion tactic	14	Dollar_or_unit: 125, Frequency: 57, Competitive_targeting: 47

Numerical Design Features

Variable	# Missing	Mean	S.D.	Min	P25	Median	P75	Max
Points awarded	0	2,587	2,305	350	1000	2,000	3,500	20,000
Minimum required spending ($) to redeem the reward points	240	18.9	11.4	5.0	10.0	15.0	30.0	50.0
Minimum required quantity to redeem for the reward	34	1.37	0.614	1	1	1	2	4
Total number of times the offer can be redeemed	0	1.84	1.5	1	1	1	2	10
Maximum redemptions allowed per transaction	0	1.41	1.11	1	1	1	1	5
Duration of the promotion (in days)	0	43.8	30.6	5	28	31	60	176

Note. For instance, the offer "Spend $30 on selected sizes of diapers and earn 3,000 points" is characterized as a single-transaction, non-stackable offer. It specifies the criteria as *purchase from selected items* along with a *spending* threshold for claiming the reward. The variable "Is the offer stackable?" indicates whether a purchase related to this offer can also count towards another offer. The variable "promotion tactics" describes the strategy behind the offer. For instance, "Dollar_or_unit" is designed to increase the dollars spent or units purchased of specified items, while "Frequency" aims to enhance the purchase frequency of the items.

4.2.2 Customers

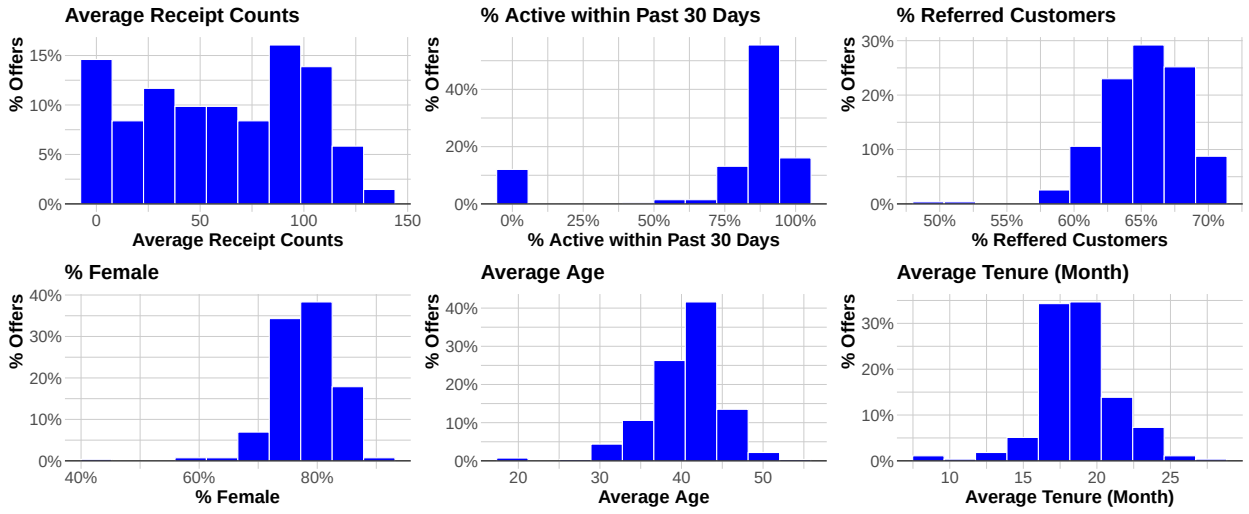
Our sample contains information from 6,884,833 distinct customers, each of whom was included in at least one experiment. In total, we collect 40 pre-treatment covariates for capturing heterogeneous customer reactions to a promotional offer. Below is an overview of each covariate group:

1. **User attributes:** Customer characteristics such as gender, age, duration of membership on the platform (measured at the time they received the offer), and whether they joined through a referral.
2. **Non-contextual behaviors:** Summaries of purchase behaviors 90 days before an experiment, including recency (whether any receipt was uploaded 7, 30, or 60 days

prior to the promotion), frequency (total number of uploaded receipts), and monetary metrics (average order value per receipt and average item count per receipt), as well as preferences for the product categories or merchants, defined by the proportion of receipts including a specific product category or those associated with a particular merchant.

3. **Contextual behaviors:** Purchase behaviors related to the promoted items 90 days prior to the experiment, including recency, frequency, and monetary metrics.

Figure 2: Distribution of Average Covariate Value across the Eligible Customer Base



4.2.3 Experiments

In each of the 274 experiments, a specific promotional offer was tested on a target population (i.e., the eligible customer pool).[3] Eligible customers of each offer were randomly divided into treatment and control groups, with 10% of the participants assigned as control customers. These control customers did not receive the offer, allowing for an assessment of the promotion's causal impact. Detailed randomization checks are provided in Web Appendix C.1.[4]

[3] At any point in time, customers are also exposed to a handful of other promotional offers not subject to any experimentation. The randomization process ensured that both the treated and control groups within an experiment were exposed to comparable non-tested offers, facilitating the identification of the focal offer's causal effect.

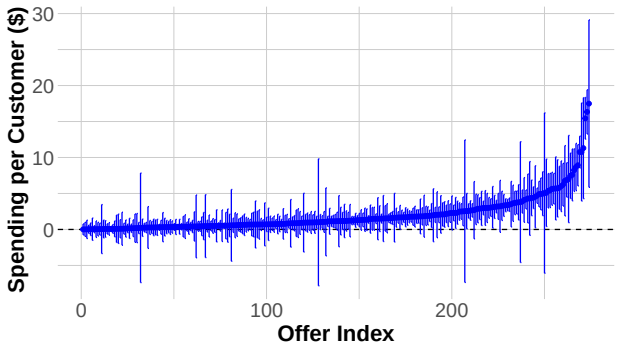
[4] Although a small portion of customers received more than one tested offer during our observation window, we believe that the potential interference is minimal. First, on average, only 7% of customers received multiple tested offers simultaneously, and none of

The focal company tests its offers on different target populations, resulting in diverse distributions of customer covariates across offers. Figure 2 displays the distribution of average values for selected covariates among the eligible customer base for each offer. For example, the top middle figure indicates that approximately 12% of the offers target dormant customers, defined as those who have not uploaded any receipts in the past 30 days. Meanwhile, the bottom middle figure reveals that around 10% of the offers specifically aim at younger customers, characterized by an average age of under 35. These patterns highlight the varying target populations for different offers.

4.3 Variations in Average Treatment Effects

The primary goal of the focal company is to increase total dollar spending on promoted items during the effective period of the offers. Therefore, this metric has been used as the outcome variable for estimating treatment effects throughout this research. Figure 3 illustrates the ATE of each offer along with its 95% confidence interval. Using a two-tailed t-test at a 5% significance level, we find that 43% of the offers have a statistically significant positive impact on dollar spending on promoted products.

Figure 3: Average Treatment Effects of 274 Promotional Offers



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

In Web Appendix G.1, we further explore how variations in ATEs correlate with observable characteristics. Overall, we find a significant positive correlation between some

these offers promoted identical items. Second, both the treatment and control groups for the focal offer were equally likely to receive offers other than the focal one, reducing the likelihood of offer interference. In Web Appendix C.2, we provide empirical evidence demonstrating that the interference from multiple treatments is minimal.

design features (such as reward points) and average values of customer covariates for each experiment (such as offer-related average order values). These results suggest the potential for the platform to utilize both design features and customer covariates to predict the treatment effects of promotional offers across customers.

5 Model Implementation and Predictive Validity

In this section, we detail the implementation of the proposed IRL model within this context and assess its performance relative to existing benchmarks in the literature.

5.1 Implementation of IRL

We implement the IRL model by calculating the doubly robust score and then developing a deep multi-task representation learning model to predict the score. (Please refer to Web Appendix D for detailed model specifications described in this section.)

5.1.1 Doubly Robust Score Estimation

We calculate the doubly robust score as follows: For each promotional offer, we construct separate conditional outcome models for the treatment and control groups using the 40 pre-treatment covariates detailed in Section 4. We use XGBoost (Chen and Guestrin 2016) with 10-fold cross-fitting to generate $\widehat{\mu}_{o,1}^{[-i]}$ and $\widehat{\mu}_{o,0}^{[-i]}$ in Equation (3), ensuring that the prediction error from these models does not introduce bias into the CATE estimation (Chernozhukov et al. 2018). For the propensity score calculation, we use the proportion of customers treated in each experiment, leveraging the random treatment assignment. These predicted outcomes and propensity scores are then incorporated into Equation (3) to calculate the doubly robust score for each offer.

5.1.2 Network Architectures and Parameter Calibration

For the encoding networks, we utilize a five-layer fully-connected neural network architecture.[5] The input layers accommodate $D = 46$ dimensions for design features and

[5]We also tested shallower and deeper network architectures in Web Appendix F and found that while the predictive accuracy is consistent across various network depths, targeting effectiveness is poorer with shallower architectures, although they still significantly outperform individual modeling.

$P = 40$ dimensions for customer covariates, respectively. Each network comprises three hidden layers, each containing 10 hidden units. The output layers of the networks produce K_1-dimensional and K_2-dimensional representations for design features and customer covariates, respectively. We set $K_1 = K_2 = 10$ because the dimensions of design features and customer covariates are similar. The Rectified Linear Unit (ReLU) activation function ($\sigma(x) = \max(0, x)$) is employed across all units. [6]

For the prediction network, we also employ a five-layer fully-connected neural network architecture. The input to this network is the concatenated representations from the design features and customer covariates, resulting in an input layer of 20 dimensions. Each of the hidden layers contains 10 hidden units. The network's output layer is tasked with generating the CATE prediction. We use the ReLU activation function for all hidden units and a linear activation function for the final prediction.

We calibrate the incrementality representation network by optimizing the loss function detailed in Equation (4) through the adaptive moment estimation (Adam) algorithm (Kingma and Ba 2014). This method is widely applied in deep learning for its ability in handling sparse gradients and adapting the learning rate for each parameter, facilitating efficient and effective model training. We also utilize mini-batch training to ensure computational efficiency and stabilize the training process.

5.2 Predictive Validity

To evaluate our model's performance relative to existing methods, we first assess its ability to accurately predict and correctly rank the treatment effects for each of the 274 offers. While this validation step is secondary to our primary objective of informing the targeting and tailoring of new offers, it aligns with established benchmarks in the field for evaluating different targeting models (Ascarza 2018, Simester et al. 2020, Yoganarasimhan et al.

[6] We find a similar targeting performance when choosing different values of K_1 and K_2 (see Web Appendix F for results when choosing $\{10, 20, 30, 40, 50\}$ dimensions). The results indicate no significant differences in performance with varying representation dimensions up to 40, while the performance typically drops when $K_1 = K_2 = 50$. We also find that deeper neural networks tend to perform slightly better with fewer representation dimensions compared to higher dimensions.

2023, Hitsch et al. 2023, Huang and Ascarza 2024). This assessment also shows that IRL can enhance targeting policies for existing offers on customers from the same populations.

5.2.1 Methods for Comparison

We evaluate the IRL model's performance by comparing it with two types of benchmarks.

Individual CATE Models. Our first benchmark involves developing *a distinct CATE model for each promotional offer*, using the doubly robust score computation method outlined in Section 5.1.1. We then utilize a deep neural network designed to predict these scores, excluding design features in the model since they are constant for each specific offer. This network consists of eight fully connected layers, with a total parameter count comparable to that of the network architecture described in Section 5.1.2.[7]

Joint CATE Models. Our second benchmark compares the proposed incrementality representation learning architecture to other joint CATE models trained using data from all 274 experiments. Specifically, we implement DR-learners using data pooled from these experiments (Ellickson et al. 2024). All deep learning models employed have a parameter count comparable to our architecture detailed in Section 5.1.2. In the main text, we consider two types of joint CATE models:[8]

1. **Deep DR-learner**: We develop a traditional deep neural network to predict the doubly robust score *without constructing separate representations for design features and customer covariates*. This benchmark enables us to quantify the benefits of the proposed model architecture compared to traditional deep learning models.
2. **Deep DR-learner with No Design Feature**: This method adopts the previous approach but *utilizes only customer covariates* for the deep neural network. This model helps evaluate the value created solely by learning across experiments, excluding the design features information, in comparison to individual modeling.

[7] We also evaluate other CATE models, including the DR-learner (Kennedy 2023) and T-learner (Künzel et al. 2019), with results showing consistent performance. Details are in Web Appendix E.

[8] Similar to individual modeling, we have estimated various types of CATE models using pooled data. Detailed information on these models and the full set of results can be found in Web Appendix E.

5.2.2 Performance metrics

We focus on two performance metrics: predictive accuracy of treatment effects and targeting effectiveness, both evaluated at the offer level. Predictive accuracy is crucial because managers often depend on the predicted treatment effect to evaluate the expected profit from an offer. Targeting effectiveness is equally important, as it informs decisions about who should receive the offer.

Predictive Accuracy. We evaluate the accuracy of a CATE model ($\widehat{\tau}$) in predicting treatment effects for each offer o. To do this, we divide the test set of each offer into $Q = 50$ equally sized groups based on their predicted CATEs. Next, we compute the average treatment effect for each group q_o based on (i) the predicted CATEs, that is, $\widehat{\tau}_{q_o} \equiv \frac{1}{|\#\{i \in q_o\}|} \sum_{i \in q_o} \widehat{\tau}(\mathbf{Z}_o, \mathbf{X}_{o,i})$, and (ii) the actual outcomes observed in the experiments, i.e.,

$$\tau_{q_o} \equiv \frac{1}{|\#\{i : i \in q_o, W_{o,i} = 1\}|} \sum_{i \in q_o, W_{o,i}=1} Y_{o,i} - \frac{1}{|\#\{i \in q_o, W_{o,i} = 0\}|} \sum_{i \in q_o, W_{o,i}=0} Y_{o,i}.$$

Finally, we compute the average root mean squared error (RMSE) across all offers, i.e., $\text{RMSE}(\widehat{\tau}) = \frac{1}{O} \sum_{o=1}^{O} \sqrt{\sum_{q_o=1}^{Q} \left(\widehat{\tau}_{q_o} - \tau_{q_o}\right)^2}$.

Targeting Performance. We assess targeting performance by analyzing each model's ability to distinguish between customers with high treatment effects and those with low treatment effects. Specifically, we choose the Area Under the Targeting Operating Characteristic curve (AUTOC) (Yadlowsky et al. 2021), a common metric that evaluate a model's ability to correctly rank individuals based on their treatment effects. To calculate that value, we first estimate the *Targeting Operating Characteristic* (TOC) for offer o, which is defined as the difference in treatment effects between individuals within the top $\phi \times 100\%$ predicted CATE tier and all individuals. That is,

$$\begin{aligned}\text{TOC}_o(\phi; \widehat{\tau}) = \mathbb{E}\left[Y_{o,i}(W_{o,i} = 1) - Y_{o,i}(W_{o,i} = 0) | F_{\widehat{\tau}}(\widehat{\tau}(\mathbf{Z}_o, \mathbf{X}_{o,i})) \geq 1 - \phi\right] - \\ \mathbb{E}\left[Y_{o,i}(W_{o,i} = 1) - Y_{o,i}(W_{o,i} = 0)\right],\end{aligned}$$

where $F_{\widehat{\tau}}$ is the cumulative distribution function of the predicted CATEs. Then, the AUTOC is defined as $\text{AUTOC}_o(\widehat{\tau}) = \int_0^1 \text{TOC}_o(\phi; \widehat{\tau}) d\phi$. Note that a model $\widehat{\tau}$ outperforms another model $\widehat{\tau}'$ in identifying customers in the top $\phi \times 100\%$ CATE group if $\text{TOC}_o(\phi; \widehat{\tau}) > \text{TOC}_o(\phi; \widehat{\tau}')$. Therefore, a higher AUTOC value suggests that the CATE model is more successful at identifying customers who demonstrate the strongest sensitivity toward the intervention, leading to more effective targeting for offer o. Finally, we compute the average AUTOC value across all the offers: $\text{AUTOC}(\widehat{\tau}) = \frac{1}{O} \sum_{o=1}^{O} \text{AUTOC}_o(\widehat{\tau})$.

5.2.3 Results

We implement a bootstrap validation scheme, similar to the one described in Ascarza (2018), to assess the performance of our model. Specifically, we create $B = 20$ pairs of training (70% of the data from each experiment) and test splits (30% of the data from each experiment) for the 274 offers. For each split b, we train a CATE model $\widehat{\tau}^{(b)}$ using the training set and then assess its performance ($\text{RMSE}(\widehat{\tau}^{(b)})$ and $\text{AUTOC}(\widehat{\tau}^{(b)})$) on the corresponding test set. This procedure is replicated B times, allowing us to calculate and report the bootstrap mean and standard deviation for key performance metrics. Table 2 presents the bootstrap means and standard deviations of average RMSE and AUTOC values for each approach, across the 274 offers.

Table 2: Predictive Validity for 274 Tested Offers

Model	RMSE	AUTOC
IRL (10-dim)	7.78 (0.82)	0.77 (0.23)
Joint-DR-DNN	8.29 (1.52)	0.45 (0.20)
Joint-DR-DNN (No Design Features)	8.62 (1.11)	0.48 (0.13)
Individual-DR-DNN	8.57 (1.41)	0.05 (0.20)

Note: We calculated the mean average RMSE and AUTOC values across 274 offers over 20 splits, with standard deviations shown in parentheses. The results demonstrate the effectiveness of models based on deep neural networks. For a complete set of results from other CATE models, please refer to Table App-2 in Web Appendix E.

Several findings are particularly noteworthy. First, the IRL approach (the first row) achieves the lowest RMSE and the highest AUTOC value, demonstrating its exceptional

predictive accuracy and targeting precision. In particular, the IRL model significantly outperforms the joint doubly robust DNN model (listed in the second row), suggesting that its superior performance is largely attributable to its network architecture. This supports the argument in Section 5.1.2 that constructing separate representations for offers and customers can enhance predictive accuracy.

Second, all the joint models that leverage information across experiments (rows one to three) outperform the method that estimates the CATE for each offer individually (the final row). Notably, the individual CATE models exhibit a significantly higher RMSE and limited targeting ability (as the mean AUTOC value is close to zero). This highlights the benefits of synthesizing knowledge from multiple experiments.

Third, the deep DR-learner with design features (the second row) performs similarly to the model without them (the third row). This suggests that the traditional deep learning architecture may not effectively capture the generalizable patterns of design features that influence treatment effect heterogeneity.

6 Application: Targeting and Tailoring New Promotions

Having demonstrated the predictive validity, we now examine how effectively the IRL model extends its predictions to untested offers and customer segments. We also show how managers can use the IRL model to enhance profitability through tailored offers. Specifically, we show that the model can identify key features and covariates from learned representations, helping managers set optimal levels of these design features for specific customer segments, thus optimizing promotional designs and boosting profitability.

6.1 Using IRL for Targeting

We explore two decision scenarios for targeting with our model. First, we assess the model's ability to select customer targets for new, untested offers. This evaluation seeks to determine the model's generalizability across different promotional designs, enabling managers to enhance the effectiveness of new offers without prior testing. Second, we

use the model to target existing offers to new, untested customers (i.e., customers who were originally ineligible for such offers). This exercise aims to evaluate the model's adaptability to new customer segments, thereby enabling managers to effectively expand the reach of existing offers.

6.1.1 Targeting for Untested Offers

We start by examining a scenario where the focal company aims to evaluate and make targeting decisions for offers that have not yet been tested. As a case study, we select 55 offers related to personal care products (one of the largest product categories in the data) as the *holdout* offers and use the remaining 219 offers (none of which promoted personal care products) to train the IRL model as well as the other joint-CATE benchmarks.[9] We compare the performance of all joint models against the traditional approach, in which the company conducts experiments for each of the 55 holdout offers and estimates individual CATE models using the actual experimental data.

To implement bootstrap validation, we first divide the actual experimental data for these 55 offers into a 70% training set and a 30% test set for each iteration. We then use the training set to estimate the individual CATE models and evaluate their performance on the test sets. For the IRL and other joint models, we train using data from the 219 offers and evaluate performance on the test set of each of the 55 holdout offers. We conduct 20 splits and report the means and standard deviations of the key performance metrics.

Table 3 presents the bootstrap means and standard deviations of average RMSE and AUTOC values for each CATE model across 55 holdout offers. Three key findings emerge. First, the proposed IRL method consistently surpasses other methods in targeting performance, demonstrating superior generalization to untested interventions. Second, the three joint models (rows one to three) significantly outperform the individual CATE models from actual experiment data (last row), highlighting the benefits of leveraging past

[9]Essentially, we are assessing whether attribute shifts and concept shifts are significant concerns in our empirical application. Regarding concept shifts, the true CATE function for one category may differ substantially from other categories. For attribute shifts, it is important to note that the distribution of design features and eligible customers for the 55 offers varies from the other offers (see Web Appendix G.2 for additional details).

experiments. Third, while RMSE values are similar across models using past experiments, the deep learning model integrating design features and customer covariates (second row) shows poorer targeting performance (smaller AUTOC) compared to the model excluding design features (third row). This confirms the advantage of separating design and customer representations for better generalization.

Table 3: Targeting Performance for Offers Related to Personal Care Products

Model	RMSE	AUTOC
IRL (10-dim)	3.28 (0.56)	0.67 (0.15)
Joint-DR-DNN	3.32 (0.56)	0.31 (0.10)
Joint-DR-DNN (No Design Features)	3.62 (0.39)	0.54 (0.11)
Actual (Individual-DR-DNN)	5.44 (0.60)	0.02 (0.14)

Note: We calculated the mean average RMSE and AUTOC values across 55 holdout offers, with standard deviations shown in parentheses. The results demonstrate the effectiveness of models based on deep neural networks. For a complete set of results from other CATE models, please refer to Table App-3 in Web Appendix E.

6.1.2 Targeting for Untested Customers

Next, we investigate a scenario in which the focal company aims to predict treatment effects and make targeting decisions for customer segments who were previously ineligible for certain offers. As a case study, we focus on "customers over the age of 40 and offers in the wine category," and consider a scenario where the company had not previously targeted this demographic for their wine-related promotions. Our IRL model uses data from past wine promotions targeted at younger customers, combined with offers from other product categories that were tested across various age groups, to predict the treatment effects of wine promotions on this older customer segment.

Specifically, we use experimental samples from wine offers involving customers over the age of 40 (from both treatment and control groups) as our *holdout* data. This dataset includes 31 distinct offers, comprising a total of 783,671 observations. Similar to Section 6.1.1, we apply the bootstrap validation approach with 20 splits for performance evaluation. We first divide the holdout data into a 70% training set and a 30% test set for

each bootstrap split. Then, the training set is used to develop the individual CATE models, whose performance is then assessed on the test sets. For IRL and other joint models, we use all the non-holdout data for training and then assess their performance on the test set derived from the holdout data.

Table 4 shows the performance metrics of various models on the holdout data. First, the IRL model (the first row) achieves the best results, while the joint model using a traditional deep learning architecture (the second row) displays comparable performance. Second, models that leverage data across experiments (rows one to three) outperform the individual CATE models for each offer (the final row). Notably, the model that excludes design features is the least effective among the joint models but still significantly outperforms the individual CATE models. These findings suggest that the IRL model and (other joint models) can enhance targeting performance on new target populations from previously tested offers by synergizing past experiments.

Table 4: Targeting Performance for Wine Offers on Customers with Age over 40

Model	RMSE	AUTOC
IRL (10-dim)	5.58 (0.77)	1.27 (0.36)
Joint-DR-DNN	5.75 (0.67)	1.16 (0.37)
Joint-DR-DNN (No Design Features)	6.27 (0.83)	0.98 (0.39)
Actual (Individual-DR-DNN)	6.65 (0.98)	0.15 (0.45)

Note: We calculated the mean average RMSE and AUTOC values using holdout data, which includes $783,671$ observations from 31 distinct offers, with standard deviations shown in parentheses. The results demonstrate the effectiveness of models based on deep neural networks. For a complete set of results from other CATE models, please refer to Table App-4 in Web Appendix E.

6.2 Using IRL for Tailoring

Typically, managers tailor their promotions through a two-step process: first, they create customer segments based on observed covariates, such as category preferences or buying behaviors. Then, they create targeted offers for each segment based on their objectives (e.g., customers who usually make smaller purchases should receive greater discounts to incentivize them to buy more) or prior beliefs (e.g., offers with stricter dollar requirements

usually work for customers who have made larger purchases). In this section, we demonstrate how the IRL model can provide a data-driven approach to this decision-making process, allowing companies to tailor promotional offers more effectively to different customer segments and enhance profitability.

6.2.1 Insights from Incrementality Representations

A key feature of the IRL model is its ability to generate low-dimensional *incrementality representations* for both offer features ($\boldsymbol{\zeta}_o$) and customer covariates ($\boldsymbol{\chi}_{o,i}$). We illustrate how managers can use these representations to identify key decision factors for optimal promotion design. Specifically, we conduct segmentation analyses of the incrementality representations, providing insights that can further inform managers' tailoring decisions. (The results discussed below are derived from one bootstrap split detailed in Section 5.2.3).

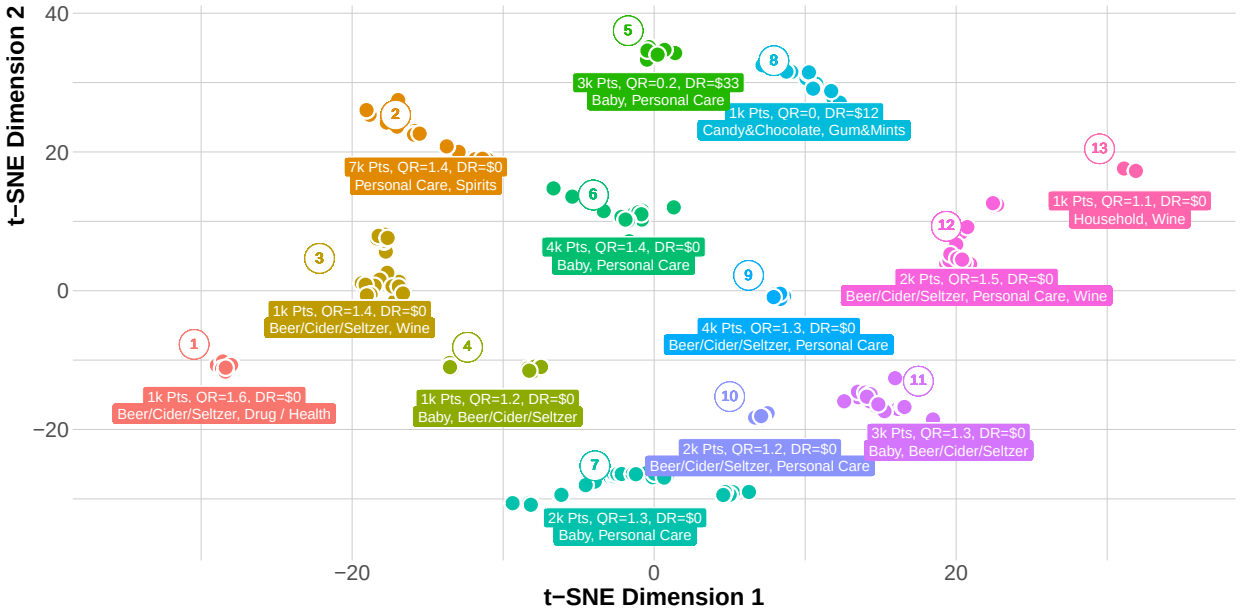
Offer Design Features. By design, offers with similar incrementality representations exhibit comparable response patterns from the same customer group (as the model generates similar treatment effects for these offers when applied to the same customer). Consequently, managers can utilize these representations to (i) identify which previously tested offers are effective for the same customer and (ii) discern the key design features influencing these patterns. With this knowledge, managers can concentrate on the design features that significantly impact incrementality representations, and therefore treatment effect heterogeneity, enabling them to tailor promotions more effectively.

To achieve this, we extract the learned representation for the design of the 274 existing offers (denoted as $\boldsymbol{\zeta}_o$ following the notation in Figure 1). We then employ t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton 2008) to project these representations into a two-dimensional space and use the DBSCAN algorithm (Ester et al. 1996) to identify "offer clusters" with similar incrementality representations. Each cluster is profiled based on the design features ($\mathbf{Z}_o$) of the offers it contains. For continuous fea-

tures (e.g., points awarded), we calculate the cluster average; for categorical features (e.g., product category), we identify the most frequently occurring labels within the cluster.

Figure 4 displays the t-SNE map and the clusters generated using incrementality representations from the proposed IRL model. Each point on the map represents a promotional offer, color-coded according to the clusters identified by the DBSCAN algorithm. Labels with colored backgrounds provide key information for each cluster, including average reward points, required quantities, required dollar amounts, and the top two promoted product categories.

Figure 4: t-SNE Maps for Learned Representations (ζ_o)



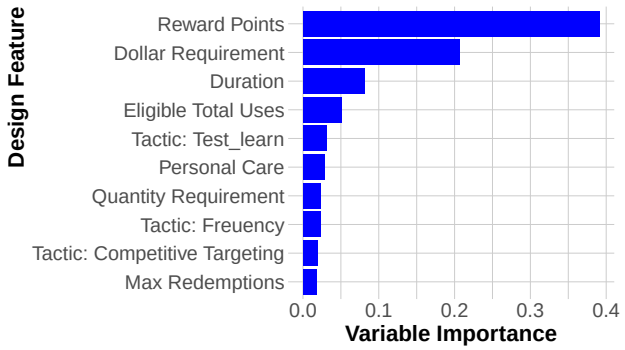
Note. Each point on the map corresponds to a promotional offer, with colors indicating the clusters determined by the DBSCAN algorithm. Labels with colored backgrounds show the average reward points (Pts, set as zero for offers requiring minimum dollar spending), required quantities (QR), required dollar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

Figure 4 suggest that clusters based on incrementality representations provide a nuanced perspective on offers with similar patterns of heterogeneous treatment effects. For instance, consider the clusters related to baby and personal care products. For these categories, the incrementality representations identify three distinct clusters (clusters 5, 6, and 7 in Figure 4). Cluster 7 is characterized by lower reward points (2,000 points) and

quantity requirement (QR = 1.3), while Clusters 5 and 6 feature higher reward points (3,000 points and 4,000 points, respectively), differing mainly in their reward claim requirements. Specifically, Cluster 6 demands a higher quantity requirement (QR = 1.4), whereas Cluster 5 requires purchases to meet a certain dollar amount (DR = $33).

To further understand what are driving the differences behind the incrementality representations, we conduct a statistical analysis to identify the features determining cluster membership. Specifically, we use XGBoost to construct a multi-class classification model that predicts the cluster membership of each offer based on its design features. Then, the ten features with the highest variable importance are shown in Figure 5. The classification models show that clusters of incrementality representations are influenced by a diverse set of factors, with reward points accounting for 40% of the importance, requirements making up 20%, and additional features such as duration, product categories, and promotional tactics collectively contributing 35%.

Figure 5: Variable Importance of Top 10 Design Features for Cluster Classification Models



Note. Variable importance is measured by the average gain in accuracy across all splits where the variable is utilized.

For comparison, we also conduct a similar analysis with the original design features ($\mathbf{Z}_o$) and present the results in Web Appendix G.3. We find that clustering offers using raw features provides less discriminative power in separating offers (Figure App-4). In partciular, clusters based on raw design features (Figure App-5) are overwhelmingly influenced by just two features: reward points, which constitute 80% of the classification outcomes, and dollar requirements, which account for 17%.

Customer Covariates. We conduct a similar analysis comparing clusters from the learned representations of customer covariates ($\boldsymbol{\chi}_{o,i}$) to those derived directly from (rescaled) customer covariates ($\mathbf{X}'_{o,i}$).[10] Given the significantly larger data size (millions of customers), we opt for K-means clustering over density-based methods to enhance computational efficiency. In this analysis, we choose three distinct customer segments for illustrative purposes. We perform this analysis using both (a) the incrementality representations from the test data covariate samples[11] and (b) the rescaled covariate values ($\mathbf{X}'_{o,i}$). Once the segments are established, we profile each segment by averaging the customers' past activities. We summarize both "general" behavior, which includes all receipts uploaded by the customer in the 90 days prior to receiving the offer, and "offer-related" behavior, which pertains specifically to receipts that included products from each focal offer. Table 5 presents both sets of results, with segments sorted by the average total number of receipts (first column).

Table 5: Summary of Covariate Segments Based on (a) Incrementality Representations ($\boldsymbol{\chi}_{o,i}$) and (b) Rescaled Covariated Values ($\mathbf{X}'_{o,i}$)

(a) Segments Based on Incrementality Representations ($\boldsymbol{\chi}_{o,i}$)

General		Offer-related		Proportion
Receipts	AOV	Receipts	AOV	
18.8	$ 38	0.34	$ 3.8	55%
73.8	$ 44	0.36	$ 2.2	40%
225.7	$ 125	0.86	$ 5.2	5%

(b) Segments Based on Scaled Covariate Values ($\mathbf{X}'_{o,i}$)

General		Offer-related		Proportion
Receipts	AOV	Receipts	AOV	
23.1	$ 37	0.31	$ 2.6	23%
51.3	$ 51	0.39	$ 3.6	49%
71.0	$ 48	0.40	$ 3.0	28%

Upon initial examination, both approaches appear to categorize customers based on their engagement level with the platform, as indicated by the disparities in the number of receipts shown in the first column of Tables 5a and 5b). However, segmentation via incrementality representations uncovers more nuanced distinctions. Notably, Table5a shows a broader spread in the total number of receipts compared to Table 5b, suggesting that incrementality representations offer a stronger discriminative power. Furthermore, the

[10] To prevent variables with larger scales from dominating the clustering results, we standardize all continuous variables using the Z-score transformation.

[11] We use the test data for this analysis as the subsequent section validates the treatment effect heterogeneity across these derived customer segments using the test set.

incrementality representations distinguish customers by average order value (AOV), as detailed in the second column of Tables 5a. In contrast, segmentation based on rescaled covariate values fails to differentiate customers by AOV, as illustrated in the second column of Table 5b. This distinction is important, given that customers with higher AOVs are more likely to meet quantity or spending requirements, making AOV a critical factor in influencing treatment effects (as illustrated in Section 6.2.2 below).

Moreover, segments derived from incrementality representations display distinct variations in offer-related behaviors, as detailed in columns 3 and 4 of Table 5a. This contrasts sharply to segments based on original covariate values, which show no significant differences in uploaded receipts for promoted items, as observed in columns 3 and 4 of Table 5b.

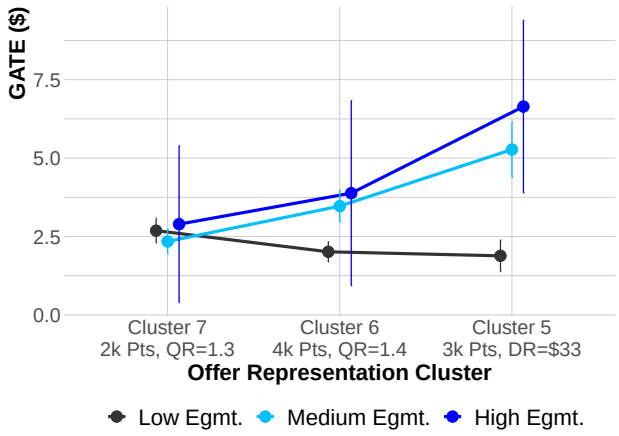
6.2.2 Assessing Treatment Effect Heterogeneity

So far, our cluster analysis has shown that incrementality representations capture more nuanced variations in design features and customer covariates. We now focus on illustrating *how treatment effect heterogeneity varies across the offer clusters and customer segments* identified by the incrementality representations. To achieve this, we label each observation in the *test data* with its corresponding offer cluster from Figure 4 and engagement segment from Table 5a. We then calculate the group average treatment effect (GATE) for each combination of offer cluster and engagement segment, using the actual outcomes from the test data.

Figure 6 shows the GATEs for three clusters associated with "baby and personal care" products across three engagement segments, illustrating distinct treatment effect patterns among different segments. First, the medium and high engagement segments generally display higher incremental spending than the low engagement segment, except in scenarios with relatively low reward points (Cluster 7), where the GATEs show no significant differences across engagement levels. Second, increasing the reward points or requirements leads to significant differences among customer segments. For the medium and high engagement segments, higher points or requirements (Clusters 5 and 6) enhance

the treatment effect, whereas they result in less effective offers for low engagement customers. Finally, dollar requirements prove more effective than quantity requirements for medium and high engagement segments, as demonstrated by higher treatment effects for offers in Cluster 5 compared to those in Cluster 6.

Figure 6: Treatment Effects by Offer Clusters and Customer Segments Derived from Incrementality Representations $(\zeta_o, \chi_{o,i})$



Note. Each point reports the group average treatment effect together with two-standard-error interval. The offer representation clusters align with those outlined in Figure 4, and the engagement segments correspond to those outlined in Table5a.

In summary, the analysis demonstrates that clustering offers and customers based on their incrementality representations from the IRL model uncovers meaningful patterns related to design features, customer behaviors, and treatment effects, which are not as evident when using the original features and covariates. By segmenting samples based on the key variables identified through these incrementality representations, companies can more effectively detect variations in treatment effects and identify key design features and customer segments driving such variations.[12]

6.2.3 Tailoring Promotions Using Segment Incrementality Plots

Building on our analysis, we identified two critical design features — reward points and dollar requirements — that influence treatment effect heterogeneity across three customer

[12]For comparison, an analysis for clusters based on original design features and scaled customer covariates is detailed in Web Appendix G.4. The results show minimal significant differences across defined clusters and segments. In contrast, as depicted in Figure 6, incrementality representations more effectively elucidate treatment effect heterogeneity.

segments. In this section, we develop an interpretable machine learning tool for the IRL model, enabling managers to tailor promotions to maximize predicted profitability rather than relying on business intuition.

Suppose a manager is considering a specific offer, like a "30-day, one-time-use baby product offer that can be combined with other offers," but is unsure about the optimal reward points and required dollar amount due to varied customer responses. We recommend using a post-hoc model interpretation method. By leveraging predictions from the IRL model, this approach can determine the optimal reward points and dollar requirements for different customer segments, tailoring the offer to maximize effectiveness.

Specifically, we introduce a model visualization framework named the *Segment Incrementality Plot* (SIP). This plot is created by calculating the predicted treatment effects for different customer segments using the IRL model, considering variations in reward points and dollar requirements for the offer. We then calculate the predicted incremental profits across segments based on these treatment effects. The construction of the SIP involves the following steps:

1. Randomly sample covariate observations from pre-defined customer segments for computational efficiency. Here, we use the low, medium, and high engagement segments identified in Table 5a for illustration (denoted as $\mathcal{S}_{\text{Low}},\ \mathcal{S}_{\text{Medium}},$ and $\mathcal{S}_{\text{High}}$).
2. Create hypothetical offers with a range of reward points $\{Z_{\text{new}}^{\text{point}}\}$ and dollar requirements $\{Z_{\text{new}}^{\text{dollar}}\}$, while keeping all other design features constant ($\overline{\mathbf{Z}}_{\text{new}}$).
3. Compute predicted treatment effects for every combination of $Z_{\text{new}}^{\text{point}}$, $Z_{\text{new}}^{\text{dollar}}$ and $\mathbf{X}_{o,i}$ within $\mathcal{S}_{\text{Low}},\ \mathcal{S}_{\text{Medium}},\ \mathcal{S}_{\text{High}}$, that is, $\widehat{\tau}\left(\widehat{\mathbf{\Phi}}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \overline{\mathbf{Z}}_{\text{new}}), \widehat{\mathbf{\Psi}}(\mathbf{X}_{o,i})\right)$.
4. Calculate the average predicted treatment effect for each combination of $Z_{\text{new}}^{\text{point}}$, $Z_{\text{new}}^{\text{point}}$ and each segment (i.e., the segment incrementality):

$$\text{SI}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \mathcal{S}_{(\cdot)}) = \frac{1}{|\mathcal{S}_{(\cdot)}|} \sum_{i \in \mathcal{S}_{(\cdot)}} \widehat{\tau}\left(\widehat{\mathbf{\Phi}}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \overline{\mathbf{Z}}_{\text{new}}), \widehat{\mathbf{\Psi}}(\mathbf{X}_{o,i})\right).$$

5. Calculate the predicted incremental profits of the new offer on each customer using $\text{SI}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \mathcal{S}_{(\cdot)})$. [13] Then, visualize $\text{Profit}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \mathcal{S}_{(\cdot)})$ for each segment.

The SIP offers significant advantages as a decision support tool. First, unlike partial dependence plots (Friedman 2001), which average effects across all customers, the SIP allows for the identification of differential interactions between customer engagement levels and offer design features. Second, in contrast to individual conditional expectation plots (Goldstein et al. 2015), which focus on predictions for each individual, the SIP provides insights at the segment level. This facilitates the interpretation of results by illustrating how different segments respond to various promotional strategies. Third, the SIP works for any predefined customer segments. Therefore, managers can apply the SIP to segments from any method and evaluate whether treatment effect heterogeneity exists for those segments based on the SIP.

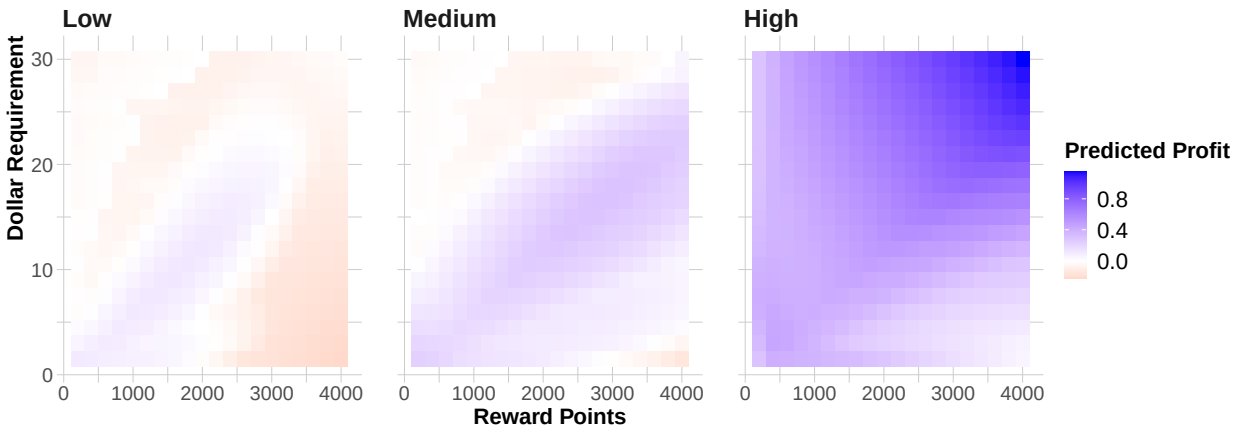
We now illustrate how to apply the SIP to tailor promotional offers for customer segments identified using incrementality representations. Figure 7 shows the SIP of profitability for a baby product offer, generated using the IRL model, where reward points range from 200 to 4000 and the dollar requirement ranges from \$1.5 to \$30. The blue zones in the plots indicate profitable regions, which differ across customer segments. (We also present the SIP of the predicted treatment effects in Web Appendix G.5.)

The SIP suggests that for the high-engagement segment (right-most figure), higher dollar requirements significantly increases the predicted incremental profit. These customers are more tolerant of higher spending thresholds, making offers with substantial rewards more profitable. In contrast, designing promotions for low- and medium-engagement customers requires a nuanced approach. Offers with substantial rewards and small spending requirements are often unprofitable for these groups, as the addi-

[13] Since profitability data were not directly available from the focal company, we employ the following formula as a proxy for profit: $\text{Profit}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \mathcal{S}_{(\cdot)}) = \text{SI}(Z_{\text{new}}^{\text{point}}, Z_{\text{new}}^{\text{dollar}}, \mathcal{S}_{(\cdot)}) \times \text{Avg. Rev.} - Z_{\text{new}}^{\text{point}} \times \text{Cost per Point} \times \text{Reward Claim Rate}(\mathcal{S}_{(\cdot)})$, where Avg. Rev. represents the average revenue from one dollar of incremental purchase, Cost per Point indicates the actual dollar cost associated with each reward point, and $\text{Reward Claim Rate}(\mathcal{S}_{(\cdot)})$ reflects the segment-specific reward claim rates, which are derived from previously observed reward claim behaviors across each engagement segment.

tional spending does not offset the promotional costs. Additionally, moderate rewards with high spending requirements tend to attract only low-value "cherry pickers," further diminishing profits in these segments.

Figure 7: Segment Incrementality Plot of Profitability for the Focal Offer Design



Note. We set the reward points within the range of $\{200, \cdots, 4000\}$ and the dollar requirement within $\{\$1.5, \cdots, \$30\}$.

Combining the results from the three figures, it becomes clear that optimal offer designs vary across segments: For the low-engagement segment, an offer of 1,600 points with a minimum $9 purchase yields a modest profit of $0.11 per customer. In the medium-engagement segment, 3,000 points with a minimum $16.5 purchase generates $0.30 profit per customer, while in the high-engagement segment, 4,000 points with a minimum $30 purchase results in $1.15 profit per customer. It is crucial to note that applying the strategy designed for the high-engagement segment to the lower segments would not only fail to generate profit but would actually incur losses of about $0.02 per customer. This underscores the importance of tailoring promotional offers to distinct customer segments to achieve optimal profit outcomes.

In conclusion, the discriminative power of the incrementality representations offers a nuanced understanding of how to optimize promotional personalization for diverse customer populations. These insights not only help understand treatment effect heterogeneity but also boost profitability by navigating the complexities of promotional design. By

leveraging these decision-support tools, marketers can implement tailored strategies that ensure each customer segment receives an optimized offer, effectively aligning marketing efforts with individual preferences and behaviors.

7 Conclusion and Future Directions

In pursuit of personalization, companies aim to target and tailor marketing interventions to maximize incremental value. Traditional experimental approaches are limited by the capacity to test only a few interventions within specific segments. To overcome this, we introduce the IRL framework. This approach leverages data from past experiments to target and tailor personalized interventions more effectively. By extracting low-dimensional representations that capture generalizable patterns in treatment effect heterogeneity, the IRL framework accurately predicts CATEs for tested interventions and generalizes to new, untested interventions and customer segments. This method helps companies address key personalization challenges, such as tailoring interventions for specific segments to maximize impact and identifying responsive segments for new interventions.

We demonstrate the superior performance of the IRL model over traditional methods through its application to CPG promotional campaigns. By synergizing past experiments, the IRL method not only enhances the accuracy of treatment effect predictions and targeting effectiveness for previously tested interventions but also adeptly extends to untested interventions or segments. Essentially, the IRL model addresses two major challenges in personalizing interventions: managing the complexity of a high-dimensional decision space and overcoming the cold-start problem for untested interventions. Through this empirical application, we also show how firms can use the learned representations to identify critical design features and customer characteristics that significantly influence the effectiveness of promotions. Furthermore, we introduce a model interpretation tool that assists companies in designing optimally tailored promotions for various customer segments, further enhancing the practical utility of the IRL framework.

While our research provides a valuable tool for researchers and managers to design personalized interventions, it also presents several limitations that open avenues for future research. First, a practical concern identified in Theorem 2 is managing both *attribute shift* and *concept shift*, which can undermine the accuracy of our IRL model when applied to untested interventions or customer segments that deviate from historical data. To mitigate these issues, one can apply active learning techniques (e.g., Jesson et al. 2021, Toth et al. 2022) and target new data collection efforts toward offers or customer segments with high uncertainty in their CATE predictions, thereby enhancing the model's generalization capabilities. Furthermore, model fine-tuning strategies in transfer learning (e.g., Alshalali and Josyula 2018, Shen et al. 2021) could be employed to refine predictions based on newly collected, small-scale experiments, ensuring the model's continued accuracy in dynamic environments.

Second, beyond improving generalizability through new data collection, future research could also investigate optimal strategies for determining which interventions should and should not be learned together, and what information should be transferred between them. Negative transfer can occur when interventions that elicit markedly different responses from the same customers are combined in learning processes (Wang et al. 2019); in such cases, learning from individual experiments may outperform learning across multiple experiments. The existing literature (e.g., Ying et al. 2018, Jang et al. 2019) provides potential strategies to mitigate this issue by specifying (a) which parameters should be shared across experiments, and (b) which features or covariates should be used for experiment, and to what extent this should be transferred. Integrating these approaches into our IRL model could further improve its accuracy and targeting effectiveness.

Third, while we provide a flexible tool for managers to interpret model behaviors, it currently supports only post-hoc exploration of model predictions at the segment level. Future research could investigate the application of various explainable machine learning tools to enable managers to better understand treatment effect heterogeneity. For

instance, one could assess how different design features and customer covariates most significantly contribute to the treatment effect using frameworks such as feature attribution (e.g., Ribeiro et al. 2016, Lundberg and Lee 2017) or counterfactual explanation (e.g., Pawelczyk et al. 2020, Albini et al. 2022), thereby providing deeper insights into the behaviors of the IRL model.

Finally, while we have shown the benefits of synergizing experiments in a promotional setting, exploring this approach's applicability across various marketing applications could be highly valuable. For instance, an e-commerce company could use insights from email or push notification experiments to gauge customer responsiveness to promotional offers. Additionally, responses to different product recommendation algorithms could help better predict the treatment effects of various email communications and tailor them more effectively. Investigating the effectiveness of our framework in diverse contexts could reveal broader implications for when and how personalized marketing strategies can significantly benefit from integrated experimental insights.

We hope that these limitations inspire further research on synergizing experiments and incrementality representation learning.

References

Albini E, Long J, Dervovic D, Magazzeni D (2022) Counterfactual shapley additive explanations. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1054–1070.

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).

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

Arora N, Henderson T (2007) Embedded premium promotion: Why it works and how to make it more effective. *Marketing Science* 26(4):514–531.

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.

Balestriero R, Pesenti J, LeCun Y (2021) Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485* .

Bárány I, Füredi Z (1988) On the shape of the convex hull of random points. *Probability Theory and Related Fields* 77:231–240.

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.

Baxter J (2000) A model of inductive bias learning. *Journal of artificial intelligence research* 12:149–198.

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.

Ben-David S, Schuller R (2003) Exploiting task relatedness for multiple task learning. *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, 567–580 (Springer).

Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, et al. (2021) On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* .

Bonnasse-Gahot L (2022) Interpolation, extrapolation, and local generalization in common neural networks. *arXiv preprint arXiv:2207.08648* .

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 T, Guestrin C (2016) Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

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.

Chintagunta PK, Huang L, Miao W, Zhang W (2023) Measuring seller response to buyer-initiated disintermediation: Evidence from a field experiment on a service platform. *Available at SSRN 4423917* .

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

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.

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

Degtiar I, Rose S (2023) A review of generalizability and transportability. *Annual Review of Statistics and Its Application* 10:501–524.

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.

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).

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

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

Goldstein A, Kapelner A, Bleich J, Pitkin E (2015) Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics* 24(1):44–65.

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

Golrezaei N, Nazerzadeh H, Rusmevichientong P (2014) Real-time optimization of personalized assortments. *Management Science* 60(6):1532–1551.

Gordon BR, Moakler R, Zettelmeyer F (2023) Predictive incrementality by experimentation (pie) for ad measurement. *arXiv preprint arXiv:2304.06828* .

Hill J, Su YS (2013) Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics* 1386–1420.

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).

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

Jang Y, Lee H, Hwang SJ, Shin J (2019) Learning what and where to transfer. *International Conference on Machine Learning*, 3030–3039 (PMLR).

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`.

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

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).

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.

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.

Pawelczyk M, Broelemann K, Kasneci G (2020) Learning model-agnostic counterfactual explanations for tabular data. *Proceedings of The Web Conference 2020*, 3126–3132.

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 .

Ribeiro MT, Singh S, Guestrin C (2016) " why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

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

Shen Z, Liu Z, Qin J, Savvides M, Cheng KT (2021) Partial is better than all: Revisiting fine-tuning strategy for few-shot learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9594–9602.

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.

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

Toth C, Lorch L, Knoll C, Krause A, Pernkopf F, Peharz R, Von Kügelgen J (2022) Active bayesian causal inference. *Advances in Neural Information Processing Systems* 35:16261–16275.

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.

Vafaeikia P, Namdar K, Khalvati F (2020) A brief review of deep multi-task learning and auxiliary task learning. *arXiv preprint arXiv:2007.01126* .

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

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.

Wang Z, Dai Z, Póczos B, Carbonell J (2019) Characterizing and avoiding negative transfer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11293–11302.

Watkins A, Ullah E, Nguyen-Tang T, Arora R (2024) Optimistic rates for multi-task representation learning. *Advances in Neural Information Processing Systems* 36.

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 (2021) Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966* .

Ye Z, Zhang Z, Zhang D, Zhang H, Zhang RP (2023) Deep learning based causal inference for large-scale combinatorial experiments: Theory and empirical evidence. *Available at SSRN 4375327* .

Ying W, Zhang Y, Huang J, Yang Q (2018) Transfer learning via learning to transfer. *International Conference on Machine Learning*, 5085–5094 (PMLR).

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.

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 J, Wedel M (2009) The effectiveness of customized promotions in online and offline stores. *Journal of marketing research* 46(2):190–206.

Zhang WW, Misra S (2022) 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.

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 Theoretical Guarantees

In this appendix, we provide the theoretical guarantees for the proposed deep IRL framework.

Web Appendix A.1 Formalizing the IRL Problem

We begin by formalizing the incrementality representation learning problem. First, we assume that the true CATE function can be decomposed into two distinct functions: the representation function, which transforms raw design features and customer covariates into a representation, and the prediction function, which maps this representation to the actual CATE. This assumption is articulated as follows:

Assumption App-1 (Data Generating Process) *Assume that the true CATE function defined in Equation* (1) *can be written as*

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

where $\mathbf{h}^\star : \mathcal{Z} \times \mathcal{X} \to \mathbb{R}^K$ *denotes the representation function and* $f^\star : \mathbb{R}^K \to [-r, r], r < \infty$ *is the prediction function for each offer. Also, assume that the probability of the firm choosing a specific design feature* $\mathbf{Z} \in \mathcal{Z}$ *and observing specific customer covariate values* $\mathbf{X} \subset \mathcal{X}$ *are non-zero for an experiment.*

Note that this assumption does not require the existence of more predictive low-dimensional representations. Specifically, if the original design features and customer covariates already constitute the most meaningful representation, then we have $\mathbf{h}^\star(\mathbf{Z}_o, \mathbf{X}_{o,i}) = (\mathbf{Z}_o, \mathbf{X}_{o,i})$.

Assume that we have access to an unbiased proxy $\widetilde{\tau}_{o,i}$ for the actual CATE (e.g., estimated through the doubly robust score in Theorem 1). Without loss of generality and for simplicity, we assume that each past experiment has the same sample size N. The objective of IRL is to construct $\widehat{\mathbf{h}}$ and $\widehat{f}$ that best recover $\mathbf{h}^\star$ and $f^\star$ by learning from $\widetilde{\tau}_{o,i}$, that is,

$$(\widehat{f}, \widehat{\mathbf{h}}) = \underset{f \in \mathcal{F},\, \mathbf{h} \in \mathcal{H}}{\arg\min} \; \frac{1}{ON} \sum_{o=1}^{O} \sum_{i \in \mathcal{C}_o^E} \ell\left(f \circ \mathbf{h}(\mathbf{Z}_o, \mathbf{X}_{o,i}), \widetilde{\tau}_{o,i}\right). \tag{App-1}$$

Here, ℓ represents a loss function, while $\mathcal{H}$ and $\mathcal{F}$ denote the classes for the representation and prediction models, respectively (such as deep neural networks). In the following analysis, we focus on the case where $\ell(y, y') = (y - y')^2$ is the squared-loss function.

Web Appendix A.2 Model Complexity

Before establishing the bound on the prediction error, we first introduce a widely used metric to assess the complexity of a function class in machine learning theory, known as *Gaussian complexity* (Bartlett and Mendelson 2002). This measure quantifies the capacity of a model class to learn various patterns from data.

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, $\bar{\mathbf{x}} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)$, the* empirical Gaussian complexity *of the function class is defined as*

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

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

Essentially, Gaussian complexity measures how closely a function class $\mathcal{Q}$ can correlate with white noise. A higher Gaussian complexity signifies a greater ability of the function class to fit random noise, indicating an increased risk of overfitting.

Web Appendix A.3 Bounding the Prediction Error of IRL

We now provide the prediction error bound for the IRL problem and its specific application to IRL models utilizing deep neural networks.

Prediction Error Bound for Generic IRL Model. We characterize the prediction error as the average mean squared error between the predictions of an IRL model and the true CATE:

$$\text{Error}\,(f, \mathbf{h}) = \mathbb{E}\left[\frac{1}{O} \sum_{o=1}^{O} \ell\left(f \circ \mathbf{h}(\mathbf{Z}_o, \mathbf{X}_{o,i}), f^\star \circ \mathbf{h}^\star(\mathbf{Z}_o, \mathbf{X}_{o,i}) \right) \right].$$

Throughout the analysis, we impose the following standard regularity assumptions on the loss function and CATE functions.

Assumption App-2 (Assumption 1 in Tripuraneni et al. (2020))

1. *The loss function ℓ is nonnegative and B-bounded, i.e., $0 \leq \ell(y', y) \leq B$ for any $y, y' \in \mathcal{R}$.*
2. *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 $\mathbf{a}, \mathbf{a}' \in dom(f)$.*
3. *The composed function $f \circ \mathbf{h}$ is D-bounded, i.e., $\sup_{\mathbf{Z},\mathbf{X}} |f \circ \mathbf{h}(\mathbf{Z}, \mathbf{X})| \leq D$.*

We now present the upper bound for the prediction error for a generic IRL model.

Theorem App-1 (Upper Bound for Prediction Error) *Assuming that (i) the regularity conditions in Assumption App-2 hold, and (ii) $f^\star \in \mathcal{F}$ and $\mathbf{h}^\star \in \mathcal{H}$[1], then with probability at least $1 - \delta$, the prediction error of the IRL model $(\widehat{f}, \widehat{\mathbf{h}})$ derived from* (App-1) *can be bounded as follows:*

$$\text{Error}\left(\widehat{f}, \widehat{\mathbf{h}}\right) \leq \underbrace{\frac{C_1}{(ON)^2} + \log(ON)\left[C_2 \mathfrak{G}_{ON}(\mathcal{H}) + \frac{1}{\sqrt{O}}\overline{\mathfrak{G}}_{ON}(\mathcal{F})\right]}_{\textit{Overfitting Potential for the Model Classes}} + C_3\sqrt{\frac{\log(2/\delta)}{ON}}, \quad \text{(App-2)}$$

where C_1, C_2, and C_3 are some constants.

Proof. See Web Appendix B.3. ■

Theorem App-1 indicates that the prediction error of an IRL model is limited by the Gaussian complexity related to (i) the class of potential representation functions ($\mathcal{H}$) and (ii) the class of potential prediction functions ($\mathcal{F}$). Therefore, for the IRL model to accurately replicate the true functions $(f^\star, \mathbf{h}^\star)$ with large O (number of past experiments) and N (sample size of each experiment), it is crucial that the Gaussian complexity of both the representation and prediction model classes declines to zero at a sufficiently fast rate.

Prediction Error Bound for DNN-based IRL Model. Next, we present the prediction error bound for the model classes constituted by depth-d vector-valued neural networks. Specifically, a neural network is formulated as:

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

[1] Although this assumption might appear strong at first glance, it is actually reasonable for deep neural networks. Previous research has demonstrated the universal approximation property of neural networks (e.g., Cybenko 1989, Hornik 1991, Pinkus 1999, Liang and Srikant 2016, Kidger and Lyons 2020), highlighting their capability to approximate a wide variety of function classes effectively.

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-1 (Prediction Error Bound of Deep IRL) *Consider the case when $\mathcal{H}$ is the class of d_1-layer neural networks with 1-Lipschitz activation functions and K'-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-1) *can be bounded as follows:*

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

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

Proof. See Web Appendix B.4. ■

Corollary App-1 elucidates several critical insights regarding the application of deep neural networks within the IRL framework. Firstly, it establishes a theoretical guarantee, indicating that the upper bound of the prediction error for a deep learning-based IRL model diminishes to zero as both O (the number of offers) and N (the sample size) increase. Secondly, it highlights the advantage of dimension reduction in deep representation learning. The prediction error bound improves with a reduction in K', provided the neural network of K' dimensions can accurately capture the true representation function $\mathbf{h}^\star$.

Thirdly, it demonstrates the benefits of utilizing separate network structures for offer and customer representations. Specifically, such neural network architectures often result in many zero values in the parameter matrix $\mathbf{W}_j$ that links the nodes between the offer representation and customer representation networks. In scenarios with separate network structures, the maximum possible value of the Frobenius norm of $\mathbf{W}_j$ is generally lower compared to a fully connected network that permits interactions between design features and customer covariates, assuming the same number of nodes. Consequently, the prediction error bound is tighter for models with separate network structures as opposed to fully connected configurations, maintaining equal depth and node count in each

layer. This rationale extends to the comparison between shared-parameter prediction networks and offer-specific prediction networks, with the former typically achieving more favorable error bounds under similar conditions.

Web Appendix B Proofs

Web Appendix B.1 Proof for Identification Results in Equation (2)

By the unconfoundedness and no interference assumption, we have

$$\mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}|W_{o,i}=1,\mathbf{X}_{o,i}]=\mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}(\mathbf{Z}_o)|\mathbf{X}_{o,i}],$$
$$\mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}|W_{o,i}=0,\mathbf{X}_{o,i}]=\mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}(\mathbf{0})|\mathbf{X}_{o,i}].$$

The above equations hold for all the customers in the target population $\mathcal{C}_o^E$ by the overlap assumption.

Then, by the stability assumption, we have

$$\mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}(\mathbf{Z}_o)|\mathbf{X}_{o,i}]=\mathbb{E}_{i\in\mathcal{C}_o}[Y_{o,i}(\mathbf{Z}_o)|\mathbf{X}_{o,i}],$$
$$\mathbb{E}_{i\in\mathcal{C}_o^E}[Y_{o,i}(\mathbf{0})|\mathbf{X}_{o,i}]=\mathbb{E}_{i\in\mathcal{C}_o}[Y_{o,i}(\mathbf{0})|\mathbf{X}_{o,i}].$$

Combining these two results, we prove the identification results in Equation (2).

Web Appendix B.2 Proof for Theorem 1

Under Assumption 1, we have $\mu_{o,1}(\mathbf{X}_{o,i})=\mathbb{E}_{\mathcal{C}_o}[Y_{o,i}(\mathbf{Z}_o)|\mathbf{X}_{o,i}]$ and $\mu_{o,0}(\mathbf{X}_{o,i})=\mathbb{E}_{\mathcal{C}_o}[Y_{o,i}(\mathbf{0})|\mathbf{X}_{o,i}]$. Therefore, we have

$$\tau(\mathbf{Z}_o,\mathbf{X}_{o,i})=\mu_{o,1}(\mathbf{X}_{o,i})-\mu_{o,0}(\mathbf{X}_{o,i}).$$

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

$$\mathbb{E}_{\mathcal{C}_o}\left[W_i|\mathbf{X}_{o,i},\mathbf{Z}_o\right]=\mathbb{P}_{\mathcal{C}_o}\left[W_i=1|\mathbf{X}_{o,i},\mathbf{Z}_o\right]\cdot 1+0=\pi_o(\mathbf{X}_{o,i}).$$

Under Assumption 1, we have

$$\mathbb{E}_{\mathcal{C}_o^E}\left[Y_{o,i}|\mathbf{X}_{o,i},\mathbf{Z}_o\right]=\mu_{o,W_{o,i}}(\mathbf{X}_{o,i}).$$

These two results implies that

$$\mathbb{E}\left[\frac{W_{o,i}-\pi_o(\mathbf{X}_{o,i})}{\pi_o(\mathbf{X}_{o,i})[1-\pi_o(\mathbf{X}_{o,i})]}\left[Y_{o,i}-\mu_{o,W_{o,i}}(\mathbf{X}_{o,i})\right]\middle|\,\mathbf{X}_{o,i},\mathbf{Z}_o\right]=0.$$

Finally, combining these results proves Theorem 1.

Web Appendix B.3 Proof for Theorem App-1

We now prove Theorem App-1, which follows closely the proof for Theorem 1 in Tripuraneni et al. (2020). For a given f and $\mathbf{h}$, we write the population loss as

$$l(f,\mathbf{h},f^\star,\mathbf{h}^\star)=\frac{1}{O}\,\mathbb{E}\left[\sum_{o=1}^{O}\ell\left(f\circ\mathbf{h}(\mathbf{Z}_o,\mathbf{X}_{o,i}),\widetilde{\tau}_{o,i}\right)-\ell\left(f^\star\circ\mathbf{h}^\star(\mathbf{Z}_o,\mathbf{X}_{o,i}),\widetilde{\tau}_{o,i}\right)\right],$$

where the expectation is taken over the data collection mechanism for past experiments.

Claim. $l(f,\mathbf{h},f^\star,\mathbf{h}^\star)=\frac{1}{O}\,\mathbb{E}\left[\sum_{o=1}^{O}\ell\left(f\circ\mathbf{h}(\mathbf{Z}_o,\mathbf{X}_{o,i}),f^\star\circ\mathbf{h}^\star(\mathbf{Z}_o,\mathbf{X}_{o,i})\right)\right]$.

Proof. We first note that

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

for given f and $\mathbf{h}$ since $\widetilde{\tau}_{o,i}$ is unbiased. Similarly, we can show that

$$\mathbb{E}\left[\ell\left(f^\star\circ\mathbf{h}^\star(\mathbf{Z}_o,\mathbf{X}_{o,i}),\widetilde{\tau}_{o,i}\right)|\mathbf{Z}_o,\mathbf{X}_{o,i}\right]=\mathrm{Var}\left[\widetilde{\tau}_{o,i}|\mathbf{Z}_o,\mathbf{X}_{o,i}\right].$$

Combining these two observations, we have

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

By taking expectation of the above results for $\mathbf{Z}_o,\mathbf{X}_{o,i}$, we can prove the claim. ■

Next, we can define the empirical loss as:

$$\widehat{l}(f, \mathbf{h}, f^\star, \mathbf{h}^\star) = \frac{1}{ON} \sum_{o=1}^{O} \sum_{i \in \mathcal{C}_o^E} \ell\left(f \circ \mathbf{h}(\mathbf{Z}_o, \mathbf{X}_{o,i}), \widetilde{\tau}_{o,i}\right) - \mathbb{E}\left[\ell\left(f^\star \circ \mathbf{h}^\star(\mathbf{Z}_o, \mathbf{X}_{o,i}), \widetilde{\tau}_{o,i}\right)\right].$$

Now, let $\widehat{f}$ and $\widehat{\mathbf{h}}$ be derived from the empirical risk minimization problem in (App-1) of the main document. Then, we have

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

Note that the inequality holds as $b < 0$ by the fact that $\widehat{f}$ and $\widehat{\mathbf{h}}$ are derived by minimizing $\widehat{l}(\widehat{f}, \widehat{\mathbf{h}}, f^\star, \mathbf{h}^\star)$.

We can now establish bounds for the population prediction error of the estimator. Using the standard generalization bounds based on Rademacher complexity (Bartlett and Mendelson 2002, Wainwright 2019), we can bound the two components a and c as:

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

with probability at $1 - \delta$, where $\mathcal{F}(\mathcal{H}) = \{(f, \mathbf{h}) : f \in \mathcal{F} \text{ and } \mathbf{h} \in \mathcal{H}\}$, and $\mathfrak{R}_{ON}(\ell(\mathcal{F}(\mathcal{H})))$ 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 (Wainwright 2019).

Then, by Inequality (11) in the Appendix of Tripuraneni et al. (2020), we have

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

Note that $\mathfrak{R}_{ON}(\mathcal{F}(\mathcal{H})) \leq \sqrt{\frac{\pi}{2}}\mathfrak{G}_{ON}(\mathcal{F}(\mathcal{H}))$. Combining the above results together, we can establish the upper bound for the population prediction error:

$$\begin{aligned} l(\widehat{f}, \widehat{\mathbf{h}}, f^\star, \mathbf{h}^\star) &\leq a + c \leq 4\mathfrak{R}_{ON}(\ell(\mathcal{F}(\mathcal{H}))) + 4B\sqrt{\frac{\log(2/\delta)}{ON}} \\ &\leq 8L\mathfrak{R}_{ON}(\mathcal{F}(\mathcal{H})) + 8B\sqrt{\frac{\log(2/\delta)}{ON}} \\ &\leq 16L\mathfrak{G}_{ON}(\mathcal{F}(\mathcal{H})) + 8B\sqrt{\frac{\log(2/\delta)}{ON}}. \end{aligned} \tag{App-3}$$

The last inequality holds since the Rademacher complexity is upper bounded by the Gaussian complexity (Ledoux and Talagrand 2013): $\mathfrak{R}_{ON}(\mathcal{F}(\mathcal{H})) \leq \sqrt{\frac{\pi}{2}}\mathfrak{G}_{ON}(\mathcal{F}(\mathcal{H})) < 2\mathfrak{G}_{ON}(\mathcal{F}(\mathcal{H}))$.

Now, we can further decompose the Gaussian complexity based on Theorem 7 in Tripuraneni et al. (2020). Note that the Gaussian complexity is $\mathfrak{G}_{ON}(\mathcal{F}(\mathcal{H}))$ in our setting instead of $\mathfrak{G}_{ON}(\mathcal{F}^{\otimes O}(\mathcal{H}))$ in Tripuraneni et al. (2020). Therefore, when bounding the covering number for the Dudley entropy integral bound (page 18 of Tripuraneni et al. (2020)), we calculate the cadinality for the covering $\mathcal{C}_{\mathcal{F}(\mathcal{H})}$ of the set $\mathcal{F}(\mathcal{H})$, rather than the covering $\mathcal{C}_{\mathcal{F}^{\otimes O}(\mathcal{H})}$ of the set $\mathcal{F}^{\otimes O}(\mathcal{H})$. Therefore, the last inequality in page 18 of Tripuraneni et al. (2020) becomes:

$$\log N_{2,\mathbf{X}}\left(\varepsilon_1 \cdot \mathcal{L}(\mathcal{F}) + \varepsilon_2, d_{2,\mathbf{X}}, \mathcal{F}(\mathcal{H})\right) \leq \log N_{2,\mathbf{X}}\left(\varepsilon_1, d_{2,\mathbf{X}}, \mathcal{H}\right) + \max_{\mathbf{z}\in\mathcal{Z}} \log N_{2,\mathbf{z}}\left(\varepsilon_2, d_{2,\mathbf{z}}, \mathcal{F}\right).$$

Applying this result to the rest of proof for Theorem 7 in Tripuraneni et al. (2020) gives the following inequality:

$$\mathfrak{G}_{ON}(\mathcal{F}(\mathcal{H})) \leq 64\frac{D}{(ON)^2} + 128C(\mathcal{F}(\mathcal{H}))\log(ON), \tag{App-4}$$

where $C(\mathcal{F}(\mathcal{H})) = L\mathfrak{G}_{ON}(\mathcal{H}) + \frac{1}{\sqrt{O}}\max_{\mathbf{q}\in\{\mathbf{h}(\mathbf{Z}_o,\mathbf{X}_{o,i}):\mathbf{h}\in\mathcal{H}}\ \widehat{\mathfrak{G}}_{\mathbf{q}}(\mathcal{F})$.

Finally, combining (App-4) and (App-3) gives

$$\text{Error}\left(\widehat{f}, \widehat{\mathbf{h}}\right) \leq \underbrace{\frac{C_1}{(ON)^2} + \log(ON)\left[C_2\mathfrak{G}_{ON}(\mathcal{H}) + \frac{1}{\sqrt{O}}\overline{\mathfrak{G}}_{ON}(\mathcal{F})\right]}_{\text{Overfitting Potential for the Model Classes}} + C_3\sqrt{\frac{\log(2/\delta)}{ON}},$$

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

Web Appendix B.4 Proof for Corollary App-1

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

$$\widehat{\mathfrak{R}}_{ON}(\mathcal{NN}) \leq \frac{2K\sqrt{d+1+\log(r)}\prod_{j=1}^{j} M(j)}{\sqrt{ON}}, \tag{App-5}$$

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 $\widehat{\mathfrak{G}}_{ON}(\mathcal{NN}) \leq 2\sqrt{\log(ON)}\,\widehat{\mathfrak{R}}_{ON}(\mathcal{NN})$ (page 97 in Ledoux and Talagrand (2013)) together with (App-5) and taking expectation on the left side, we have

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

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:

$$\mathfrak{G}_{ON}(\mathcal{H}) \leq A\frac{\sqrt{\log(ON)}K'\prod_{j=1}^{j} M(j)}{\sqrt{ON}}, \overline{\mathfrak{G}}_{ON}(\mathcal{F}) \leq B\sqrt{d+1+\log(K')}\frac{\sqrt{\log(ON)}\prod_{j=1}^{j} M(j)}{\sqrt{ON}}.$$

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

Web Appendix B.5 Proof for Theorem 2

To prove the theorem, we first note that

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

For the first term, by the reverse triangle inequality — $|d(x,y) - d(x,z)| \leq d(y,x)$ 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

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

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

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

Combining these results together proves the theorem.

Web Appendix C Randomization and Interference Checks

Web Appendix C.1 Randomization Check

We perform covariate balance checks across 274 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 10,960 (experiment, covariate) pairs, only 0.3% exhibit SMDs greater than 0.2, and 1% show SMDs exceeding 0.1. This indicates that the randomization process carried out by the focal company was properly implemented.

Table App-1: Standarized Mean Differences Across Customer Covariates

Variable	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
receipt_count	0.001	0.013	-0.051	-0.005	0.007	0.062
aov	0.002	0.012	-0.047	-0.005	0.009	0.046
item_per_order	0.003	0.012	-0.041	-0.003	0.010	0.041
has_past_7	0.001	0.006	-0.016	-0.001	0.004	0.023
has_past_30	0.001	0.004	-0.012	-0.001	0.003	0.014
has_past_60	0.001	0.003	-0.011	-0.001	0.002	0.013
offer_receipt_count	0.034	0.060	-0.066	0.000	0.058	0.340
offer_aov	0.040	0.067	-0.059	0.000	0.067	0.320
offer_item_per_order	0.044	0.075	-0.063	0.000	0.074	0.400
offer_has_past_7	0.008	0.015	-0.022	0.000	0.011	0.069
offer_has_past_30	0.014	0.026	-0.016	0.000	0.021	0.130
offer_has_past_60	0.018	0.034	-0.019	0.000	0.027	0.170
7.ELEVEN	0.000	0.013	-0.051	-0.006	0.007	0.053
AMAZON	0.002	0.015	-0.063	-0.006	0.009	0.058
COSTCO	-0.001	0.012	-0.032	-0.008	0.004	0.034
CVS	0.001	0.015	-0.056	-0.005	0.007	0.056
KROGER	0.001	0.014	-0.038	-0.007	0.009	0.072
MCDONALD'S	-0.001	0.013	-0.056	-0.007	0.005	0.074
TARGET	0.001	0.013	-0.045	-0.007	0.008	0.051
THE HOME DEPOT	-0.001	0.015	-0.063	-0.008	0.007	0.050
WALGREENS	0.002	0.012	-0.039	-0.005	0.009	0.042
WALMART	0.002	0.013	-0.043	-0.005	0.009	0.047
Appetizers & Sides	0.001	0.014	-0.074	-0.006	0.008	0.067
Bakery & Bread	-0.001	0.012	-0.049	-0.007	0.006	0.042
Bath & Body	0.002	0.014	-0.039	-0.005	0.010	0.049
Beer, Wine & Spirits	0.002	0.013	-0.042	-0.005	0.009	0.048
Beverages	0.002	0.013	-0.046	-0.006	0.009	0.055
Candy	0.001	0.013	-0.046	-0.006	0.009	0.048
Cat	0.003	0.014	-0.056	-0.003	0.010	0.051
Dairy	0.000	0.013	-0.057	-0.006	0.006	0.039
Dog	0.001	0.014	-0.039	-0.005	0.008	0.058
Hair Care	0.001	0.014	-0.052	-0.006	0.009	0.046
Household, Paper & Plastic	0.001	0.012	-0.039	-0.006	0.008	0.042
Produce	0.000	0.012	-0.037	-0.007	0.007	0.040
is_referral	0.001	0.007	-0.021	-0.003	0.004	0.037
age	0.000	0.014	-0.046	-0.009	0.009	0.049
tenure	0.002	0.016	-0.052	-0.007	0.012	0.061
is_male?	0.000	0.005	-0.015	-0.003	0.003	0.022
is_femare?	0.000	0.006	-0.023	-0.003	0.004	0.020
is_non_binary?	0.000	0.002	-0.008	-0.001	0.001	0.011

Web Appendix C.2 Testing the No Interference Assumption

This appendix presents empirical evidence for minimal interference caused by multiple offers. Since we observe a correlation between the number of offers a customer receives and their purchasing behavior (likely due to the eligibility criteria specified by the company), we apply the double machine learning framework (Chernozhukov et al. 2018) to examine if the number of additional offers a customer receives causally affects treatment effects.

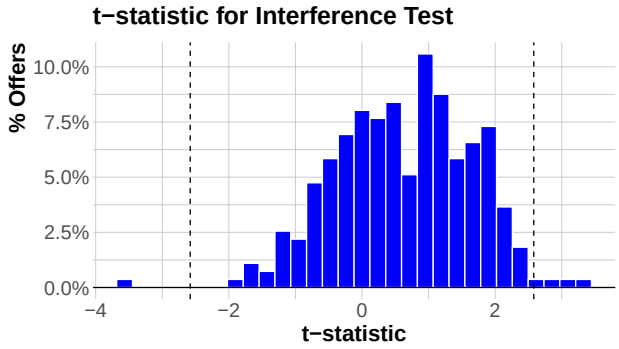
Specifically, for every offer, we first calculate the doubly robust score $\widetilde{\tau}_{o,i}$ in the same way as in the IRL model. Following this, we apply the debiased machine learning approach to estimate the following partially linear model:

$$\widetilde{\tau}_{o,i} = \beta_o N_{o,i} + m(\mathbf{X}_{o,i}) + \varepsilon_{o,i},$$
$$N_{o,i} = g(\mathbf{X}_{o,i}) + \eta_{o,i},$$

where $N_{o,i}$ denotes the number of effective offers available to customer i when offer o was effective. Here, we use XGBoost with default settings as implemented in the `DoubleML` package (Bach et al. 2022) with five-fold cross-fitting to estimate the control functions m and g. If interference across offers is absent, we would anticipate that $\beta_o = 0$ for most offers.

Figure App-1 presents the distribution of t-statistics for the estimated β_o across 274 offers. It is noteworthy that only 1.4% of offers have a significantly non-zero at a 1% significance level, suggesting that potential interference is minimal in our empirical context.

Figure App-1: t-statistics for Interference Test across 274 Offers



Note. The dashed line indicates the threshold for rejecting the null hypothesis that $\beta_o = 0$ at a 1% significance level.

Web Appendix D Model Specification for Main Analysis

Web Appendix D.1 Doubly Robust Score Estimation

To construct the doubly robust score for each offer, we estimate the nuisance models $\widehat{\mu}_{o,1}^{[-i]}$ and $\widehat{\mu}_{o,0}^{[-i]}$ using XGBoost with ten-fold cross-fitting. For tuning parameters, we utilize the cross-validation function in the R XGBoost package (Chen and Guestrin 2016) to select the depth of a single tree from $\{3, 4, 5\}$, the learning rate $\eta \in \{0.5, 1, 1.5\}$, and the number of boosting rounds from $\{20, 25, 30\}$. The results suggest that a depth of 3, $\eta = 1$, and 30 rounds yield the best predictive accuracy. However, the results are relatively insensitive to these tuning parameters. For propensity scores, since the treatment assignment is random, we simply use the percentage of customers receiving the offer as the propensity score for each offer.

Web Appendix D.2 Deep Learning Models

As described in Section 5.1.2, we implement two five-layer fully-connected neural networks for representation learning and one five-layer fully-connected neural network for prediction. Each hidden layer contains ten nodes. With the representation $K_1 = K_2 = 10$ the IRL model comprises a total of 1,981 parameters. When $K_1 = K_2 = 50$, the model has 2,401 parameters in total.

For joint models in Section 5.2.1 (i.e., the Deep DR-learners with and without design features), we employ an eight-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 2,291 parameters for the Deep DR-learner with design features and 1,491 parameters for the Deep DR-learner without design features. For individual modeling, we employ the same neural network architecture as the deep DR-learner without design features, for each offer.

To calibrate deep neural networks, we use mini-batch stochastic gradient descent with a batch size of 100 for joint modeling and 30 for individual modeling. We adaptively determine the learning rate using the adadelta algorithm (Zeiler 2012) implemented in Keras (Chollet et al. 2015). We train the neural networks for 10 epochs, as the validation performance stabilizes after this training period.

Web Appendix E Targeting Performance with Additional Models

In this appendix, we expand upon the results discussed in Section 6.1 by including other benchmark models. We begin by introducing these additional benchmark models and then present key findings for the three targeting scenarios.

Web Appendix E.1 Additional Benchmark Models

First, we construct IRL models with varying dimensions of representations, ranging from $K_1 = K_2 = 10$ to 50. Second, for both individual and joint CATE models, we consider the following benchmark models in addition to utilizing deep learning models for predicting the doubly robust scores.

1. **DR-learner with XGBoost**: Similar to the Deep DR-learner, we first calculate the doubly robust score. However, instead of employing a deep learning model, we opt for XGBoost to predict the doubly robust score. We conduct five-fold cross-validation and select a tree depth of 5, a learning rate of 1, and 30 rounds, as these parameters achieve the lowest approximation error for the proxy.

2. **T-learner with Deep Neural Network**: We construct two deep learning-based outcome models: one for treated customers and another for control customers. Subsequently, we determine the CATE predictions by calculating the difference between the predicted outcomes from the treatment model and the control model. We use the same neural network architecture as the individual deep DR-learner and hyperparameters for model calibration.

Finally, we explore targeting strategies based on predicted purchase amounts, focusing on baseline predictions rather than incremental purchases. To achieve this, we develop a deep neural network (with the same architecture as the joint deep DR-learner) that utilizes offer design features and customer covariates to predict future purchases across training data in each scenario.

Web Appendix E.2 Targeting for Existing Offers

Table App-2 presents the average RMSE and AUTOC across 274 offers, replicating the results from Table 2. Firstly, the IRL model consistently outperforms other methods, demonstrating superior targeting effectiveness. The performance of the IRL models remains relatively consistent across varying dimensions (rows 1 to 5). Secondly, although

the joint DR-learner using XGBoost (shown in row 7) displays relatively strong targeting performance (slightly below IRL, but not significantly), it also has a higher prediction error compared to all the joint deep DR-learners. This result suggests that although the joint DR-learner using XGBoost may be appropriate for the focal firm to use for targeting, it may not be the best tool for evaluating the profitability of a new campaign due to its significantly larger predictive error.

Thirdly, the deep T-learner (listed in row 10) is identified as the least effective, potentially harmful due to its negative average AUTOC. This is likely a result of the high imbalance in the experiments, with only 10% of customers in the control group, leading to a less accurate outcome model for control group customers.[2]

Fourthly, we observe that targeting customers with high predicted purchases is ineffective as it results in negative AUTOCs, indicating that a higher purchase potential does not always correlate with a higher treatment effect. Finally, the individual DR-learner using XGBoost shows better targeting performance but significantly higher RMSE.

Table App-2: Predictive Validity for 274 Existing Offers

Model	RMSE	AUTOC
IRL (10-dim)	7.78 (0.82)	0.77 (0.23)
IRL (20-dim)	8.07 (0.85)	0.77 (0.12)
IRL (30-dim)	7.74 (0.57)	0.78 (0.11)
IRL (40-dim)	7.68 (0.76)	0.82 (0.11)
IRL (50-dim)	7.48 (0.69)	0.74 (0.14)
Joint-DR-DNN	8.29 (1.52)	0.45 (0.20)
Joint-DR-XGBoost	10.9 (1.26)	0.65 (0.10)
Joint-DR-DNN (No Design Features)	8.62 (1.11)	0.48 (0.13)
Joint-t-DNN	15.91 (1.31)	−0.21 (0.11)
High Predicted Purchase	—	−0.10 (0.04)
Individual-DR-DNN	8.57 (1.41)	0.05 (0.20)
Individual-DR-XGBoost	164.42 (32.21)	0.54 (0.11)
Individual-t-DNN	8.27 (0.49)	0.06 (0.21)

Note: We calculated the mean average RMSE and AUTOC values across 274 offers over 20 splits, with standard deviations shown in parentheses.

Web Appendix E.3 Targeting for Untested Offers

Table App-3 presents the average RMSE and AUTOC across 55 offers related to personal care products, replicating the results from Table 3. Firstly, the IRL models with dimen-

[2]Such imbalance is common in practice, as firms aim to minimize untreated individuals to conserve resources while maintaining a small control group for estimation purposes.

sions $K_1 = K_2 = 10$ and 20 demonstrate better targeting performance compared to models with higher dimensions, where AUTOC tends to decrease as the dimensions of the learned representations increase, underscoring the benefits of dimension reduction for generalization. Secondly, while the Joint-DR-XGBoost (row 7) exhibits the highest RMSE of all models at 10.27, suggesting potential overfitting, it still maintains a moderately good AUTOC. Thirdly, targeting customers with high predicted purchases yields positive AUTOCs, indicating that higher purchase potential correlates with a higher treatment effect for personal care products. Finally, all the individual DR-models underperform compared to the joint models, highlighting the advantages of synergizing information from past experiments.

Table App-3: Targeting Performance for Offers Related to Personal Care Products

Model	RMSE	AUTOC
IRL (10-dim)	3.28 (0.56)	0.67 (0.15)
IRL (20-dim)	3.43 (0.41)	0.64 (0.13)
IRL (30-dim)	3.88 (0.43)	0.59 (0.11)
IRL (40-dim)	3.26 (0.36)	0.48 (0.15)
IRL (50-dim)	3.73 (0.39)	0.50 (0.14)
Joint-DR-DNN	3.32 (0.56)	0.31 (0.10)
Joint-DR-XGBoost	10.27 (0.36)	0.52 (0.12)
Joint-DR-DNN (No Design Features)	3.62 (0.39)	0.54 (0.11)
Joint-t-DNN	10.07 (0.35)	0.37 (0.11)
High Predicted Purchase	—	0.31 (0.15)
Actual (Individual-DR-DNN)	5.44 (0.6)	0.02 (0.14)
Actual (Individual-DR-XGBoost)	9.72 (0.59)	0.29 (0.14)
Actual (Individual-t-DNN)	5.95 (0.69)	0.18 (0.16)

Note: We calculated the mean average RMSE and AUTOC values across 55 holdout offers, with standard deviations shown in parentheses.

Web Appendix E.4 Targeting for Untested Customers

Table App-4 presents the average RMSE and AUTOC across 31 wine offers targeting customers over 40 years old, replicating the results from Table 4. We observe that the IRL models with lower dimensions generally outperform the other models. Additionally, both the joint deep DR-learner and strategies targeting high predicted purchases also show reasonably good targeting performance. However, while the joint model using XGBoost achieved the lowest RMSE, its AUTOC is significantly lower than that of other methods. This suggests that the IRL model, particularly at lower dimensions, provides the most effective robust results and targeting effectiveness.

Table App-4: Targeting Performance for Wine Offers on Customers with Age over 40

Model	RMSE	AUTOC
IRL (10-dim)	5.58 (0.77)	1.27 (0.36)
IRL (20-dim)	5.78 (0.61)	1.20 (0.38)
IRL (30-dim)	6.25 (0.78)	1.14 (0.34)
IRL (40-dim)	5.73 (0.79)	1.17 (0.35)
IRL (50-dim)	5.93 (0.72)	1.16 (0.35)
Joint-DR-DNN	5.75 (0.67)	1.16 (0.37)
Joint-DR-XGBoost	4.53 (0.60)	0.70 (0.34)
Joint-DR-DNN (No Design Features)	6.27 (0.83)	0.98 (0.39)
Joint-t-DNN	9.71 (0.66)	0.44 (0.36)
High Predicted Purchase	—	1.11 (0.31)
Actual (Individual-DR-DNN)	6.65 (0.98)	0.15 (0.45)
Actual (Individual-DR-XGBoost)	19.01 (1.42)	0.57 (0.40)
Actual (Individual-t-DNN)	6.97 (1.03)	0.05 (0.45)

Note: We calculated the mean average RMSE and AUTOC values using holdout data, which includes $783,671$ observations from 31 distinct offers, with standard deviations shown in parentheses.

Web Appendix F Robustness Check

Web Appendix F.1 Depth of Neural Networks

In this appendix, we demonstrate that the predictive validity of the IRL model is robust across various depths of neural network architectures. Specifically, we construct DNN-based models using shallow architectures by removing two hidden layers from each of the neural networks described in Web Appendix D. Additionally, we construct models with deep architectures by adding two additional hidden layers (each with ten nodes) to each of the neural networks outlined in Web Appendix D.

Table App-5: Predictive Validity for 274 Existing Offers

Model	RMSE	AUTOC
Shallow Neural Network Architectures		
IRL (10-dim)	6.90 (1.73)	0.58 (0.21)
IRL (20-dim)	7.51 (0.71)	0.56 (0.19)
IRL (30-dim)	7.69 (0.57)	0.62 (0.13)
IRL (40-dim)	7.62 (0.59)	0.68 (0.14)
IRL (50-dim)	7.33 (0.51)	0.54 (0.16)
Joint-DR-DNN	7.99 (0.80)	0.25 (0.22)
Joint-DR-DNN (No Design Feature)	7.98 (0.45)	0.32 (0.23)
Joint-t-DNN	13.05 (2.50)	−0.24 (0.12)
High Predicted Purchase	—	−0.04 (0.10)
Individual-DR-DNN	9.29 (0.58)	0.12 (0.16)
Individual-t-DNN	13.18 (0.66)	0.15 (0.22)
Medium Neural Network Architectures (Main Analysis)		
IRL (10-dim)	7.78 (0.82)	0.77 (0.23)
IRL (20-dim)	8.07 (0.85)	0.77 (0.12)
IRL (30-dim)	7.74 (0.57)	0.78 (0.11)
IRL (40-dim)	7.68 (0.76)	0.82 (0.11)
IRL (50-dim)	7.48 (0.69)	0.74 (0.14)
Joint-DR-DNN	8.29 (1.52)	0.45 (0.20)
Joint-DR-DNN (No Design Features)	8.62 (1.11)	0.48 (0.13)
Joint-t-DNN	15.91 (1.31)	−0.21 (0.11)
High Predicted Purchase	—	−0.10 (0.04)
Individual-DR-DNN	8.57 (1.41)	0.05 (0.20)
Individual-t-DNN	8.27 (0.49)	0.06 (0.21)
Deep Neural Network Architectures		
IRL (10-dim)	7.22 (0.16)	0.75 (0.16)
IRL (20-dim)	7.17 (0.56)	0.72 (0.17)
IRL (30-dim)	7.10 (0.36)	0.76 (0.14)
IRL (40-dim)	7.57 (2.39)	0.49 (0.32)
IRL (50-dim)	7.04 (0.31)	0.55 (0.11)
Joint-DR-DNN	7.59 (0.45)	0.38 (0.18)
Joint-DR-DNN (No Design Feature)	7.95 (0.54)	0.37 (0.10)
Joint-t-DNN	18.19 (5.73)	−0.11 (0.17)
High Predicted Purchase	25.88 (1.72)	−0.07 (0.05)
Individual-DR-DNN	5.69 (0.26)	0.2 (0.29)
Individual-t-DNN	6.52 (0.38)	0.05 (0.17)
Non-DNN Model		
Joint-DR-XGBoost	10.90 (1.26)	0.65 (0.10)
Individual-DR-XGBoost	164.42 (32.21)	0.54 (0.11)

Note: We calculated the mean average RMSE and AUTOC values across 274 offers over 20 splits, with standard deviations shown in parentheses.

Web Appendix F.2 Including Offers with Negative Treatment Effects

In this Appendix, we present the same analysis as in Section 6, but this time including offers with negative treatment effects. Table App-6 compares the predictive accuracy and targeting effectiveness across different methods. The key findings are consistent with those in Table App-2, where the IRL model outperforms other methods, although the AUTOC value is smaller and the standard deviation is higher. This is likely because it may not be possible to learn any useful targeting rules for offers with negative treatment effects.

Table App-6: Predictive Validity for 362 Existing Offers

Model	RMSE	AUTOC
IRL (10-dim)	7.52 (0.55)	0.38 (0.20)
IRL (20-dim)	7.54 (0.49)	0.38 (0.22)
IRL (30-dim)	7.86 (0.43)	0.40 (0.17)
IRL (40-dim)	7.74 (0.52)	0.37 (0.14)
IRL (50-dim)	7.42 (1.00)	0.34 (0.16)
Joint-DR-DNN	7.76 (0.70)	0.30 (0.18)
Joint-DR-XGBoost	9.16 (0.03)	0.25 (0.14)
Joint-DR-DNN (No Design Feature)	8.79 (0.64)	0.30 (0.26)
Joint-t-DNN	16.00 (0.98)	0.05 (0.22)
High Predicted Purchase	—	0.13 (0.23)
Actual (Individual-DR-DNN)	6.29 (0.22)	0.01 (0.15)
Actual (Individual-DR-XGBoost)	122.37 (1.21)	0.27 (0.10)
Actual (Individual-t-DNN)	8.24 (0.25)	−0.08 (0.20)

Note: We calculated the mean average RMSE and AUTOC values across 362 offers over 20 splits, with standard deviations shown in parentheses.

Web Appendix F.3 Additional Example of Targeting for Untested Offers

In this Appendix, we provide an additional example to demonstrate the ability of the proposed IRL method in generalizing to untested offers. Here, we replicate the analysis in Section 6.1.1, but replace the personal care product category with beer and seltzer products as the holdout offers (the largest product category for offers). Table App-7 presents the average RMSE and AUTOC across 57 offers related to beer and seltzer products, replicating the results from Table App-3. The findings are consistent with those for personal care products: the IRL method outperforms all other methods, and IRL models with lower representation dimensions perform slightly better than those with higher dimensions. The only difference here is that the joint DNN model also performs well in targeting

in this situation. Nonetheless, this example highlights the robust performance of the IRL model.

Table App-7: Targeting Performance for Offers Related to Beer and Seltzer Products

Model	RMSE	AUTOC
IRL (10-dim)	4.51 (0.3)	0.77 (0.20)
IRL (20-dim)	5.28 (0.72)	0.70 (0.21)
IRL (30-dim)	4.53 (0.31)	0.66 (0.18)
IRL (40-dim)	4.88 (0.34)	0.70 (0.18)
IRL (50-dim)	4.75 (0.35)	0.64 (0.20)
Joint-DR-DNN	5.67 (0.85)	0.65 (0.20)
Joint-DR-XGBoost	20.47 (0.72)	0.44 (0.17)
Joint-DR-DNN (No Design Feature)	5.9 (1.59)	0.57 (0.19)
High Predicted Purchase	—	0.26 (0.21)
Joint-t-DNN	18.32 (8.23)	0.12 (0.22)
Actual (Separate-DR-DNN)	9.87 (5.85)	-0.04 (0.25)
Actual (Separate-DR-XGBoost)	13.4 (0.59)	0.39 (0.17)
Actual (Separate-t-DNN)	9.55 (6.71)	0.08 (0.23)

Note: We calculated the mean average RMSE and AUTOC values across 57 offers over 20 splits, with standard deviations shown in parentheses.

Web Appendix F.4 Additional Example of Targeting for Untested Segments

In this Appendix, we provide an additional example to demonstrate the ability of the proposed IRL method to generalize to untested customer segments for certain offers. Here, we replicate the analysis in Section 6.1.2, but now select female customers and offers related to baby products as the holdout set. Table App-8 presents the average RMSE and AUTOC across 52 offers related to baby products, replicating the results from Table Table App-4. The findings are consistent with the main analysis: the IRL method outperforms all other methods, and IRL models with lower representation dimensions perform slightly better than those with higher dimensions.

There are two main differences compared to the results in Table App-4. First, the joint model with traditional deep learning architectures (Rows 6 and 7) performed significantly worse in targeting performance than the other methods. Second, building individual models using XGBoost on the actual experiment data achieves the best targeting performance (although insignificantly different from IRL models with less than 30 dimensions), while the predictive error of the treatment effects (RMSE) is still significantly higher than IRL. In summary, the IRL model remains the best for decision-making as it achieves the lowest RMSE and high AUTOC.

Table App-8: Targeting Performance for Baby Offers on Female Customers

Model	RMSE	AUTOC
IRL (10-dim)	5.83 (0.66)	0.42 (0.23)
IRL (20-dim)	6.00 (0.73)	0.45 (0.28)
IRL (30-dim)	5.90 (0.68)	0.42 (0.21)
IRL (40-dim)	6.08 (0.63)	0.20 (0.28)
IRL (50-dim)	6.40 (0.70)	0.34 (0.28)
Joint-DR-DNN	6.03 (0.74)	−0.15 (0.30)
Joint-DR-XGBoost	6.85 (0.76)	−0.13 (0.25)
Joint-DR-DNN (No Design Feature)	6.61 (0.90)	−0.42 (0.24)
High Predicted Purchase	22.04 (0.91)	−0.22 (0.27)
Joint-t-DNN	—	−0.29 (0.3)
Actual (Separate-DR-DNN)	7.48 (0.76)	0.05 (0.23)
Actual (Separate-DR-XGBoost)	16.82 (1.08)	0.47 (0.25)
Actual (Separate-t-DNN)	7.62 (1.03)	0.10 (0.18)

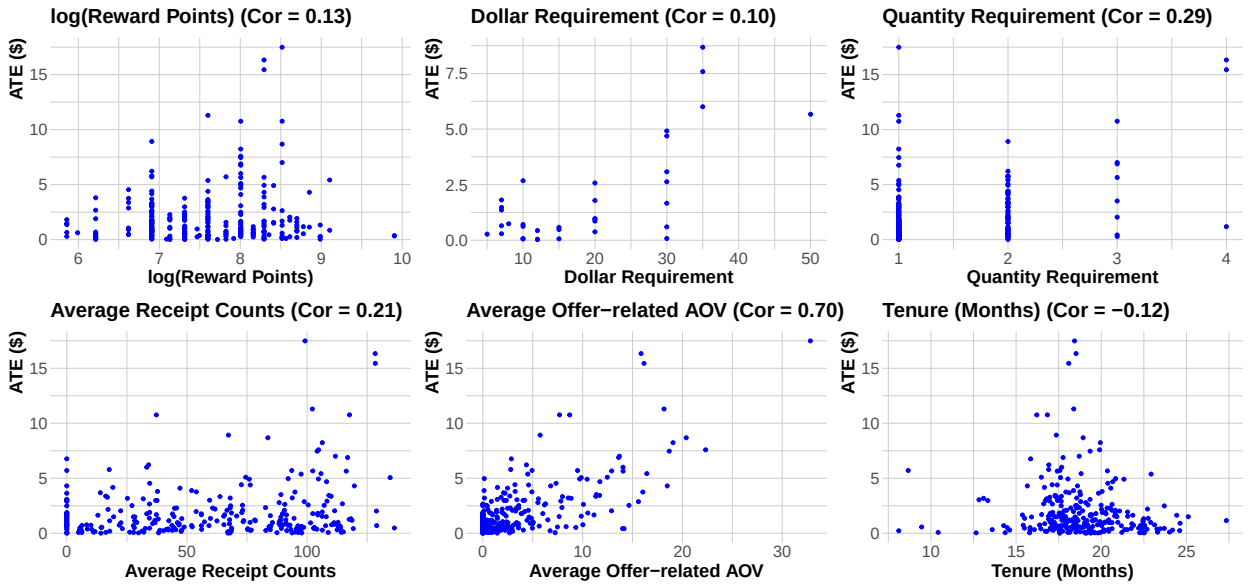
Note: We calculated the mean average RMSE and AUTOC values across 52 offers over 20 splits, with standard deviations shown in parentheses.

Web Appendix G Additional Details for Empirical Results

Web Appendix G.1 Correlations between Average Treatment Effects and Observed Vairables

We explore how variations in ATEs correlate with observable characteristics. Figure App-2 depicts the relationship between various design features, the average characteristics of eligible customers for each offer, and their corresponding ATE. In these subplots, the x-axis represents offer attributes, including design features like reward points and customer covariates such as average age. The y-axis measures the ATE for each offer. The patterns observed indicate that (i) both design features and customer covariates provide some predictive value for an offer's ATE, yet (ii) there are significant variations in ATE, even when specific design features or customer profile variables are accounted for.

Figure App-2: Overview of ATE, Selected Design Features, and Selected Covariate Averages

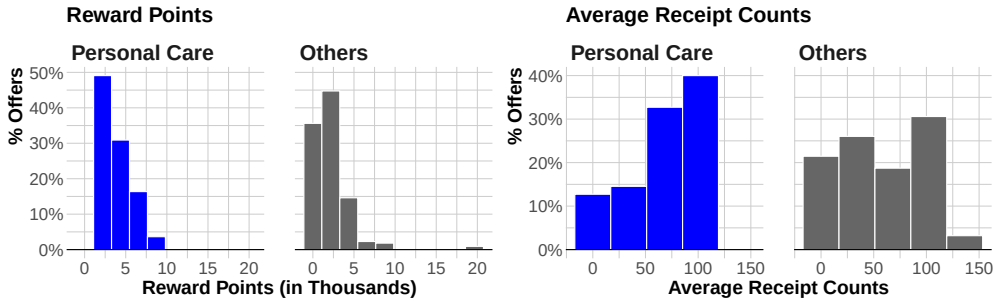


Note. Each point reports the average treatment effect of a promotional offer and its corresponding design feature or covariate average.

Web Appendix G.2 Comparison of Personal Care Product Offers with Other Offers

In Section 6.1.1, we assess the targeting performance of personal care offers when the IRL model is trained using offers unrelated to personal care products. This evaluation not only tests the concept shift problem also examines attribute shift. Figure App-3 illustrates the distribution differences for a key design feature (reward points) and a key customer covariate (average receipt counts). The results indicate that personal care offers tend to offer fewer reward points and target customers with higher receipt counts, demonstrating the presence of attribute shifts in this case study.

Figure App-3: Comparison of Personal Care Product Offers with Other Offers

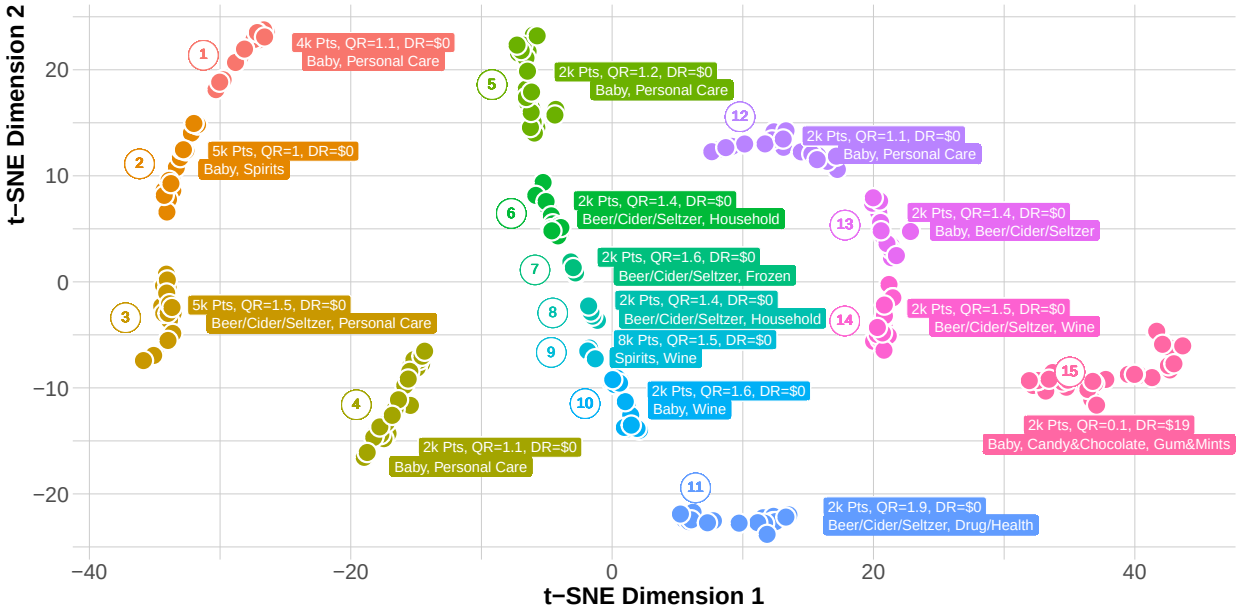


Web Appendix G.3 Offer Clusters Using Raw Design Features

In this appendix, we demonstrate that offer clusters derived from incrementality representations provide a more nuanced understanding than those derived from raw design features. For comparison, we also conduct a similar cluster analysis in Section 6.2.1 using the original design features ($\mathbf{Z}_o$). We adjust the DBSCAN parameters to obtain a comparable number of clusters, resulting in 13 clusters for the incrementality representations and 15 clusters for the design features. Figure App-4 presents the t-SNE map and cluster results. Unlike the clusters formed using raw features, which primarily differentiate baby and personal care products (clusters 1, 5, and 12 in Figure App-4) based mainly on reward points, the incrementality representation clusters identified in Figure 4 provide a more detailed consideration of reward claim requirements.

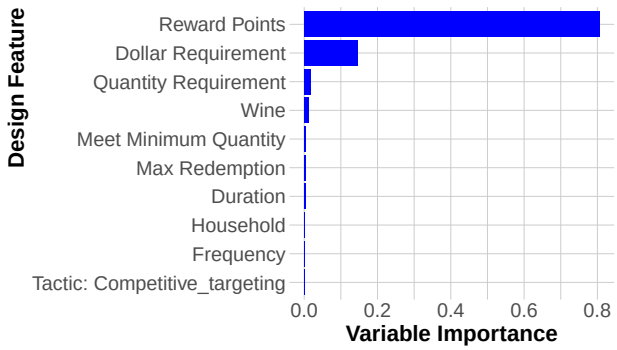
To further validate that the incremental representation captures more nuanced information, we conduct a statistical analysis to identify the features determining cluster membership. The variable importance plot is shown in Figure App-5. In contrast to Figure 5, we find that clusters based on raw design features are overwhelmingly influenced by just two features: reward points, which constitute 80% of the classification outcomes, and dollar requirements, which account for 17% of the outcomes.

Figure App-4: t-SNE Maps for Offer Design Features (Z_o)



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, set as zero for offers requiring minimum dollar spending), required quantities (QR), required dollar amounts (DR, set as zero for offers requiring certain product quantities), and the top two promoted categories for each cluster.

Figure App-5: Variable Importance Plot of Top 10 Design Features for Cluster Classification Models



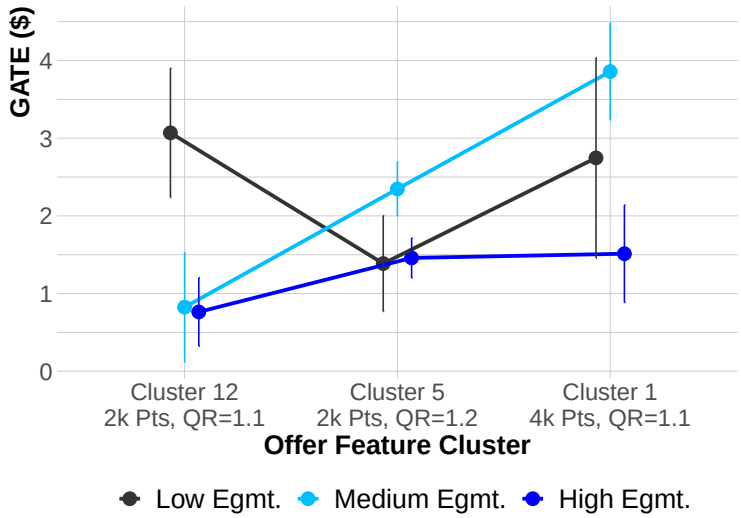
Note. Variable importance is measured by the average gain in accuracy across all splits where the variable is utilized.

Web Appendix G.4 Treatment Effect Heterogeneity Using Original Design Features and Customer Covariates

Section 6.2.2 shows that offer and customer clusters based on incrementality representations elucidate clear treatment effect heterogeneity. In this Appendix, we show that clusters based on the original design features and customer covariates cannot explain treatment effect heterogeneity as effectively. Specifically, we replicate Figure 6 with offer and customer clusters derived using the original design features and customer covariates.

Figure App-6 shows the GATEs for three clusters associated with 'baby and personal care' products derived using raw design features (clusters 12, 5, and 1 in Figure App-4)) across three engagement segments identified using (scaled) customer covariates (Table5b). The results show minimal differences across defined clusters and segments. In contrast, as depicted in Figure 6 in the paper, incrementality representations more effectively elucidate treatment effect heterogeneity. This significant difference highlights the value of key design features and customer segments identified using incrementality representations for tailoring decision making.

Figure App-6: Treatment Effects by Offer Clusters and Customer Segments Derived from $(\mathbf{Z}_o, \mathbf{X}'_{o,i})$

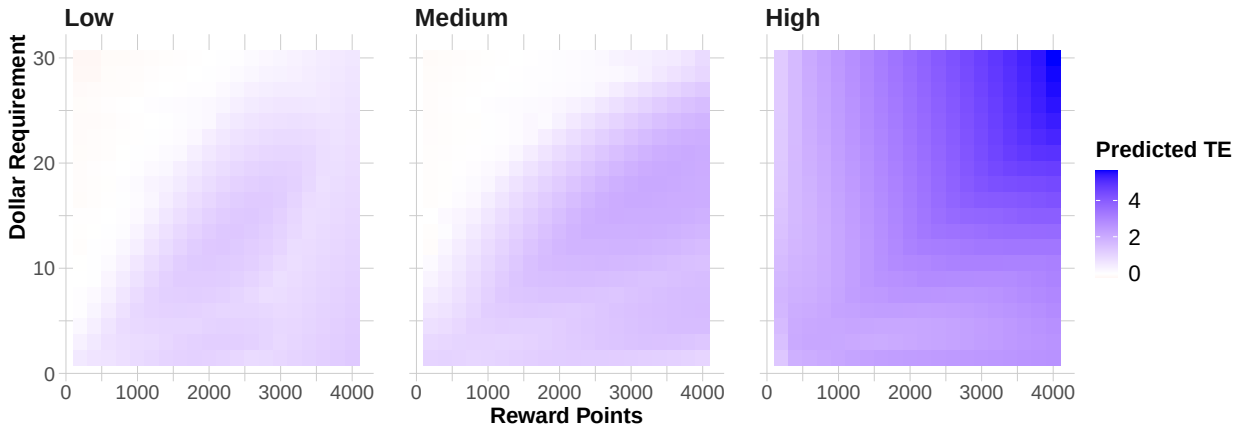


Note. Each point reports the group average treatment effect together with two-standard-error interval. The offer representation clusters align with those outlined in Figure App-4, and the engagement segments correspond to those in Table5b.

Web Appendix G.5 SIP for Predicted Treatment Effects

In this appendix, we create the SIP of predicted treatment effects on offer-related spending across customer segments identified using incrementality representations. Figure App-7 shows the SIP for the previously described baby product offer, generated using the IRL model. Here, reward points vary within $\{200, \cdots, 4000\}$ and the dollar requirement ranges from $\{\$1.5, \cdots, \$30\}$. The results suggest that promotional designs should be tailored to different engagement segments. For the low engagement segment, offering 2,400 reward points with a minimum $14 purchase leads to the highest incremental purchase ($1.35) per customer. For the medium engagement segment, 3,200 reward points with a minimum $18 purchase maximize incremental purchases ($4.28) per customer. Additionally, for high engagement customers, 4,000 reward points with a minimum $30 purchase achieve the greatest incremental purchase ($5.65) per customer.

Figure App-7: Segment Incrementality Plot for the Focal Offer Design



Note. We set the reward points within the range of $\{200, \cdots, 4000\}$ and the dollar requirement within $\{\$1.5, \cdots, \$30\}$.