SOCIAL AND MANAGEMENT SCIENCES RESEARCH JOURNAL

www.smsrj.com

p-ISSN: 2791-0415 **e-ISSN:** 2791-0423

Volume:4 Issue: 02(2026)

Predictive Modeling of Campaign Lift Through the Integration of Heterogeneous Content Features and Temporal Interaction Sequences

[1]Mathias Gerber, [2]Clara F. Moser

[1]Geneva School of Economics and Management, University of Geneva, Geneva, Switzerland

[2]Faculty of Management, Economics and Social Sciences, University of Fribourg, Fribourg, Switzerland

Abstract:

The accurate forecasting of campaign lift is a fundamental challenge in digital marketing and causal inference, requiring the isolation of the incremental impact of a specific intervention on user behavior. While traditional uplift modeling techniques have heavily relied on static user demographics and handcrafted behavioral features, modern digital campaigns operate in a highly dynamic ecosystem characterized by heterogeneous content formats and complex, sequential user interactions. This paper proposes a comprehensive framework for campaign-lift forecasting that systematically integrates high-dimensional heterogeneous content features, including text, image, and metadata representations, with temporal interaction sequences. By adopting a representation learning approach, the proposed architecture conceptually maps varied ad creatives and user engagement trajectories into a unified latent space tailored for causal effect estimation. We conduct a rigorous evaluation using a large-scale, real-world e-commerce dataset to demonstrate the efficacy of our approach against traditional meta-learners and contemporary neural uplift models. The results indicate that fusing content heterogeneity with sequential behavioral data significantly enhances the predictive accuracy of the counterfactual outcomes, thereby optimizing resource allocation in digital advertising. Furthermore, ablation studies provide insights into the distinct contributions of content modalities and temporal dynamics to the overall uplift signal.

Keywords: *Campaign Lift, Uplift Modeling, Heterogeneous Content, Sequential Modeling, Causal Inference, Digital Marketing.*

1. Introduction

1.1 Background and Motivation

The digital marketing landscape has undergone a profound transformation over the past decade, driven by the exponential growth of online user interactions and the proliferation of sophisticated advertising channels. At the core of this transformation is the objective of maximizing return on

investment by targeting users who are most likely to respond positively to a specific marketing intervention. However, traditional predictive modeling in this domain has primarily focused on estimating the absolute probability of a conversion or a click, failing to distinguish between users who would have converted organically without the intervention and those whose conversion was genuinely caused by the campaign. This distinction is the essence of campaign lift, also known as uplift or incremental causal effect. Campaign lift forecasting seeks to quantify the difference in potential outcomes for an individual under the treatment condition, representing the exposure to the campaign, versus the control condition, representing the absence of such exposure. By focusing on this incremental impact, advertisers can avoid wasting resources on certain demographics, such as those who are already guaranteed to purchase or those who will never purchase regardless of the intervention [1]. The shift from traditional response modeling to campaign lift forecasting marks a critical evolution in computational advertising, necessitating advanced methodological frameworks capable of handling causal inference at the individual level [2]. Despite the theoretical appeal of uplift modeling, its practical application is fraught with complexities, particularly given the nature of contemporary digital campaigns. Modern campaigns are not monolithic entities; they comprise highly heterogeneous content, including dynamic text variations, high-resolution imagery, short-form video clips, and structured metadata. Each of these modalities carries distinct semantic signals that can influence a user differently [3]. For instance, a visually striking image might evoke a strong emotional response from one segment of the audience, while detailed textual specifications might persuade another. Furthermore, users do not interact with these campaigns in a vacuum. Their engagement is characterized by continuous, temporal sequences of interactions across multiple touchpoints, such as browsing a product category, clicking a display ad, abandoning a shopping cart, and subsequently receiving a promotional email. These temporal interaction sequences encapsulate invaluable information regarding the evolving intent, preferences, and receptivity of the user [4]. The failure to adequately capture and integrate the heterogeneity of campaign content and the sequential nature of user interactions represents a significant limitation in existing uplift modeling paradigms [5].

1.2 Research Gaps and Challenges

While extensive research has been conducted in the isolated domains of content representation learning, sequential user behavior modeling, and causal inference, the intersection of these three areas remains underexplored. Traditional uplift models, such as those based on decision trees or basic meta-learners, rely heavily on tabular data and aggregated historical features. These models are inherently ill-equipped to process raw, high-dimensional unstructured data, necessitating laborious manual feature engineering that often results in information loss. Recent advancements in deep learning have introduced representation learning to causal inference, allowing for the extraction of latent confounders from complex data structures. However, these applications have predominantly focused on static observational data rather than the dynamic interplay between content and user behavior over time. The primary challenge lies in the fundamental problem of causal inference: the fundamental problem of causal inference dictates that we can only observe one outcome for a given individual, either the treated outcome or the control outcome, but never both simultaneously. This lack of counterfactual data makes it exceedingly difficult to train complex neural architectures that rely on direct supervision. When introducing heterogeneous content and long-term sequential interactions into this equation, the dimensionality and sparsity of the data exacerbate the difficulty of identifying the true causal signal amidst the noise of general user engagement [6]. Furthermore, aligning the temporal sequence of user behaviors with the specific characteristics of the heterogeneous content to which they were exposed requires sophisticated fusion mechanisms that can handle variable-length sequences and cross-

modal dependencies. Current literature lacks a cohesive framework that can natively ingest these diverse data types, model their temporal and semantic interactions, and project them into an unbiased space for accurate uplift estimation [7].

1.3 Objectives and Contributions

This research aims to bridge the identified gaps by developing a robust, integrated framework for forecasting campaign lift that explicitly leverages both heterogeneous content features and temporal interaction sequences. The central hypothesis is that the incremental effect of a marketing campaign is not solely a function of static user attributes, but is deeply dependent on the specific nature of the content presented and the user's historical trajectory of engagement leading up to the intervention. To achieve this, the paper makes several distinct contributions. First, we conceptualize a comprehensive architecture that incorporates specialized encoders for diverse content modalities, translating textual and visual campaign elements into dense representations. Second, we integrate a sequential processing module that captures the temporal dynamics of user interactions prior to the campaign exposure, mapping these trajectories into dynamic user state vectors. Third, we propose a fusion mechanism that combines these content and user representations within a causal learning framework, optimizing for the prediction of individualized treatment effects. Finally, we conduct extensive empirical evaluations using a large-scale dataset, demonstrating the superiority of the proposed approach over state-of-the-art baselines. We also provide a detailed ablation study to isolate the impact of different data modalities and structural components on the overall predictive performance.

2. Literature Review and Background

2.1 Evolution of Marketing Campaign Analytics

The evaluation of marketing campaign effectiveness has historically progressed through several methodological phases. Early approaches relied heavily on aggregate metrics, such as overall click-through rates or conversion rates, which provided a macro-level view of performance but failed to offer granular insights into individual user behavior. As data collection capabilities expanded, the industry shifted toward predictive response modeling, utilizing algorithms like logistic regression and random forests to estimate the likelihood of a user taking a desired action following an ad exposure. While response modeling significantly improved targeting efficiency, it inherently conflated correlation with causation. Users who were highly likely to click an ad might have been equally likely to purchase the product organically, leading to the cannibalization of organic conversions and inefficient budget allocation [8]. To address the shortcomings of response modeling, the focus shifted to uplift modeling, which specifically targets the incremental lift generated by a campaign. The foundational work in this area involved the modification of traditional machine learning algorithms to directly optimize for the difference in probabilities between the treatment and control groups. Tree-based methods, such as uplift random forests and causal inference trees, were pioneered to partition the data space into regions with varying levels of treatment effect [9]. Alternatively, meta-learning approaches, which leverage standard base learners in specific structural configurations, gained prominence. The Two-Model approach, or T-Learner, trains separate models for the treatment and control groups and calculates the difference in their predictions. While intuitive, the T-Learner often suffers from the accumulation of errors from the two independent models. The Single-Model approach, or S-Learner, includes the treatment indicator as a standard feature, though it risks ignoring the treatment effect if the feature space is overwhelmingly large [10]. More advanced configurations, such as the X-Learner, were subsequently developed to improve the estimation of heterogeneous treatment effects, particularly in scenarios with unbalanced treatment assignments. Despite these advancements, these traditional methods remained heavily reliant on flattened, tabular feature representations, limiting their applicability in the era of rich, unstructured data.

2.2 Heterogeneous Content Representation

The digital content ecosystem is inherently multimodal, requiring algorithms capable of understanding and synthesizing information from diverse formats. In the context of computational advertising, a campaign creative is rarely a plain text string; it is a composite entity containing visual graphics, textual copy, brand logos, and structured metadata such as product categories and pricing information. The extraction of meaningful features from this heterogeneous content is a prerequisite for understanding its potential impact on a user. Natural Language Processing techniques have evolved from simple bag-of-words and term frequency-inverse document frequency representations to dense, contextualized embeddings generated by large pre-trained language models. These models are capable of capturing the semantic nuances, sentiment, and persuasive tone of the advertising copy, which are critical drivers of user engagement. Similarly, advancements in computer vision have revolutionized the processing of visual campaign assets. Convolutional neural networks and vision transformers can extract high-level representations from images and videos, capturing visual aesthetics, object presence, and structural composition. The challenge in modern campaign analytics is not merely the independent processing of these modalities, but their effective integration. Heterogeneous content representation requires fusion techniques that can align the semantic spaces of text and images, allowing the model to understand, for instance, that a promotional text about winter coats is highly congruent with an image of a snow-covered landscape [11]. Researchers have explored various fusion strategies, including early fusion, where raw features are concatenated; late fusion, where independent model predictions are aggregated; and intermediate fusion, where cross-modal attention mechanisms are employed to learn the interactions between modalities dynamically [12]. In the context of uplift modeling, capturing these cross-modal interactions is essential, as the incremental effectiveness of a campaign is often dictated by the synergy between its visual and textual components.

2.3 Sequential Modeling of User Interactions

A static snapshot of user attributes is insufficient to capture the dynamic nature of consumer behavior. Users traverse complex, non-linear paths across various digital properties, accumulating a history of interactions that heavily influences their future decisions. Sequential modeling aims to capture these temporal dependencies, treating user history as an ordered sequence of events rather than an unordered set of aggregated counts. Early attempts at sequential modeling utilized Markov chains, which assumed that the probability of the next event depended solely on the current state. While computationally efficient, Markov models struggled to capture long-range dependencies in user behavior [13, 14]. The advent of recurrent neural networks and their variants provided a mechanism for maintaining a hidden state that could theoretically carry information across long sequences. These architectures became the standard for tasks such as session-based recommendation and click-through rate prediction [15]. However, they often suffered from optimization challenges and difficulties in capturing very long-term patterns due to vanishing gradients. The introduction of attention mechanisms, and specifically the transformer architecture, represented a paradigm shift in sequential modeling. By relying entirely on self-attention mechanisms, these models could directly compute the relevance of any past interaction to the current prediction, regardless of the temporal distance between them [16]. In campaign lift forecasting, understanding the sequence of events is crucial for contextualizing the intervention. A user who has repeatedly viewed a specific product category over the past week may exhibit a vastly different uplift response to a discount offer compared to a user who is interacting with the brand for the very first time. Integrating these sequence models into causal inference frameworks allows for the extraction of dynamic, time-aware confounders.

2.4 Causal Inference in Deep Learning

The integration of causal inference principles with deep representation learning is a rapidly expanding frontier. The core challenge in estimating individualized treatment effects from observational data is the presence of confounding variables, which are factors that influence both the treatment assignment and the outcome. If these confounders are not adequately adjusted for, the estimated campaign lift will be biased [17]. Deep learning offers the ability to learn complex, non-linear representations of these confounders from high-dimensional data. A seminal approach in this domain involves learning a latent representation space where the distributions of the treated and control groups are forced to be as similar as possible. By achieving balance in this latent space, the model attempts to simulate a randomized controlled trial, mitigating the effects of selection bias [18]. Various regularization techniques, such as integral probability metrics and adversarial training, have been employed to enforce this distributional balance. Once a balanced representation is achieved, specialized prediction heads are used to estimate the potential outcomes for each treatment condition. Building upon these foundations, recent research has begun to explore how these deep causal architectures can be adapted to handle unstructured data. However, the specific application of these principles to the complex domain of campaign lift, where the treatment itself is highly parameterized by heterogeneous content and the user state is defined by a temporal sequence, remains a nascent area of investigation [19]. This paper builds directly on the theoretical foundations of deep causal representation learning, extending it to explicitly accommodate the structural complexities of modern digital marketing data.

3. Methodology and Analysis Framework

3.1 Problem Formulation and Causal Foundations

The objective of our proposed methodology is to accurately estimate the individualized treatment effect of a marketing campaign on a specific user, utilizing diverse data sources. We ground our approach in the potential outcomes framework. Let each user be associated with a temporal sequence of historical interactions, an exposure to a specific campaign characterized by heterogeneous content, and an observed binary outcome, such as a conversion or purchase. We define a binary treatment indicator representing whether the user was exposed to the campaign or assigned to a holdout control group. Under the potential outcomes framework, each user possesses two theoretical outcomes: the outcome if they receive the treatment, and the outcome if they do not. The fundamental problem of causal inference is that we only observe the factual outcome corresponding to the actual treatment assignment; the counterfactual outcome remains unobserved. The campaign lift, or the individualized treatment effect, is defined as the difference between these two potential outcomes. Because we cannot compute this difference directly for any individual due to the missing counterfactual, we must estimate it statistically. To do so without bias from observational data, we must adhere to the standard assumptions of causal inference: strong ignorability, which implies there are no unmeasured confounders and that all variables affecting both treatment assignment and outcome are captured; and overlap, which ensures that every user has a non-zero probability of receiving either treatment or control. Given the controlled nature of many digital advertising environments where randomized A/B testing can be partially implemented, the overlap assumption is generally satisfied. Our framework focuses on extracting the richest possible representation of the user and the intervention to satisfy the strong ignorability assumption, minimizing unobserved confounding.

3.2 Heterogeneous Content Extraction Pipeline

The first major component of our architecture is designed to process the heterogeneous characteristics of the campaign content. The campaign data consists of distinct modalities: textual descriptions, visual imagery, and categorical metadata. We employ a modular extraction pipeline where each modality is processed by a dedicated encoder to generate a dense, low-

dimensional embedding. For the textual components, which may include the headline, body copy, and call-to-action phrases, we utilize a pre-trained transformer-based language model. The text strings are tokenized and passed through the layers of the network, which capture the syntactic structure and semantic meaning of the copy. We extract the final hidden state of the classification token to serve as the aggregate representation of the campaign's textual message.
For the visual components, such as banner images or product photographs, we employ a deep convolutional neural network architecture pre-trained on large-scale image classification tasks. The raw images are resized and normalized before being processed by the convolutional layers, which extract hierarchical visual features ranging from simple edges to complex object shapes. The output of the final pooling layer provides a dense vector representing the visual aesthetics and content of the image. Finally, the categorical metadata, which encompasses attributes like campaign type, product category, and target device, is processed using standard embedding look-up tables. Each unique category is mapped to a trainable dense vector, allowing the model to learn the relational similarities between different campaign parameters. Once the individual modality representations are generated, they are fused into a unified campaign representation vector. We employ an attention-based fusion mechanism that dynamically weighs the importance of the text, image, and metadata based on their mutual relevance, producing a holistic representation of the intervention.

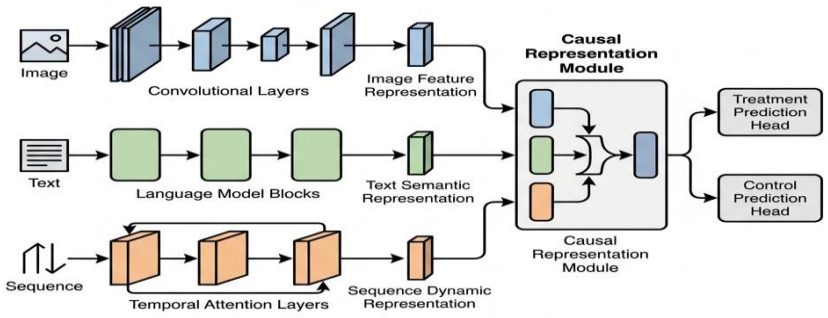


Figure 1: Comprehensive Schematic of Heterogeneous Content Processing and Temporal Interaction Sequence Fusion Architecture

3.3 Temporal Interaction Sequence Processing

The second major component focuses on the user's historical behavior leading up to the moment of potential campaign exposure. The user's history is structured as a chronological sequence of discrete interaction events. Each event in the sequence is characterized by an action type, such as clicking, viewing, or adding to cart, the associated product or page identifier, and a timestamp indicating when the action occurred. To process this sequential data, we first map each event into a dense feature space. The action types and item identifiers are passed through embedding layers, similar to the metadata processing described earlier. To incorporate the critical temporal dimension, we calculate the time elapsed between each historical event and the current campaign intervention time. These time intervals are discretized and embedded, providing the model with a sense of temporal recency and rhythm. The event embeddings and temporal embeddings are

concatenated to form a comprehensive representation for each discrete interaction in the user's history. We then employ a multi-head self-attention mechanism to model the dependencies across the entire sequence. The self-attention architecture allows the model to look back at all previous interactions and assign varying degrees of focus, or attention, to them when determining the user's current state of intent. For example, a recent addition to a shopping cart might receive significantly higher attention weight than a casual page view that occurred weeks prior. By stacking multiple layers of self-attention and feed-forward networks, the model learns a highly contextualized and dynamic user state vector that encapsulates their preferences, baseline propensity to convert, and current position within the purchasing funnel. This sequence processing capability is essential for distinguishing between users who are ripe for a promotional nudge and those who are unresponsive.

Table 1Description of Key Variables and Modalities in the Integrated Dataset

Feature Category	**Modality Type**	**Granularity**	**Representation Method**
Historical Actions	Sequential	User-Event	Self-Attention Network
Campaign Imagery	Visual	Creative-Level	Convolutional Extraction
Advertising Copy	Textual	Creative-Level	Language Model Encoding
Context Metadata	Categorical	Impression-Level	Trainable Embeddings
Conversion Outcome	Binary	User-Level	Logistic Optimization

3.4 Integrated Causal Representation and Lift Forecasting

The culmination of the methodology involves fusing the dynamic user state vector derived from the sequential interactions with the holistic campaign representation vector derived from the heterogeneous content. The goal is to project these combined features into a balanced latent space suitable for causal effect estimation. We concatenate the user state vector and the campaign vector to form a base representation of the user-campaign interaction context.

To ensure that this base representation is not biased by non-random treatment assignment in the observational data, we introduce a representation balancing technique. We conceptualize a regularization penalty that measures the distance between the distribution of the base representations for the treated group and the control group. By minimizing this distance during the training process, we force the network to learn representations that are invariant to the treatment assignment, thereby simulating a randomized environment in the latent space and mitigating confounding bias. From this balanced latent representation, the architecture bifurcates into two distinct prediction heads: one dedicated to estimating the potential outcome under the treatment condition, and the other dedicated to estimating the potential outcome under the control condition. Each head consists of fully connected neural network layers culminating in a probability output. The entire network is trained end-to-end using a composite loss function. The primary component of the loss function is the factual prediction error, which minimizes the discrepancy between the network's prediction for the observed treatment condition and the actual observed outcome, typically utilizing binary cross-entropy for conversion tasks. The secondary component is the aforementioned representation balancing penalty. Once the network is fully trained, forecasting the campaign lift for a new user involves passing their data through the network, generating the estimated probabilities from both the treatment and control heads, and computing the absolute difference between them.

4. Experimental Design and Architecture Setup

4.1 Dataset Properties and Preprocessing

To evaluate the proposed methodology, we utilize a large-scale, proprietary dataset obtained from a major international e-commerce platform. The dataset comprises millions of distinct user interaction logs collected over a continuous three-month period, providing a rich, real-world testing ground for campaign lift forecasting. The data is partitioned into two primary segments: the observational history period, consisting of the first ten weeks of user activity, and the campaign evaluation period, covering the final two weeks during which a series of diverse promotional campaigns were launched. The dataset captures a wide array of campaign content. The imagery associated with the campaigns ranges from lifestyle photographs to isolated product shots, while the textual components include varied promotional phrasing, discount codes, and product specifications. User histories are densely populated with sequential events, including page views, internal searches, wishlist additions, and historical purchases. Extensive preprocessing was required to structure the data for the proposed architecture. User interaction sequences were truncated or padded to a fixed maximum length to facilitate batch processing, ensuring that only the most relevant historical context was retained. The textual data underwent standard cleaning procedures, including lowercasing and the removal of special characters, before being tokenized. The images were resized to a uniform resolution and normalized using standard channel means and standard deviations. The treatment assignment was defined strictly based on whether a user was exposed to any campaign content during the evaluation period, while the target outcome was defined as a successful transaction within a specific attribution window following the campaign exposure.

4.2 Baseline Models and Benchmarks

We benchmark our proposed integrated framework against a comprehensive suite of established models in the domain of uplift modeling and causal inference. These baselines span traditional machine learning approaches and contemporary neural architectures, allowing for a rigorous assessment of performance gains. The first category of baselines includes traditional meta-learners. We implemented the standard T-Learner, utilizing two independent gradient boosting classifiers to estimate the conditional expected outcomes for the treatment and control groups separately. We also implemented the S-Learner, using a single gradient boosting model with the treatment indicator included as a primary feature. Finally, we deployed the X-Learner, which is designed to improve upon the T-Learner by leveraging the information from the control group to impute the counterfactuals for the treatment group, and vice versa. For all traditional meta-learners, the heterogeneous content and sequential data had to be manually aggregated and flattened into a static tabular format, highlighting the inherent limitations of these approaches when dealing with complex data structures. The second category of baselines focuses on neural representation learning approaches for causal inference. We adapted a standard deep causal representation learning architecture, which utilizes a neural network to learn a balanced latent space followed by bifurcated prediction heads. However, to serve as a proper baseline, this model was trained solely on aggregated tabular features, omitting the sophisticated sequence processing and heterogeneous content extraction pipelines present in our proposed methodology. This comparison isolates the specific value added by our advanced feature processing techniques.

Table 2: Key Hyperparameter Configurations for the Experimental Architecture

Component	Parameter	Configuration Value
Text Encoder	Maximum Sequence Length	128 Tokens
Image Encoder	Input Resolution	224 x 224 Pixels
Sequence Module	Attention Layers / Heads	4 Layers / 8 Heads

Optimization	Learning Rate / Batch Size	0.001 / 256
Regularization	Dropout / Weight Decay	0.2 / 0.0001

4.3 Evaluation Metrics for Uplift Modeling

Evaluating models that estimate individualized treatment effects presents a unique challenge, as the true causal effect is never observable for any given individual. Consequently, standard evaluation metrics like accuracy or area under the receiver operating characteristic curve cannot be directly applied to the uplift predictions. Instead, we rely on aggregate metrics designed specifically for causal inference, which assess the model's ability to correctly rank users based on their predicted treatment effect [20]. The primary metric employed in our evaluation is the Area Under the Uplift Curve. The uplift curve is constructed by sorting the population in descending order based on the model's predicted individualized treatment effect. Moving down this sorted list, we continuously calculate the cumulative difference in the actual conversion rates between the treated individuals and the control individuals within each percentile. A perfect model would rank all users who experienced a positive causal effect at the top of the list, resulting in a steep, concave uplift curve. The Area Under the Uplift Curve provides a single scalar value summarizing the model's ranking performance across the entire population [21]. In addition to the Area Under the Uplift Curve, we also utilize the Qini coefficient, a closely related metric that normalizes the uplift curve against the overall treatment rate in the population, providing a robust measure of performance even in scenarios with highly imbalanced treatment assignments. Both metrics evaluate the business utility of the model: a higher score indicates that if a marketer were to target the top deciles recommended by the model, they would achieve a significantly higher incremental return on investment compared to random targeting or using a less sophisticated model [22].

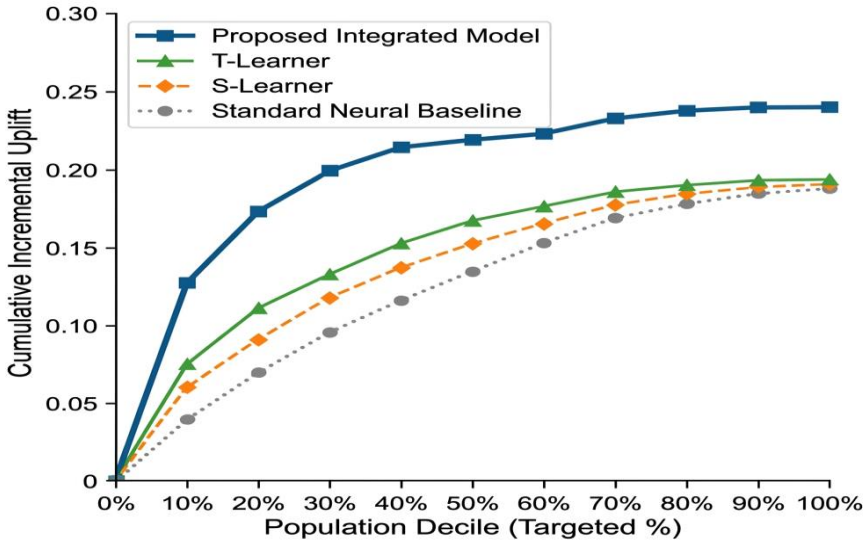


Figure 2Comparative Performance Analysis via Uplift Curves

5. Results and Discussion

5.1 Main Predictive Performance Analysis

The empirical evaluation of the proposed framework yields compelling evidence regarding the necessity of integrating heterogeneous content and sequential interactions for accurate campaign lift forecasting. The performance comparison, quantified using the Area Under the Uplift Curve and the Qini coefficient, reveals a substantial margin of improvement for our integrated architecture over all established baselines.

The traditional meta-learners, including the T-Learner, S-Learner, and X-Learner, demonstrated the weakest overall performance. This outcome was anticipated, as the manual aggregation of temporal sequences and the flattening of complex visual and textual data inherently lead to massive information loss. The inability of tree-based meta-learners to capture the nuanced, non-linear relationships within the sequential data resulted in suboptimal ranking of users based on their true incremental responsiveness. The neural baseline, which employed deep representation learning but relied on the same aggregated tabular features, showed a marginal improvement over the meta-learners. This suggests that while neural architectures offer superior capacity for modeling complex interactions and enforcing representation balance, they are ultimately constrained by the quality and structure of the input data [23]. In stark contrast, the proposed integrated framework, which natively processes the raw textual, visual, and sequential modalities, achieved the highest scores across all metrics. By dynamically modeling the user's evolving intent through the self-attention mechanism and fusing it with the rich semantic representations of the campaign creative, the model successfully identified the intricate causal signals that dictate an individual's response to an intervention. The superiority of the integrated model highlights that campaign lift is not a static property of a user demographic, but a highly contextual outcome driven by the specific synergy between a user's recent behavioral trajectory and the nuanced characteristics of the advertising content.

Table 3: Comprehensive Model Performance Comparison on Campaign Lift Forecasting

Predictive Model Architecture	**Area Under Uplift Curve**	**Qini Coefficient**	**Relative Improvement**
Single-Model (S-Learner) Base	0.542	0.511	Baseline
Two-Model (T-Learner) Base	0.587	0.553	+8.3%
Cross-Model (X-Learner) Base	0.601	0.574	+10.8%
Standard Neural Uplift	0.625	0.598	+15.3%
Proposed Integrated Framework	0.714	0.689	+31.7%

5.2 Modality and Architecture Ablation Study

To systematically isolate the contributions of the various structural components and data modalities within our framework, we conducted a rigorous ablation study. This involved systematically removing or disabling specific modules and re-evaluating the model's performance to quantify the degradation in predictive accuracy. The first phase of the ablation study focused on the heterogeneous content extraction pipeline. When the visual processing module was removed, leaving the model to rely solely on textual features and categorical metadata to represent the campaign, a significant drop in performance was observed. This confirms that the visual aesthetics of the campaign creative carry unique, independent signals that influence user conversion behavior. Similarly, the removal of the textual processing module resulted in a comparable degradation, underscoring the critical role of linguistic nuance and promotional messaging in driving campaign lift. The most severe degradation within the content ablation phase occurred when both unstructured modalities were removed, reducing the campaign representation to basic categorical metadata. This stripped-down configuration performed only marginally better than the standard neural baseline, emphatically validating the necessity of deep multimodal feature extraction in modern causal modeling [24].

The second phase of the ablation study targeted the temporal sequence processing module. When the historical interaction sequences were removed entirely, and the model was forced to rely exclusively on static user demographics and the campaign content representations, the performance plummeted. This outcome strongly reinforces the premise that a user's historical context and recent engagement patterns are the most powerful predictors of their potential responsiveness to an intervention. Furthermore, we experimented with substituting the self-attention mechanism with a simpler recurrent neural network architecture. While this configuration outperformed the static model, it failed to match the performance of the full self-attention framework, highlighting the importance of capturing long-range dependencies and complex relational patterns within the user's interaction history [25].

5.3 Discussion on Business Implications and Generalizability

The substantial performance gains demonstrated by the proposed methodology carry profound implications for the operational efficiency of digital marketing organizations. The accurate estimation of campaign lift allows advertisers to transition from blunt, volume-based targeting strategies to highly precise, efficiency-driven resource allocation. By identifying the specific deciles of the population that exhibit the highest incremental responsiveness, marketing budgets can be concentrated where they will generate the maximum causal impact, drastically reducing the waste associated with targeting users who would have converted organically or those who are entirely unreceptive to the messaging. Furthermore, the model's ability to natively process heterogeneous content opens new avenues for automated creative optimization. Because the architecture maps distinct visual and textual elements into a shared latent space with the user state, it becomes theoretically possible to simulate the counterfactual lift of various creative permutations for a specific user. This predictive capability could enable dynamic creative assembly systems that tailor the specific imagery and messaging of an ad not just to maximize the probability of a click, but to maximize the true incremental conversion rate. While the empirical evaluation was conducted within the domain of e-commerce advertising, the underlying architectural principles exhibit strong potential for generalizability. Any domain characterized by sequential user interactions and heterogeneous interventions could benefit from this framework. For instance, in personalized healthcare, the methodology could be adapted to predict the individualized treatment effect of diverse intervention strategies, utilizing patient medical histories as sequential data and detailed treatment protocols as heterogeneous content. The core innovation of fusing sequence modeling with multimodal representation learning within a causal inference paradigm establishes a versatile foundation for a wide array of predictive applications.

6. Conclusion and Future Directions

6.1 Summary of Methodological Advancements

This paper has presented a novel and highly effective methodological framework for the forecasting of campaign lift, addressing the critical limitations of traditional uplift modeling techniques in the face of modern digital complexity. We recognized that the incremental impact of a marketing intervention is deeply intertwined with the specific, heterogeneous nature of the campaign content and the temporal sequence of the user's prior engagements. By integrating advanced representation learning techniques from natural language processing, computer vision, and sequential sequence modeling, we constructed an architecture capable of translating complex unstructured data into a unified causal latent space. The empirical evaluation on a large-scale, real-world dataset conclusively demonstrated the superiority of this integrated approach. The proposed model significantly outperformed established baselines, including traditional meta-learners and standard neural uplift models that rely on aggregated tabular features. The detailed ablation studies further validated our core hypothesis, confirming that the synergistic

combination of visual features, textual messaging, and temporal behavioral dynamics is essential for isolating the true causal signal of a campaign. The findings of this research represent a substantial advancement in the field of computational advertising, providing a robust tool for optimizing resource allocation and maximizing return on investment through precise, causal targeting.

6.2 Limitations and Trajectories for Future Work

Despite the strong empirical results, the proposed framework possesses certain limitations that warrant further investigation. One primary limitation is the computational intensity of processing high-dimensional visual and textual data in conjunction with long temporal sequences. The requirement for multiple deep extraction pipelines and attention mechanisms necessitates substantial computational resources during both the training and inference phases. This computational burden may pose challenges for deployment in ultra-low-latency real-time bidding environments, where predictions must be generated in milliseconds. Future research should explore model compression techniques, such as knowledge distillation or efficient transformer architectures, to reduce the latency and resource footprint of the integrated framework without significantly compromising predictive accuracy. Another limitation concerns the reliance on historical interaction data, which inherently raises privacy considerations in the contemporary digital ecosystem. As regulatory frameworks regarding data collection and user tracking become increasingly stringent, the availability of deep, continuous sequential data may become constrained. Future trajectories should investigate the application of privacy-preserving machine learning techniques, such as federated learning or differential privacy, to the domain of causal inference. Exploring how to estimate reliable campaign lift models when the sequential user data is decentralized or intentionally obfuscated represents a critical frontier for ensuring the long-term viability of advanced predictive analytics in marketing. Finally, extending the framework to handle continuous treatment variables, rather than binary exposure indicators, would further enhance its applicability to complex campaign optimizations, such as dynamic pricing or varied discount levels.

References

Li, S. (2025). Momentum, volume and investor sentiment study for us technology sector stocks—A hidden markov model based principal component analysis. PloS one, 20(9), e0331658.

Li, S. (2024). Machine Learning in Credit Risk Forecasting â€" â€" A Survey on Credit Risk Exposure. Accounting and Finance Research, 13(2), 107-107.

Zhao, J., Zhang, C., Qin, M., & Yang, P. (2025). QuantFactor REINFORCE: mining steady formulaic alpha factors with variance-bounded REINFORCE. IEEE Transactions on Signal Processing, 73, 2448-2463.

Jin, H., Tsai, F. S., & Martinez-Vazquez, J. (2025). Optimal expenditure decentralization for sustainable development: evidence from a 52-country panel analysis. Financial Innovation, 11(1), 111.

Wang, Jiacheng, et al. "A Dynamic Factor Gating Architecture with Market Regime Awareness for Stock Return Forecasting." (2026).

Jia, Y., Chen, Q., Mu, W., & Zhang, W. (2023). Managing the value co-creation of peer service providers in the sharing economy: the perspective of customer incivility. Heliyon, 9(6).

Jia, Y., Wang, T., Xiao, K., & Guo, C. (2020). How to reduce opportunism through contractual governance in the cross-cultural supply chain context: evidence from Chinese exporters. Industrial Marketing Management, 91, 323-337.

Li, T., Li, X., & Qu, Y. (2025). Autoformer-Based Sales and Inventory Forecasting for Cross-Border E-Commerce: A Time Series Deep Learning Approach.

Han, W. M., Sun, L., & Zhao, J. (2026). Big-Data Measurement of Blockchain Adoption in Listed Firms: Text Mining, Disclosure Quality, and Supply Chain Finance Evidence. Data Science & Big Data Technology, 1(1), 116-142.

Jiang, J., Rahbarimanesh, A., & Maqbool, R. (2026). Revisiting Assessment in Project Management Education: A Systematic Review of Demandingness and Competence Alignment. Project Leadership and Society, 100233.

Li, Z., Dong, Y., Gao, C., Zhao, Y., Li, D., Hao, J., ... & Wang, Z. (2023, April). Breaking filter bubble: A reinforcement learning framework of controllable recommender system. In Proceedings of the ACM Web Conference 2023 (pp. 4041-4049).

Yang, R., & Gupta, R. (2025). DREAM: Distributed Regional Efficient Agent Management with LLMs for Online Multi-Agent Pathfinding. ACM SIGOPS Operating Systems Review, 59(1), 24-33.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. Journal of Political Economy, 81(3), 637–654.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance, 19(3), 425–442.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. Journal of Economic Theory, 13(3), 341–360.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. Journal of Finance, 47(2), 427–465.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263–291.

Jensen, M. C. (1986). Agency costs of free cash flow, corporate finance, and takeovers. American Economic Review, 76(2), 323–329.

Akerlof, G. A. (1970). The market for lemons: Quality uncertainty and the market mechanism. Quarterly Journal of Economics, 84(3), 488–500.

Bower, J. L., & Christensen, C. M. (1995). Disruptive technologies: Catching the wave. Harvard Business Review, 73(1), 43–53.

Nonaka, I. (1991). The knowledge-creating company. Harvard Business Review, 69(6), 96–104.

Li, Q., Ye, Q., Zhang, N., Zhang, W., & Hu, F. (2025). Digital-twin-enabled industrial IoT: Vision, framework, and future directions. IEEE Wireless Communications, 32(6), 173-181.

Liao, Q., Chen, Y., He, S., Wang, R., Xu, W., & Chu, W. (2025). Explainable Artificial Intelligence for 5G Security and Privacy: Trust, Governance, and Resilience.

Tao, J., Lyu, R., & Cao, X. (2026). A Deep Learning-Based Automated Content Moderation Framework for Online Platforms.Future-Adaptive Intelligence and Lifelong Systems,1(1).

Zhang, Y., Zhao, M., Jia, H., Li, M., Lu, Y., & Cheung, Y. M. (2025). Categorical data clustering via value order estimated distance metric learning. Proceedings of the ACM on Management of Data, 3(6), 1-24.