# Valid Inference for Treatment Effects under Multimodal Confounding

**Sven Klaassen** SVKLAA@GMAIL.COM
*Christian-Albrechts-Universität zu Kiel*

**Jan Teichert-Kluge** JAN.TEICHERT-KLUGE@UNI-HAMBURG.DE
*University of Hamburg*

**Philipp Bach** PHILIPP.BACH@FU-BERLIN.DE
*Freie Universität Berlin*

**Victor Chernozhukov** VCHERN@MIT.EDU
*Massachusetts Institute of Technology*

**Martin Spindler** MARTIN.SPINDLER@UNI-HAMBURG.DE
*University of Hamburg*

**Suhas Vijaykumar** SVIJAYKUMAR@UCSD.EDU
*UC San Diego*

**Editors:** Bijan Mazaheri and Niels Richard Hansen

## Abstract

This paper provides methods for the valid estimation and inference of treatment effects in the presence of unstructured, multimodal data, namely text and images, as confounders. We develop a neural network architecture that is adapted to the double machine learning (DML) framework, specifically the partially linear model. An additional contribution of our paper is a new method to generate a semi-synthetic dataset which can be used to evaluate the performance of causal effect estimation in the presence of text and images as confounders. The proposed methods and architectures are evaluated on a correspondingly generated semi-synthetic dataset and compared to standard approaches, highlighting the potential benefit of using text and images directly in causal studies. In the experiments, our methods performs well and achieves the nominal coverage. Our findings might be valuable for researchers and practitioners in economics, marketing, finance, medicine and data science in general who are interested in estimating causal quantities using non-traditional data.

**Keywords:** Causal Inference, Causal Machine Learning, Double Machine Learning, Multimodal Data, Unstructured Data, Inference

## 1. INTRODUCTION

In this paper, we propose deep learning architectures for treatment effect estimation and valid statistical inference in the presence of high-dimensional and unstructured multimodal confounders, emphasizing the utilization of deep learning techniques for handling complex nuisance parameters. In many cases, text and image data contain information that can otherwise not be accounted for in causal studies, for example in the form of sentiment in product descriptions or reviews, labels for product images in online marketplaces or health information encoded in medical images. In causal studies, this information can be very important to account for otherwise unmeasured confounding or to improve estimation precision of causal effects. Our focus is on developing methods that ensure root-$N$ consistency and valid inferential statements of the causal parameter, particularly in scenarios where traditional semi-parametric assumptions are challenged by the increasing complexity of the nuisance parameter space (Foster and Syrgkanis, 2023). The parameter of interest

will typically be a causal or treatment effect parameter, denoted by $\theta_0$. Common examples for $\theta_0$ include the average treatment effect (ATE) or the ATE for the subgroup of the Treated (ATT). We propose neural network architectures adapted to the confounding structure and DML framework, which allows for valid inference in complex settings, with unstructured, multimodal confounding variables, called DoubleML-Deep. In this setting the nuisance parameters / functions are estimated using deep learning methods, such as transformers, or large language models (LLM). These deep learning methods are capable of handling high-dimensional, unstructured covariates, like texts and images (Goodfellow et al., 2016; Zhang et al., 2023), and provide estimators of nuisance functions when these functions are highly complex. In this context, "highly complex" refers to the entropy of the parameter space associated with the nuisance parameter increases with the sample size, going beyond the conventional framework addressed in the classical semi-parametric literature (Härdle et al., 2012). The main contribution of this paper is to develop neural net architectures blended with the DML framework and to offer a general procedure for estimation and inference on $\theta_0$ that is formally valid in these highly complex settings. In Section 5.1, we also propose a method to generate semi-synthetic data in the presence of text and images as confounders. Data generating processes for unstructured data are characterized by inherent challenges, which are briefly summarized and addressed in our paper. Given the growing interest in causal inference with text and image data and the increased availability of this data, we believe that this contribution of our paper might be of independent interest. In a simulation study we compare our methods with different approaches and find that DoubleML-Deep has the smallest bias and root mean squared error (RMSE) and achieves the nominal coverage of the confidence interval.

As a lead example, we consider the following partially linear regression (PLR) model (Chernozhukov et al., 2018):

$$Y = \theta_0 D + g_0(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon|X, D] = 0 \tag{1}$$

$$D = m_0(X) + \vartheta, \qquad \mathbb{E}[\vartheta|X] = 0 \tag{2}$$

Here, $Y$ is the outcome variable, $D$ is the policy/treatment variable of interest, $X$ consists of other controls, and $\varepsilon$ and $\vartheta$ are non-degenerate disturbances with corresponding variances $\sigma_\varepsilon^2$ and $\sigma_\vartheta^2$. If the treatment $D$ is exogenous conditional on the controls $X$, the coefficient $\theta_0$ can be interpreted as treatment effect of $D$ on $Y$. Equation (2) models the confounding effect of $X$ on $D$ via $m_0$, e.g. for a binary treatment $D$, the conditional expectation $m_0(X)$ corresponds to the propensity score. As $m_0$ characterizes the effect of confounders $X$ on $D$, it can be used to remove regularization bias. Not correctly accounting for all confounding factors, e.g. by not including all relevant confounders $X$, may lead to biased estimates of the target parameter $\theta_0$. In real world applications the confounding factors might be very complex and hard to observe or measure (e.g. product quality or medical information). Text and image data can be helpful to control for these complex confounding factors, for instance product images and descriptions are usually a good indicator of product quality. Consequently, causal models can benefit from text and image data to remove confounding bias. We leverage deep learning methods to fit functions representing the conditional expectations of the output variable $Y$ and the treatment variable $D$ given our set of covariates, incoporated in DML-based neural network architectures. Specifically, we define:

$$l_0(X) := \mathbb{E}[Y|X] \tag{3}$$

$$m_0(X) := \mathbb{E}[D|X] \tag{4}$$

To construct an orthogonalized score based on the nuisance components $\hat{\eta} = (\hat{l}, \hat{m})$, define the following expression:

$$\psi(W, \theta, \hat{\eta}) := \Big(Y - \hat{l}(X) - \theta\,(D - \hat{m}(X))\Big) \cdot (D - \hat{m}(X))\,,$$

where $W = (Y, D, X)$. The score $\psi$ is based on partialling-out the effects of the controls from treatment $D$ and outcome $Y$ (Chernozhukov et al., 2018) or also known as the R-Learner (Nie and Wager, 2021). Construction of an orthogonalized score ensures the necessary orthogonality for valid statistical inference. Finally, the estimate is computed as the solution to the following empirical moment equation:

$$0 = \frac{1}{n} \sum_{i=1}^{n} \psi(W, \hat{\theta}, \hat{\eta}_0)$$

The necessary assumptions involve bounding the difference between estimated nuisance functions ($\hat{m}$ and $\hat{l}$) and true functions ($m_0$ and $\ell_0$) in relation to the sample size $N$

$$\begin{aligned} \|\hat{m}(X) - m_0(X)\|_{P,2} \cdot (\|\hat{m}(X) - m_0(X)\|_{P,2} + \\ \|\hat{l}(X) - l_0(X)\|_{P,2}) \leq \delta_N N^{-1/2}, \end{aligned} \tag{5}$$

which requires high-quality machine learning estimates $\hat{m}$ and $\hat{l}$. Under this assumption and additional regularity conditions[1], the estimator $\hat{\theta}$ converges to the true parameter $\theta_0$ at a rate of $1/\sqrt{n}$ and is approximately normally distributed:

$$\sqrt{n}(\hat{\theta} - \theta_0) \to \mathcal{N}(0, \sigma^2)$$

This methodology forms the basis for treatment effect estimation in the presence of high-dimensional unstructured confounders. In modern research, the availability of unstructured data, such as images and text, has become ubiquitous. These data types offer a rich source of information that can contribute significantly to the estimation of causal effects. Incorporating unstructured data into the causal models such as the PLR as controls introduces several advantages. First, it allows for a more comprehensive representation of the confounding structure, capturing nuances that may be missed by solely relying on structured / tabular covariates. Second, more and more deep learning models are being developed which are tailored for unstructured data, such as transformers for images or LLMs for text, can be seamlessly integrated into the estimation process, further enhancing the accuracy of nuisance parameter estimation (Chernozhukov et al., 2018). Third, we would like to note that we focus specifically on the semi-parametric PLR model in this paper. In general, our methodology is basically also applicable to other nonparametric causal models that share the key ingredients of the DML framework (cf. Section 3).

In the subsequent sections, we elaborate on the methodology for leveraging unstructured data within the PLR framework and present simulation studies based on the semi-synthetic dataset. Additionally, the appendix includes an empirical illustration.

1. $\delta_N$ is a sequence of positive constants converging to zero. For details on the regularity conditions, see Chernozhukov et al. (2018).

## 2. LITERATURE REVIEW AND EXAMPLES

The use of unstructured data such as images and text as controls is crucial for identification and estimation of causal parameters, for example for estimating price elasticities, effects of medical treatments (Chan et al., 2018; Masukawa et al., 2022), or the effect of condensation trails on the climate (Ortiz et al., 2024). While unstructured data have been used for prediction tasks for some time, e.g. deep learning techniques for clinical risk predictions (Zhang et al., 2020), the use of text and images for causal inference has been a very recent development in the scientific literature. Text as outcome and treatment variable was discussed in Egami et al. (2018) and Sridhar and Blei (2022), but on a very high / conceptual level. We focus on text (and images) as confounders. Veitch et al. (2020) consider this setting and provide results for the consistency of the causal estimate, while we integrate it into the double machine learning framework to perform valid inference, i.e. constructing valid confidence intervals and test statistics for causal parameters. In a contemporary paper, Veljanovski and Wood-Doughty (2024) integrate only text in the double machine learning framework. Melnychuk et al. (2022) apply transformer to estimate treatment effects with tabular data and time-varying covariates, but do not provide inference results. Schulte et al. (2025) consider inference with pre-trained networks. The recent literature on the use of text for causal inference is nicely summarized in Feder et al. (2022) and Jiao et al. (2024). There are also approaches to use images in causal inference (Jerzak et al., 2023a,b,c), but we are, to the best of our knowledge, not aware of any study allowing for valid inference with images and consider our approach as the first to integrate both text and images in a *multimodal* double machine learning framework.

## 3. DOUBLE MACHINE LEARNING FOR TABULAR DATA

The double machine learning method, as proposed by Chernozhukov et al. (2018), is a framework that aims to provide valid statistical inference in structural equation models while leveraging the predictive power of machine learning methods for potentially high-dimensional and non-linear nuisance functions. The DML framework consists of three key ingredients: 1. Neyman orthogonality, 2. High-quality machine learning estimation and 3. Sample splitting. Neyman orthogonality ensures that the score function $\psi(W, \theta, \hat{\eta})$ to estimate the target parameter $\theta_0$ is insensitive towards the plug-in estimates from the nuisance learners $\hat{\eta}$. This is specifically relevant for machine learning algorithms such as neural networks, as these usually trade of variance and bias to achieve high-quality predictions via regularization. High-quality machine learning estimation involves using state-of-the-art machine learning algorithms to estimate the nuisance functions in Equations 3 and 4, ensuring that they are estimated as accurately as possible. Sample splitting is employed to separate the data into estimation and inference samples, which helps to avoid overfitting, and hence to achieve valid statistical inference.

The DML framework has gained attention in various domains, including social sciences, computer science, medicine, biostatistics, and economics and finance, because of its ability to address the challenges of causal inference with high-dimensional data. Furthermore, the DML framework has been implemented in both `R` and `Python` programming languages as `DoubleML` Bach et al. (2024a), making it accessible to many users in academic and industry research. In summary, the double machine learning method, as described by Chernozhukov et al. (2018), provides a robust framework for valid statistical inference for tabular data or structured data in structural equation models by integrating high-quality machine learning estimation with sample splitting. Its versatility

and applicability across various domains make it a valuable tool for addressing causal inference challenges with complex, high-dimensional controls.

## 4. DOUBLE MACHINE LEARNING FOR TEXT AND IMAGES

In the standard causal inference literature, the controls $X$ are tabular, as discussed in the previous section, and can be represented as $X_{\text{tab}} = (X_1, \ldots, X_p)$ such that $X_{\text{tab}} \in \mathbb{R}^p,\ p \in \mathbb{N}$. However, if the controls are unstructured, e.g. in the simple form of RGB images, $X$ can be represented as a tensor such that $X_{\text{img}} \in \mathbb{R}^{3\times h\times w},\ h, w \in \mathbb{N}$. For controls in text form, $X$ can be defined as an input representation matrix including the token, segmentation and position embeddings (Devlin et al., 2019) and can therefore be written as $X_{\text{txt}} \in \mathbb{R}^{3\times S},\ S \in \mathbb{N}$ where $S$ denotes the sequence length of the input sentence.

If all input modalities are to be used together as controls, $X$ can be represented as a set consisting of $X = (X_{\text{tab}}, X_{\text{txt}}, X_{\text{img}})$ with $X_{\text{tab}} \in \mathbb{R}^p$ for the tabular input, $X_{\text{txt}} \in \mathbb{R}^{3\times S}$ for the text input and $X_{\text{img}} \in \mathbb{R}^{3\times h\times w}$ for the image input with $p, h, w, S \in \mathbb{N}$. This results in a high-dimensional input vector, which cannot be directly used for nuisance estimation, but rather can be modeled as input for Deep Learning architectures, resulting in low-dimensional representations. Two possible confounding structures with text and image data are illustrated in Figure 1. As highlighted in the

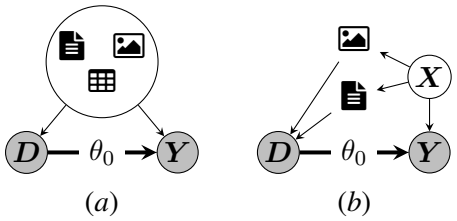



Figure 1: Examples of directed acyclic graphs (DAGs) with image and text confounding. (a) Direct confounding via image, text and tabular data as common cause. (b) Treatment take-up is driven by text and images. All backdoor paths are blocked by conditioning on both image and text data.

previous sections, the DML approach relies on high-quality estimates for the nuisance elements $m_0(X)$ and $l_0(X)$, such that the mean squared error (or product of root mean squared errors in Equation 5) converges fast enough. For tabular data these rates are achievable via standard machine learning algorithms such as e.g. lasso (Bickel et al., 2009) or boosting (Luo et al., 2022). Although for the state-of-the-art architectures formal results on the rates of convergence are still missing, deriving such results has currently high priority in research and a lot of progress in understanding the properties of different neural network architectures has been achieved recently. For feed forward neural networks Schmidt-Hieber (2020), Kohler and Langer (2021) and Farrell et al. (2021) provide results on the rates and show that the required theoretical rates are achievable. For convolutional neural nets Yang et al. (2024) provide theoretical guarantees. First results for attention mechanism are provided in Vasudeva et al. (2024) and for transformer models in Furuya et al. (2024), Gurevych et al. (2021) and Makkuva et al. (2024). Of course, for advanced architectures still some work needs to be done, but the results point into the right direction. Moreover, more complex neural network architectures as e.g. transformers achieve stunning predictive performance in the respective

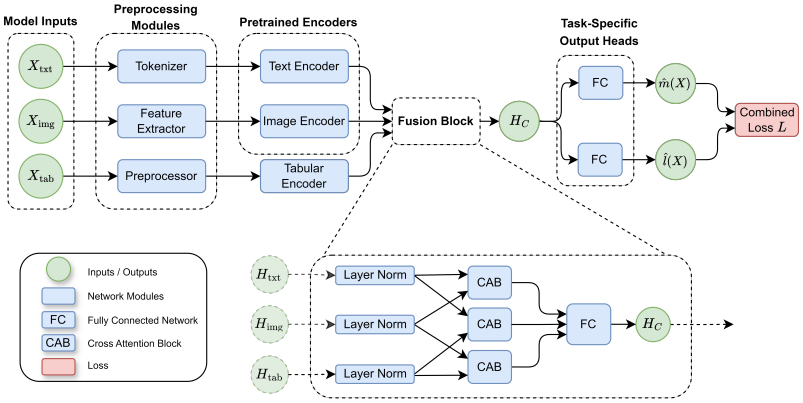



Figure 2: High-Level PLR Model Architecture. Both nuisance components are trained simultaneously on the confounding embedding $H_C$ with a combined loss.

regression or classification tasks, suggesting credibly fast convergence and therefore likely fulfilling the conditions for the double machine learning framework. Nevertheless, it should be highlighted, that the rate requirements impose regularity assumptions such as smoothness on the underlying data generating process. Veljanovski and Wood-Doughty (2024) perform a simulation study which indicates that the desired rates are actually achieved by modern architectures.

### 4.1. Model

To estimate the nuisance functions according to Equations 3 and 4 with the set of multimodal controls $X$, we propose a tailored neural network architecture which is adapted to the DML framework. The focus of this work is on multimodal models due to their promising results across different tasks. Utilizing multimodal data fusion is particularly suited for achieving the objective of estimating the causal parameter $\theta_0$ incorporating confounding influences sourced from tabular features, text and images. Multimodal models are a class of ML models that can effectively handle input data from different modalities such as text, image, video or audio. These models combine information from multiple modalities to improve their predictive power and achieve a better performance compared to individual modalities systems (Rahate et al., 2022). Multimodal data fusion seeks to extract and merge contextual information from multiple modalities in order to enhance decision-making. This is done by taking advantage of the complementary strengths of each modality (Lipkova et al., 2022). Multimodal models can, for example, combine semantic knowledge gained from texts with knowledge of spatial structures obtained from images to learn joint representations of images and texts (Miller et al., 2020). The objective of a multimodal model is to combine features of various modalities (Lee and Rho, 2022). The architecture of these models can be diverse and includes, for example, neural networks capable of processing and analyzing each modality, followed by the fusion block that concatenates the information across modalities (Miller et al., 2020).

### 4.2. Deep Learning Architecture and Implementation Details

The high-level architecture of our model for the single treatment case is shown in Figure 2. The integration of the modalities occurs through a middle fusion approach, whereby the text and image data are combined at an intermediate representation level, utilizing the embedding output. In our architecture, each modality encoder (text, image, tabular) first produces a modality-specific embedding $H_{\text{txt}}, H_{\text{img}}, H_{\text{tab}}$. Rather than simply concatenating these embeddings, we interpose a series of Cross-Attention Blocks (CABs) that allow each modality to dynamically "attend" to the other two. Concretely, in each CAB the normalized queries from one modality interact via scaled dot-product attention with keys and values drawn from another modality. By stacking CABs so that each modality both queries and is queried by the others, we learn pairwise and higher-order cross-modal correlations before projecting into the final joint representation $H_C$. Finally, the embedding $H_C \in \mathbb{R}^{d_E}$ provides a $d_E$-dimensional representation of the unstructured input data that can be used to make predictions or classify the input (Wolf et al., 2020).
We use pre-trained transformer-based models such as BERT (Vaswani et al., 2023) or variants like the robustly optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019) or LLMs like Llama (Touvron et al., 2023) for handling the text data. For the image processing, we rely on transformer-based models such as BEIT (Bao et al., 2022) or Vision Transformers (VITs) (Dosovitskiy et al., 2021). Using models like the SAINT model (Somepalli et al., 2021) for tabular data appeared to be satisfactory. It is crucial to closely monitor losses of each nuisance component

$$\|Y - \hat{l}(X)\|_{P,2} \leq \|l_0(X) - \hat{l}(X)\|_{P,2} + \sigma_\epsilon$$
$$\|D - \hat{m}(X)\|_{P,2} \leq \|m_0(X) - \hat{m}(X)\|_{P,2} + \sigma_\vartheta$$

to ensure high-quality predictions. These considerations contribute to the robustness and effectiveness of the models applied to both unstructured and tabular data, emphasizing the importance of vigilance over nuisance losses. The combined loss function for model training is defined as the product of root mean squared errors, which is directly motivated by the theoretical requirement for valid inference in Double Machine Learning:

$$L = \|D - \hat{m}(X)\|_{P_n,2} \|Y - \hat{l}(X)\|_{P_n,2}.$$

Usually the DML framework relies on cross-fitting to ensure the nuisance predictions are not overfitted, avoiding bias of the target parameter. In principle cross-fitting is also possible for neural networks but would increase the computational costs excessively. Instead, we rely on sample splitting, such that the neural network is trained on a training set and the estimation of the target parameter $\theta_0$ is performed on a seperate test set. Recent simulation results by Bach et al. (2024b) suggest that the performance of DML with a single sample split is comparable to a cross-fitting approach in large sample sizes.

## 5. SIMULATION STUDY

The validation of the proposed estimator is a crucial step in ensuring their accuracy and reliability. Unlike predictive models, the performance of causal models is not straightforward to evaluate in real world applications. The evaluation of causal estimation approaches is generally complicated by the fact that the true causal effect (unlike true labels for predicted outcomes) is not observable, which requires the use of synthetic or semi-synthetic data. In this simulation study, we generate a

semi-synthetic dataset with a known treatment effect parameter. We document a generally inherent challenge of simulating multimodal data for causal estimation: Generating credible confounding through unstructured data makes it very hard to uncover the true causal effect parameter. In our data generating process, the confounding operates through labels of the supervised learning tasks such as image classification, which cannot be perfectly predicted by the neural nets. Consequently, a part of the imposed confounding remains unexplained, which prevents exact identification and estimation of the causal parameter. Hence, we consider the true causal parameter as an ideal, but generally infeasible statistical estimate. Our analysis will include simulations to evaluate prediction performance for treatment and outcome, joint loss function values, and variance and bias of the causal effect estimate. Additionally, the appendix contains a short empirical illustration.

### 5.1. Simulating Confounding with Text and Images

To evaluate the performance of our model, we generate a semi-synthetic dataset according to the underlying model in Equations 1 and 2

$$Y = \theta_0 D + \tilde{g}_0(\tilde{X}) + \varepsilon,$$
$$D = \tilde{m}_0(\tilde{X}) + \vartheta,$$

where $\tilde{X} = (\tilde{X}_{\text{tab}}, \tilde{X}_{\text{txt}}, \tilde{X}_{\text{img}})$ with the following additive structure

$$\tilde{g}_0(\tilde{X}) = \sum_{\text{mod} \in \{\text{tab,txt,img}\}} \tilde{g}_{\text{mod}}(\tilde{X}_{\text{mod}}),$$
$$\tilde{m}_0(\tilde{X}) = \sum_{\text{mod} \in \{\text{tab,txt,img}\}} \tilde{m}_{\text{mod}}(\tilde{X}_{\text{mod}})$$

and $\varepsilon, \vartheta \sim \mathcal{N}(0, 1)$.
Each of the three modality effects is based on a publicly available (simple) non-synthetic dataset which is usually used for classification and regression tasks. All datasets contain one target $\tilde{X}_{\text{mod}}$ (outcome or label) and features $X_{\text{mod}}$ (image, text etc.). As these datasets have been shown to work well with the respective predictive task, we generate the confounded treatment $D$ and outcome $Y$ based on the targets $\tilde{X}_{\text{mod}}$ of the three different datasets instead of the respective features $X_{\text{mod}}$. This ensures a credible confounding, especially for image and text data as the confounding effect depends on content of the high-level information of the image and text which can be considered as proxies of the latent confounding structure via labels.

The tabular data is sourced from the DIAMONDS dataset (Wickham, 2016), which includes various attributes of diamonds such as carat, cut, color, clarity, depth, table, price, and measurements (x, y, z). The logarithm of the price column $\tilde{X}_{\text{tab}}$ is used to simulate confounding and the remaining variables will be utilized as tabular controls $X_{\text{tab}}$. The tabular data is preprocessed (log-transformation etc.) and downsampled to create a dataset with $N = 50,000$ observations to maintain consistency with the other data modalities of the new semi-synthetic dataset.
The text data component of the semi-synthetic dataset is derived from the IMDB dataset (Maas et al., 2011), which is a collection of movie reviews with corresponding sentiment labels. This dataset is publicly available and has been widely used for sentiment analysis tasks in natural language processing research. The semi-synthetic dataset is based on the training and test sample of the IMDB dataset to match the size of the tabular and image datasets. The binary representation of

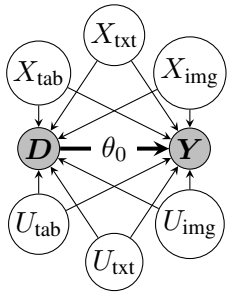



Figure 3: DAG for the semi-synthetic dataset. The confounding via the features $X = (X_{\text{tab}}, X_{\text{txt}}, X_{\text{img}})$ can be adjusted for, whereas the unexplained/noise parts $U = (U_{\text{tab}}, U_{\text{txt}}, U_{\text{img}})$ are unobserved.

the sentiment $\tilde{X}_{\text{txt}}$ is used to generate confounding while the review constitutes the text control $X_{\text{txt}}$ for the set of controls.

The image data component of the semi-synthetic dataset is sourced from the CIFAR-10 dataset (Krizhevsky et al., 2009), which is a well-known benchmark in the field of computer vision. The CIFAR-10 dataset consists of 60,000 32x32 color images in 10 different classes, with 6,000 images per class. The semi-synthetic dataset is based on the training set, which contains 50,000 images. A numerical representation of the image labels $\tilde{X}_{\text{img}}$ is used to obtain a confounding on $Y$ and $D$ while the images will be part of the set of controls as $X_{\text{img}}$.

The effect on the outcome $Y$ is generated via a standardized version of target variable

$$\tilde{g}_{\text{mod}}(\tilde{X}_{\text{mod}}) = (\tilde{X}_{\text{mod}} - \mathbb{E}[\tilde{X}_{\text{mod}}])/\sigma_{\tilde{X}_{\text{mod}}},$$
$$\text{mod} \in \{\text{tab}, \text{txt}, \text{img}\}$$

to balance the confounding impact of all modalities. Further, the impact on the treatment $D$ is defined via

$$\tilde{m}_{\text{mod}}(\tilde{X}_{\text{mod}}) = -\tilde{g}_{\text{mod}}(\tilde{X}_{\text{mod}}),$$
$$\text{mod} \in \{\text{tab}, \text{txt}, \text{img}\}$$

to ensure a strong confounding. Due to the negative sign and the additive structure, the confounding effect will ensure that higher outcomes $Y$ occur with lower treatment values $D$, creating a negative bias. Further, independence of all three original datasets and additive negative confounding results in a negative bias even if we only control for a subset of confounding factors.

The treatment effect is set to $\theta_0 = 0.5$ and both $\tilde{g}_0(X)$ and $\tilde{m}_0(X)$ are rescaled to ensure a signal-to-noise ratio of 2 for $Y$ and $D$ (given unit variances of the error terms).

A generally inherent challenge of this type of data generating processes is the dependency on the target of the modality $\tilde{X}_{\text{mod}}$, which might not be fully explained by the corresponding controls $X_{\text{mod}}$. For example, the price of the DIAMONDS dataset can not be perfectly predicted by carat, cut, color, etc., introducing a small part of confounding which can not be controlled for by using the tabular features $X_{\text{tab}}$ instead of $\tilde{X}_{\text{tab}}$ (logarithm of price), as shown in the DAG in Figure 3. Consequently, the estimate $\hat{\theta}$ might only be able to account for the part of confounding which can

be explained by the input features as

$$\tilde{X}_{\text{mod}} = \mathbb{E}[\tilde{X}_{\text{mod}}|X_{\text{mod}}] + U_{\text{mod}},$$

where $U_{\text{mod}}$ can not be controlled for. Nevertheless, since all modalities contribute a negative bias, the semi-synthetic dataset can be used as a benchmark with an oracle upper bound of an effect estimate of $\theta_0 = 0.5$. To evaluate the confounding one can evaluate a basic ordinary least squares model with outcome $Y$ on the treatment variable $D$ (excluding all confounding variables). The resulting effect estimate $\hat{\theta}_{\text{OLS}} = -0.4594$, can be interpreted as a lower bound for the effect estimate. Accordingly, all evaluated models should estimate the parameter between $-0.4594$ and $0.5$, where higher values indicate better bias correction (ignoring sampling uncertainty).
To further access the predictive performance of the nuisance models, we can rely on oracle predictions of

$$\begin{aligned}\tilde{m}_0(\tilde{X}) &:= \mathbb{E}[D|\tilde{X}] \\ \tilde{l}_0(\tilde{X}) &:= \mathbb{E}[Y|\tilde{X}] = \theta_0 \tilde{m}_0(\tilde{X}) + \tilde{g}_0(\tilde{X}).\end{aligned}$$

Evaluating the oracle predictions $\tilde{m}_0(\tilde{X})$ and $\tilde{l}_0(\tilde{X})$ results in the following upper bounds for the performance of the nuisance estimators

$$R^2(D, \tilde{m}_0(\tilde{X})) = 0.6713 \qquad\qquad R^2(Y, \tilde{l}_0(\tilde{X})) = 0.5845$$

on the whole dataset of $N = 50,000$ observations, which is to be expected due to the choice of signal-to-noise ratio (for the definition of $R^2$, see the appendix. Again, since models only have access to the features $X = (X_{\text{tab}}, X_{\text{txt}}, X_{\text{img}})$ instead of the targets $\tilde{X} = (\tilde{X}_{\text{tab}}, \tilde{X}_{\text{txt}}, \tilde{X}_{\text{img}})$, the above values represent upper bounds for $R^2$. The additive form and independent input features enable separate isolation and equally strong confounding effect for each modality. Nevertheless, this structure might not be very close to reality. In the appendix, we repeat the analysis with a similar semi-synthetic dataset with dependent modalities.

### 5.2. Results

In this section, we evaluate the performance of our proposed deep learning architecture against a comprehensive set of benchmark models on the semi-synthetic dataset. The evaluation is conducted as a Monte-Carlo experiment with $R = 100$ replications, each using a newly generated dataset of $N = 50,000$ observations. For computational efficiency, the DML-Deep model is trained on half the sample size ($N/2$). More information on simulation and training/hyperparameter tuning of the models can be found in the appendix under Simulation Details.

All models are designed to estimate the target parameter $\theta_0$ from the partially linear regression model. To ensure comparability, we use the `DoubleML` package (Bach et al., 2022) for all Double Machine Learning (DML) implementations, which only differ in how the nuisance functions $l(X) = E[Y|X]$ and $m(X) = E[D|X]$ are estimated.

We compare three classes of models:

The **Tabular-only Models** serve as baselines that ignore the unstructured data. These include standard Ordinary Least Squares with tabular controls (OLS-Tab) and DML models using only tabular features ($X_{\text{tab}}$). For the DML variants, we test both a simple linear model (DML-Tab (Lin))

and a tuned gradient boosting model (DML-Tab (Tree)) via the `LightGBM` package (Ke et al., 2017) as the nuisance learners.

The **Latent-Feature Models** represent an intermediate approach. Here, latent feature vectors are extracted from the text and image modalities using pre-trained, non-finetuned transformers. These vectors are then concatenated with the tabular features $X_{\text{tab}}$ and used as input for standard learners. This category includes OLS (OLS-Latent) and DML with both linear (DML-Latent (Lin)) and tree-based (DML-Latent (Tree)) nuisance estimators. This approach tests whether simply adding embeddings as features is sufficient to correct for confounding.

The **DML-Deep Model** is our proposed end-to-end architecture, as depicted in Figure 2. It processes multimodal features $X = (X_{\text{tab}}, X_{\text{txt}}, X_{\text{img}})$ simultaneously to generate out-of-sample predictions for $\hat{m}(X)$ and $\hat{l}(X)$. For our simulation, we use a DistilBERT model (Sanh et al., 2019) for text, a ViT model pretrained on ImageNet-21k (Wightman, 2019) for images, and a SAINT model (Somepalli et al., 2021) for tabular data.

To assess the predictive quality of the nuisance models, we define a relative $r^2$-score that benchmarks each model's predictive performance against the theoretical upper bound described in Section 5.1:

$$\begin{aligned} 0 \leq r^2(D, \hat{m}) &:= \frac{R^2(D, \hat{m}(X))}{R^2(D, \tilde{m}_0(\tilde{X}))} \leq 1 \\ 0 \leq r^2(Y, \hat{l}) &:= \frac{R^2(Y, \hat{l}(X))}{R^2(Y, \tilde{l}_0(\tilde{X}))} \leq 1. \end{aligned} \tag{6}$$

The primary results are presented in Table 1 and Figure 4.

Table 1: Key Performance Metrics by Method

| Method | Bias | RMSE | CI-Coverage | CI-Length |
|---|---|---|---|---|
| Unadjusted | -0.957 | 0.957 | 0.0% | 0.016 |
| OLS-Tab* | -0.820 | 0.820 | 0.0% | 0.017 |
| DML-Tab (Lin) | -0.820 | 0.820 | 0.0% | 0.016 |
| DML-Tab (Tree) | -0.819 | 0.819 | 0.0% | 0.016 |
| OLS-Latent* | -0.619 | 0.619 | 0.0% | 0.018 |
| DML-Latent (Lin) | -0.625 | 0.625 | 0.0% | 0.017 |
| DML-Latent (Tree) | -0.638 | 0.638 | 0.0% | 0.017 |
| DML-Deep[1] | **-0.002** | **0.012** | **91.0%** | 0.036 |
| *Oracle** | *0.000* | *0.004* | *95.0%* | *0.018* |

*: Predictive Performance evaluated in sample.
[1]: Evaluated on half the sample for efficiency.

The simulation was designed with a substantial negative confounding effect, providing a challenging scenario for bias correction.

The results in Table 1 clearly show the limitations of traditional methods. The **Tabular-only models** are heavily biased (Bias $\approx -0.82$), with their 95% confidence intervals failing to cover the true parameter $\theta_0$. The **Latent-Feature models** reduce this bias (Bias $\approx -0.62$), demonstrating that the unstructured data embeddings contain useful information about the confounders. However, this simple inclusion of latent features is insufficient for adequate bias control, and these models

also fail to achieve nominal coverage. In contrast, our **DML-Deep model** successfully eliminates the confounding bias, yielding a nearly unbiased estimate (Bias = $-0.002$) and achieving a 91% confidence interval coverage, which is close to the nominal 95% level.

Figure 4(*a*) in the appendix visualizes the distribution of the point estimates $\hat{\theta}$ across the 100 simulation runs. The distributions for the Tabular and Latent models are centered far from the true effect $\theta_0$, while the distribution for the DML-Deep model is centered directly on it. It is important to note that the confidence intervals for the biased models do not cover the true value.

Figure 4(*b*) in the appendix provides insight into *why* the DML-Deep model succeeds. It shows the relative $r^2$-scores for the nuisance functions, $l(X)$ and $m(X)$. The DML-Deep model achieves significantly higher predictive accuracy for both nuisance functions, approaching the performance of the theoretical oracle. This superior ability to model the confounding relationships is what enables the effective debiasing of the final treatment effect estimate, $\hat{\theta}$. In addition to the simulation results above, another simulation with dependent modalities and a real-world validation can be found in the appendix.

## 6. CONCLUSION

In this article, we extend the double machine learning framework to accommodate tabular data, text, and images as confounders, enabling valid causal inference in multimodal settings. With the growing availability of such data, we consider this a significant contribution to causal machine learning. Incorporating information from images or text can help reduce bias from unobserved confounding or improve estimator precision.

To validate our approach, we set up a new framework to generate semi-synthetic data with multimodal data that addresses inherent challenges in this type of simulation studies. The proposed method might be of independent interest to the scientific community. Our semi-synthetic numerical experiments demonstrated the model's ability to incorporate the confounding that is contained in the text and image data. The performance substantially improved as compared to a benchmark model that does not account for this information. While our model does not always recover the true causal effect, this reflects limitations of simulated unstructured data, not the architecture itself.

Our method is the first for valid (causal) inference with multimodal data as confounder. Moreover, our approach could also be extended to other kinds of unstructured data, like graphs, networks, audio or video data.

## Acknowledgments

This work has been supported by the German Federal Ministry of Research, Technology and Space (grant: 01IS24082). For computation the HPC-cluster Hummel-2 at University of Hamburg was used. The cluster was funded by Deutsche Forschungsgemeinschaft (project number 498394658).

## References

Philipp Bach, Victor Chernozhukov, Malte S Kurz, and Martin Spindler. Doubleml: an object-oriented implementation of double machine learning in Python. *The Journal of Machine Learning Research*, 23(1):2469–2474, 2022.

Philipp Bach, Malte S Kurz, Victor Chernozhukov, Martin Spindler, and Sven Klaassen. Doubleml: An object-oriented implementation of double machine learning in r. *Journal of Statistical Software*, 108:1–56, 2024a.

Philipp Bach, Oliver Schacht, Victor Chernozhukov, Sven Klaassen, and Martin Spindler. Hyperparameter tuning for causal inference with double machine learning: A simulation study. In *3rd Causal Learning and Reasoning*, 2024b. URL https://openreview.net/forum?id=h0ecxkungr.

Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *CoRR*, abs/2106.08254, 2021. URL https://arxiv.org/abs/2106.08254.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732, 2009. doi: 10.1214/08-AOS620. URL https://doi.org/10.1214/08-AOS620.

Alex J. Chan, Isabel Chien, Edward T. Moseley, Saad Salman, Sarah Kaminer Bourland, Daniela Lamas, Anne M. Walling, and James A. Tulsky. Deep learning algorithms to identify documentation of serious illness conversations during intensive care unit admissions. *Palliative Medicine*, 2018. doi: 10.1177/0269216318810421.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. How to make causal inferences using texts, 2018.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022. doi: 10.1162/tacl_a_00511. URL https://aclanthology.org/2022.tacl-1.66.

Dylan J. Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879 – 908, 2023. doi: 10.1214/23-AOS2258. URL https://doi.org/10.1214/23-AOS2258.

Takashi Furuya, Maarten V. de Hoop, and Gabriel Peyré. Transformers are universal in-context learners, 2024. URL https://arxiv.org/abs/2408.01367.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data?, 2022.

Iryna Gurevych, Michael Kohler, and Gözde Gül Sahin. On the rate of convergence of a classifier based on a transformer encoder, 2021. URL https://arxiv.org/abs/2111.14574.

W.K. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer Series in Statistics. Springer Berlin Heidelberg, 2012. ISBN 9783642171468. URL https://books.google.de/books?id=wqX7CAAAQBAJ.

Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar S. Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *CoRR*, abs/2012.06678, 2020. URL https://arxiv.org/abs/2012.06678.

Connor T. Jerzak, Fredrik Johansson, and Adel Daoud. Estimating causal effects under image confounding bias with an application to poverty in africa, 2023a.

Connor T. Jerzak, Fredrik Johansson, and Adel Daoud. Image-based treatment effect heterogeneity, 2023b.

Connor T. Jerzak, Fredrik Johansson, and Adel Daoud. Integrating earth observation data into causal inference: Challenges and opportunities, 2023c.

Licheng Jiao, Yuhan Wang, Xu Liu, Lingling Li, Fang Liu, Wenping Ma, Yuwei Guo, Puhua Chen, Shuyuan Yang, and Biao Hou. Causal inference meets deep learning: A comprehensive survey. *Research*, 7:0467, 2024. doi: 10.34133/research.0467. URL https://spj.science.org/doi/abs/10.34133/research.0467.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.,

2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, 2009.

Seung Jae Lee and Mina Rho. Multimodal deep learning applied to classify healthy and disease states of human microbiome. *Scientific Reports*, 12(1):824, 2022.

Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A System for Massively Parallel Hyperparameter Tuning. *arXiv (Cornell University)*, 1 2018. doi: 10.48550/arxiv.1810.05934. URL https://arxiv.org/abs/1810.05934.

Jana Lipkova, Richard J Chen, Bowen Chen, Ming Y Lu, Matteo Barbieri, Daniel Shao, Anurag J Vaidya, Chengkuan Chen, Luoting Zhuang, Drew FK Williamson, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer cell*, 40(10):1095–1110, 2022.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. Timelms: Diachronic language models from twitter, 2022.

Ye Luo, Martin Spindler, and Jannis Kück. High-dimensional $l_2$boosting: Rate of convergence, 2022.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Martin Jaggi, Hyeji Kim, and Michael Gastpar. Attention with markov: A framework for principled analysis of transformers via markov chains, 2024. URL https://arxiv.org/abs/2402.04161.

Kento Masukawa, Maho Aoyama, Shinichiroh Yokota, Junya Nakamura, Ryoka Ishida, and Masaharu Nakayama. Machine learning models to detect social distress, spiritual pain, and severe physical psychological symptoms in terminally ill patients with cancer from unstructured text data in electronic medical records. *Palliative Medicine*, 2022. doi: 10.1177/02692163221105595.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes, 2022.

Stuart J Miller, Justin Howard, Paul Adams, Mel Schwan, and Robert Slater. Multi-modal classification using images and text. *SMU Data Science Review*, 3(3):6, 2020.

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

Irene Ortiz, Ermioni Dimitropoulou, Pierre de Buyl, Nicolas Clerbaux, Javier García-Heras, Amin Jafarimoghaddam, Hugues Brenot, Jeroen van Gent, Klaus Sievers, Evelyn Otero, Parthiban Loganathan, and Manuel Soler. Satellite-based quantification of contrail radiative forcing over europe: A two-week analysis of aviation-induced climate effects, 2024. URL https://arxiv.org/abs/2409.10166.

Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81: 203–239, 2022.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv (Cornell University)*, 1 2019. doi: 10.48550/arxiv.1910.01108. URL https://arxiv.org/abs/1910.01108.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020. doi: 10.1214/19-AOS1875. URL https://doi.org/10.1214/19-AOS1875.

Rickmer Schulte, David Rügamer, and Thomas Nagler. Adjustment for confounding using pre-trained representations. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 53557–53580. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/schulte25a.html.

Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C. Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training, 2021.

Dhanya Sridhar and David M. Blei. Causal inference from text: A commentary. *Science Advances*, 8(42):eade6585, 2022. doi: 10.1126/sciadv.ade6585. URL https://www.science.org/doi/abs/10.1126/sciadv.ade6585.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Bhavya Vasudeva, Puneesh Deora, and Christos Thrampoulidis. Implicit bias and fast convergence rates for self-attention, 2024. URL https://arxiv.org/abs/2402.05738.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In Jonas Peters and David Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 919–928. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/veitch20a.html.

Marko Veljanovski and Zach Wood-Doughty. Doublelingo: Causal estimation with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 799–807, 2024.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Yunfei Yang, Han Feng, and Ding-Xuan Zhou. On the rates of convergence for learning with convolutional neural networks, 2024. URL https://arxiv.org/abs/2403.16459.

Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. Cambridge University Press, 2023. https://D2L.ai.

Dongdong Zhang, Changchang Yin, Jucheng Zeng, Xiaohui Yuan, and Ping Zhang. Combining structured and unstructured data for predictive models: A deep learning approach. *BMC Medical Informatics and Decision Making*, 2020. doi: 10.1186/s12911-020-01297-6.

## Appendix A. Definitions

Let $V$ be a random variable and $V_1, \ldots, V_n$ iid. realizations of $V$.
Define

$$\mathbb{E}_n[V] := \frac{1}{n} \sum_{i=1}^{n} V_i$$

and the correspondingly

$$\|V\|_{P,2} := \mathbb{E}[V^2]$$
$$\|V\|_{P_n,2} := \mathbb{E}_n[V^2].$$

Further, define

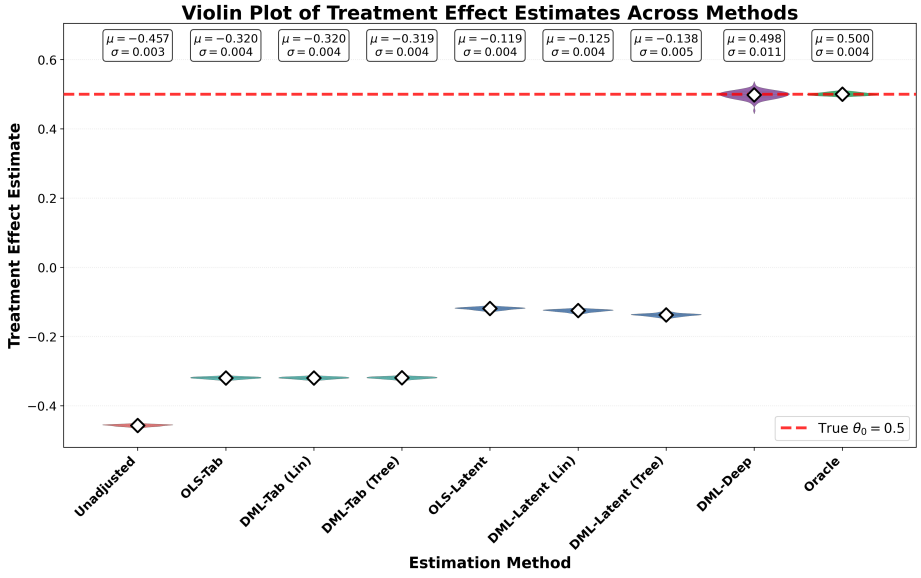
$$R^2(V, \hat{V}) = 1 - \frac{\sum_{i=1}^{n}(V_i - \hat{V}_i)^2}{\sum_{i=1}^{n}(V_i - \mathbb{E}_n[V])^2}.$$

for two random variables $V$ and $\hat{V}$.

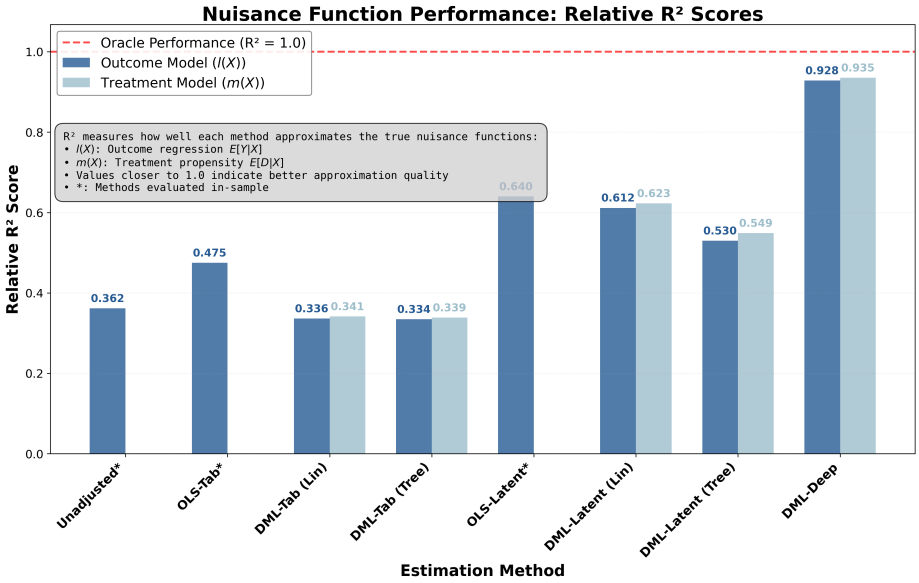
## Appendix B. Simulation Details

### B.1. Semi-Synthetic Dataset with Independent Modalities

This subsection includes further information regarding the semi-synthetic dataset. As described in the main paper, figure 4(*a*) visualizes the distribution of the point estimates $\hat{\theta}$ across the 100 simulation runs and figure 4(*b*) shows the relative $r^2$-scores for the nuisance functions, $l(X)$ and $m(X)$.




(*a*)




(*b*)

Figure 4: Parameter estimates. (a) Point estimates $\hat{\theta}$. $\theta_0$ represents the upper bound. (b) The relative $r^2$-scores across methods.

Due to the distribution of the noise and centering of the confounding components it should hold

$$\mathbb{E}[Y] = \mathbb{E}[D] = 0$$
$$\mathrm{Var}(Y) = \mathrm{Var}(D) = 3,$$

which can be observed in the corresponding descriptive statistics. This subsection includes further information regarding the semi-synthetic dataset. The data generating process is given by

$$\begin{aligned} Y &= \theta_1 D_1 + g_1(X) + \epsilon, \\ D_1 &= m_1(X) + \eta, \end{aligned}$$

where $X = (X_{\text{tab}}, X_{\text{txt}}, X_{\text{img}})$ consists of tabular, text and image data. The treatment effect is set to $\theta_1 = 0.5$. The nuisance functions $g_1(X)$ and $m_1(X)$ are constructed as a sum of modality-specific components

$$\begin{aligned} g(X) &= g_{\text{tab}}(X_{\text{tab}}) + g_{\text{txt}}(X_{\text{txt}}) + g_{\text{img}}(X_{\text{img}}) \\ m(X) &= m_{\text{tab}}(X_{\text{tab}}) + m_{\text{txt}}(X_{\text{txt}}) + m_{\text{img}}(X_{\text{img}}), \end{aligned}$$

where each component is standardized to have zero mean and unit variance. The components are based on features from the original datasets: $g_{\text{tab}}$ on the price of the diamond, $g_{\text{txt}}$ on the review sentiment and $g_{\text{img}}$ on the image class label. To induce confounding, the components for the treatment regression are set to be the negative of the outcome components, i.e., $m_{\text{modality}}(X_{\text{modality}}) = -g_{\text{modality}}(X_{\text{modality}})$ for each modality.

The final nuisance functions $g_1(X)$ and $m_1(X)$ are scaled versions of $g(X)$ and $m(X)$ to achieve pre-defined signal-to-noise ratios for the outcome $Y$ and the treatment $D_1$. With signal-to-noise ratios $SNR_Y = 2$ and $SNR_D = 2$, the nuisance functions are

$$\begin{aligned} m_1(X) &= \sqrt{\frac{SNR_D}{\text{Var}(m(X))}} \cdot m(X) \\ g_1(X) &= c_g \cdot g(X), \end{aligned}$$

To achieve the desired signal-to-noise ratio for the outcome, $\text{SNR}_Y$, the final nuisance function $g_1(X)$ is a scaled version of the composite function $g(X)$, such that $g_1(X) = c_g g(X)$. The scaling factor $c_g$ is determined by solving the following quadratic equation, which arises from the variance definition of the signal component:

$$\text{Var}(\theta_1 D_1 + c_g g(X)) = \text{SNR}_Y$$

This expands to:

$$\underbrace{(\text{Var}(g(X)))}_{a} c_g^2 + \underbrace{(2\theta_1 \text{Cov}(g(X), D_1))}_{b} c_g + \underbrace{(\theta_1^2 \text{Var}(D_1) - \text{SNR}_Y)}_{c} = 0$$

The scaling factor $c_g$ is then obtained by applying the quadratic formula and selecting the positive root:

$$c_g = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

The noise terms are drawn from a standard normal distribution, $\epsilon, \eta \sim \mathcal{N}(0, 1)$.

For the noise variation experiment, we generate 100 new samples of the treatment and outcome variables, denoted as $D_i$ and $Y_i$ for $i = 1, \ldots, 100$. For each sample $i$, we draw new noise terms $\epsilon_i, \eta_i \sim \mathcal{N}(0, 1)$ and compute

$$\begin{aligned} D_i &= m_1(X) + \eta_i \\ Y_i &= \theta_1 D_i + g_1(X) + \epsilon_i. \end{aligned}$$

This results in a dataset with 100 different realizations of the treatment and outcome for the same set of covariates $X$ and nuisance functions $m_1(X)$ and $g_1(X)$.

Table 2: Semi-synthetic dataset example descriptives statistics regarding the first realized treatment and outcome variable.

| | $Y_1$ | $D_1$ |
|---|---|---|
| count | 50000 | 50000 |
| mean | 0.000926 | -0.006157 |
| std | 1.735310 | 1.730348 |
| min | -6.543028 | -6.995684 |
| 25%-quantile | -1.207339 | -1.199605 |
| 50%-quantile | -0.004285 | 0.007363 |
| 75%-quantile | 1.191172 | 1.189492 |
| max | 6.514348 | 6.249006 |

The oracle root mean squared errors for the nuisance components are

$$\|D - \tilde{m}_0(\tilde{X})\|_{P_n,2} = 0.9994$$
$$\|Y - \tilde{l}_0(\tilde{X})\|_{P_n,2} = 1.1198.$$

The DML-Deep Model was tuned and trained on a GPU Node with the following configurations:

Table 3: GPU Node Infrastructure Specifications

| **Component** | **Specification** |
|---|---|
| GPU | 8 × NVIDIA H100 (80 GB HBM3) |
| CPU | 2 × AMD EPYC 9334 (32-Core) |
| System RAM | 1152 GB |

All other Methods / Models were tuned / trained on a CPU Node with the following configurations:

Table 4: CPU Node Infrastructure Specifications

| **Component** | **Specification** |
|---|---|
| CPU | 2 × AMD EPYC 9654 (96-Core) |
| System RAM | 768 GB |

Hyperparameters were selected by tuning using the **Validation Combined Loss** as criterion over the following Search Space with 10 trials using the `ray.tune.schedulers.AsyncHyperBandScheduler` (Li et al., 2018).

Table 5: Hyperparameter Search Space

| **Hyperparameter** | **Search Space** |
|---|---|
| Embedding Dimension | $\{256, 512\}$ |
| Fusion Head Layers | $\{[128], [256, 128], [512, 256]\}$ |
| Learning Rate | Log-Uniform($10^{-6}, 8 \times 10^{-4}$) |
| Embedding Dropout | Uniform(0.1, 0.4) |
| Learning Rate Scheduler | {None, ReduceLROnPlateau, CosineAnnealingWarmRestarts} |
| Weight Decay | Log-Uniform($2 \times 10^{-4}, 10^{-3}$) |
| Momentum | Uniform(0.80, 0.99) |
| Cross Attention Heads | $\{16, 32\}$ |
| Optimizer | {SGD, RMSprop, Adam, Adagrad, Adadelta, Adamax} |
| Text Model | {RoBERTa, DistilBERT, BERT} |
| Image Model | {ViT-Base, ViT-Small, BEiT} |

The following hyperparameters were finally selected for the simulation study:

- Optimizer: SGD (PyTorch)
  - Learning Rate: $0.00003$
  - Weight Decay: $0.0002$
  - Momentum: $0.90$
  - Scheduler: ReduceLROnPlateau (PyTorch)
- Training Precision: 16-mixed
- Dropout: $0.30$
- Batch Size: $12$
- Number of Heads in CAB: $16$
- Embedding Dimension: $256$
- Output Heads Dimensions: $\{256; 128\}$
- Early Stopping, Patience: $10$
- Max. Epochs: $100$

The models used are subject to the following licenses:

- All three models used in the experiment are licensed under the Apache 2.0 license.

The datasets used are subject to the following licenses:

- The DIAMONDS dataset is licensed under CC BY-NC-SA 4.0 license.
- The CIFAR-10 dataset is not licensed.
- The IMBD dataset is not licensed.

### B.2. Semi-Synthetic Dataset with Dependent Modalities

We further consider a semi-synthetic dataset with dependent modalities. The dataset is based on amazon marketplace data and contains a product description, a product image and several tabular features such as price or average ratings (all tabular features can be seen in Table 6). The corresponding semi-synthetic dataset is generated similar to the semi-synthetic dataset in Section 5.1

$$Y = \theta_0 D + \tilde{g}_0(\tilde{X}) + \varepsilon,$$
$$D = \tilde{m}_0(\tilde{X}) + \vartheta,$$

where $\tilde{X} = \log(\text{price})$ and $\varepsilon, \vartheta \sim \mathcal{N}(0,1)$. The outcome $Y$ and treatment $D$ are confounded via standardized versions of the of the logarithm of the price of the items via

$$\tilde{g}_0(\tilde{X}) = \frac{\log(\text{price}) - \mathbb{E}[\log(\text{price})]}{\sigma_{\log(\text{price})}}$$

and

$$\tilde{m}_0(\tilde{X}) = -\tilde{g}_0(\tilde{X}).$$

As in the previous dataset the treatment effect is set to $\theta_0 = 0.5$ and the confounding effects are scaled to ensure a signal-to-noise ratio of 2 for $Y$ and $D$ (given unit variances of the error terms).

Table 6: Semi-synthetic dataset descriptive statistics for the tabular data.

| | Y | D1 | log(price) | Rating | Review Count | Count FBA | Count FBM |
|---|---|---|---|---|---|---|---|
| count | 9207 | 9207 | 9207 | 9207 | 9207 | 9207 | 9207 |
| mean | 0.010487 | -0.000052 | 3.063914 | -0.000000 | 0.000000 | 0.000000 | 0.000000 |
| std | 1.720295 | 1.729765 | 0.782993 | 1.000054 | 1.000054 | 1.000054 | 1.000054 |
| min | -10.093060 | -6.844529 | -1.427116 | -9.344589 | -0.347623 | -0.594732 | -0.626514 |
| 25%-quantile | -1.082308 | -1.046593 | 2.638343 | -0.273828 | -0.323614 | -0.263314 | -0.626514 |
| 50%-quantile | -0.030014 | 0.023748 | 2.999226 | 0.259746 | -0.265392 | -0.263314 | -0.376738 |
| 75%-quantile | 1.094071 | 1.075214 | 3.485232 | 0.526533 | -0.101230 | -0.263314 | 0.122812 |
| max | 8.084152 | 9.062393 | 6.640359 | 1.326894 | 18.271398 | 17.964712 | 9.614275 |

Given the price (or $\log(\text{price})$) of the items the adjustment for confounding is quite straight forward and leads to well-behaved treatment effect estimates. For example, evaluating an ordinary least squares regression on the complete generated dataset with $\log(\text{price})$ as confounding results in an estimate $\hat{\theta}_{\text{oracle}} = 0.4919$ with a confidence interval of $[0.472, 0.512]$ covering the true parameter $\theta_0 = 0.5$. Instead, if we do not condition on any confounding variables the effect is heavily biased with an estimate of

$$\hat{\theta}_{\text{OLS}} = -0.4494$$

and a confidence interval $[-0.467, -0.431]$. Again, since the confounding introduces a negative bias, the semi-synthetic dataset allows the comparison of different models based resulting treatment effect estimates. Generally, all effect estimates should lie between $-0.4494$ and $0.5$, where higher

values indicate a better bias correction. We evaluate our model on all other available features, including product descriptions and images, except for price as they are able to explain a certain amount of variation within the price. The model will suffer the same identification issues as highlighted in Section 5.1. The price can not be perfectly predicted from the available features, such that a certain part of the confounding bias can not be accounted for (generally this effect will be much larger as in Section 5.1 as the signal to noise ratio in the price with respect to the input features is much smaller). To still be able to evaluate how much variation of $\log(\text{price})$, our model is able to capture we rely on the same relative $r^2$-score defined in Equation (6) based on the explained variation in treatment and outcome based on $\log(\text{price})$

$$R^2(D, \tilde{m}_0(\tilde{X})) = 0.6685$$
$$R^2(Y, \tilde{l}_0(\tilde{X})) = 0.5799$$

based on $N = 9,207$ observations. These $R^2$ values are upper bounds for the predictive performance for our model and highlight how much variation in treatment and outcome can be explained by the price. The $r^2$-score then refers to the explained variation of the model relative to the upper bound. As in Section 5.2, we evaluate the Baseline Model, Deep Model and Embedding Model.

The *Baseline Model* is a standard DML approach, relying only on tabular data. The estimation of the nuisance elements is based on the `LightGBM` package (Ke et al., 2017) only using the features $X_{\text{tab}}$. Due to the construction of the semi-synthetic data, the resulting estimate should be highly biased, as the model is only able to account for the part of the confounding, which is generated via the tabular data.

The *Embedding Model* also relies on the proposed architecture in Figure 2, but does not use the generated predictions directly. Instead the generated embedding $H_C$ of the unstructured modalities is used together with the tabular features $X_{\text{tab}}$ as input for a boosting algorithm. Since neural networks are often outperformed by tree based models, such as gradient boosted trees, on tabular data (Grinsztajn et al., 2022), the model might perform better on the tabular part of the data, while still accounting for the information contained in the image and text components.

As pretrained models, we rely on the BEiT model (Bao et al., 2021), RoBERTa model (Loureiro et al., 2022) and the TabTransformer model (Huang et al., 2020).

To evaluate the impact of using image and text data in addition to tabular data, we compare the relative $r^2$-score and the effect estimates $\hat{\theta}$ of the Deep Model and Embedding Model further remove one of the unstructured modalities as input features and re-estimate the treatment effect. The results are presented in Table 7.

Table 7: Results for the semi-synthetic dataset with dependent modalities. Reported: mean ± sd. over five random train-test splits.

| Modalites | Model | $r^2(Y, \hat{l})$ | $r^2(D, \hat{m})$ | $\hat{\theta}$ |
|---|---|---|---|---|
| tab | Baseline | $0.1664 \pm 0.0348$ | $0.1792 \pm 0.0274$ | $-0.3508 \pm 0.0182$ |
| tab & img | Deep | $0.5098 \pm 0.0692$ | $0.4915 \pm 0.0269$ | $-0.1718 \pm 0.0205$ |
| | Embedding | $0.4548 \pm 0.073$ | $0.4603 \pm 0.0223$ | $-0.1317 \pm 0.0171$ |
| tab & txt | Deep | $0.5743 \pm 0.012$ | $0.6081 \pm 0.0077$ | $-0.1515 \pm 0.0199$ |
| | Embedding | $0.5682 \pm 0.0193$ | $0.6015 \pm 0.0101$ | $-0.0948 \pm 0.0275$ |
| tab & txt & img | Deep | $\mathbf{0.5931 \pm 0.0101}$ | $\mathbf{0.6254 \pm 0.0464}$ | $-0.1089 \pm 0.0517$ |
| | Embedding | $0.5453 \pm 0.0124$ | $0.5788 \pm 0.0377$ | $\mathbf{-0.0814 \pm 0.0622}$ |

The results are very similar to the semi-synthetic dataset in Section 5.1. The Baseline Model relies only on tabular data only able to slightly reduce the confounding bias (as the price prediction is quite hard only based on the tabular features). Additionally adding unstructured data, either in form of images or text does improve predictive performance of the nuisance estimates and reduces the confounding bias. Comparing the effect of image and text data, text data seem to contain more relevant information about prices. Combining all modalities seems to slightly improve the estimates, but also increase uncertainty as standard errors become larger. The improvement might only be comparatively small as images and text could contain a lot of similar information. Surprisingly, the Embedding Model seems to perform better compared to directly using the predicted values, even with slightly lower $r^2$-scores. One possible explanation of this might be the comparatively small sample size of $N = 9,207$, which might benefit tree models such as boosting to better incorporate the information from the tabular features.

### B.3. Real World Dataset

We analyze the same dataset as in appendix B.2. In the real-world application, we consider the variable $\log(\text{price})$ as treatment and the negative logarithm of the sales rank $-\log(\text{sales rank})$ as outcome. Consider a partially linear regression model (from equation 1 and 2)

$$
\begin{aligned}
-\log(\text{sales rank}) &= \log(\text{price}) \cdot \theta_0 + g_0(X) + \varepsilon, & \mathbb{E}[\varepsilon | X, \log(\text{price})] &= 0 \\
\log(\text{price}) &= m_0(X) + \vartheta, & \mathbb{E}[\vartheta | X] &= 0.
\end{aligned}
$$

When interpreting the negative sales rank as a proxy for the quantity of the good demanded, the causal parameter can be understood as the price elasticity of demand (remark that all additional controls refer to pre-treatment features).

Table 8: Results for the semi-synthetic dataset with dependent modalities. Reported: mean ± sd. over five random train-test splits.

| Modalites | $R^2(-\log(\text{sales rank}), \hat{l})$ | $R^2(\log(\text{price}), \hat{m})$ | $\hat{\theta}$ | |
|---|---|---|---|---|
| tab | $0.588 \pm 0.0108$ | $0.2282 \pm 0.0239$ | $-0.137 \pm 0.0277$ | |
| | | | $\hat{\theta}^{Deep}$ | $\hat{\theta}^{Embedding}$ |
| tab & img | $0.4518 \pm 0.0292$ | $0.5085 \pm 0.0288$ | $-0.2721 \pm 0.0378$ | $-0.2165 \pm 0.0358$ |
| tab & txt | $0.5574 \pm 0.0255$ | $0.6189 \pm 0.0149$ | $-0.2137 \pm 0.0535$ | $-0.1984 \pm 0.035$ |
| tab & txt & img | $0.5829 \pm 0.015$ | $0.6532 \pm 0.0184$ | $-0.2794 \pm 0.0436$ | $-0.2669 \pm 0.0623$ |

Mean values of the Deep Model over five random train-test splits. Reported as mean $\pm$ standard deviation. The $R^2$ values refer to the Deep Model.

The results in Table 8 show a large predictive performance increase for the treatment ($\log(\text{price})$), but not for the outcome ($-\log(\text{sales rank})$). Generally, it seems that tabular features, such as e.g. Review Count are very good predictors for the sales rank as predictive performance is already quite good in the baseline model. Using a neural network architecture instead of tree-based models makes it harder to achieve similar predictive performance on the sales rank and is only achieved in combination with text data. In contrast to the sales rank the price predictions strongly benefit from the incorporation of further modalities (all $R^2(\log(\text{price}), \hat{m})$ are compareable to the relative $r^2(D, \hat{m})$-scores from Section B.2). Overall the estimates of price elasticity are consistently larger in a absolute value and point to a much higher elasticity than baseline estimates.