Text Rationalization for Robust Causal Effect Estimation

Anonymous authors
Paper under double-blind review

Abstract

Recent advances in natural language processing have enabled the increasing use of text data in causal inference, particularly for adjusting confounding factors in treatment effect estimation. Although high-dimensional text can encode rich contextual information, it also poses unique challenges for causal identification and estimation. In particular, the positivity assumption, which requires sufficient treatment overlap across confounder values, is often violated at the observational level, when massive text is represented in feature spaces. Redundant or spurious textual features inflate dimensionality, producing extreme propensity scores, unstable weights, and inflated variance in effect estimates. We address these challenges with Confounding-Aware Token Rationalization (CATR), a framework that selects a sparse necessary subset of tokens using a residual-independence diagnostic designed to preserve confounding information for unconfoundedness. By discarding irrelevant texts while retaining key signals, CATR mitigates observational-level positivity violations and stabilizes downstream causal effect estimators. Experiments on semi-synthetic data and a real-world study using the MIMIC-III database demonstrate that CATR yields more accurate, stable, and interpretable causal effect estimates than existing baselines.

1 Introduction

Causal inference plays a central role across social, biomedical, and computational sciences (Rubin, 1974; Holland, 1986; Pearl, 2000). A primary objective in causal inference is to estimate treatment effects, such as the average treatment effect (ATE) (Imbens & Rubin, 2015). Identification of these estimands requires the assumptions of consistency, unconfoundedness, and positivity (Pearl, 2000). To satisfy *unconfoundedness*, it is common practice to include as much relevant information as possible. Beyond structured covariates, unstructured text has emerged as an additional source of information for confounding adjustment. Text often encodes rich contextual information that simultaneously influences treatment assignment and outcome, making it a promising source for confounding adjustment by surfacing otherwise unrecorded confounders (Egami et al., 2022; Feder et al., 2022; Wei et al., 2022; Zhao et al., 2023). A motivating example arises in critical care, where randomized trials are seldom feasible and analyses rely on observational data. The Medical Information Mart for Intensive Care III database (MIMIC-III, Johnson et al., 2016) combines structured electronic health record variables (e.g., demographics, vital signs, laboratory measurements) with unstructured clinical notes. As illustrated in Figure 1, these notes capture rich contextual information about comorbidities and patient adherence, which are not encoded in structured electronic health records yet may confound both treatment assignment and outcomes.

Despite this promise, incorporating text poses unique challenges for the *positivity* assumption, which requires the probabilities of treatment assignments (also known as the propensity score) to remain bounded away from 0 and 1 for all confounder combinations. This assumption is particularly vulnerable in text settings because (i) text often contains *irrelevant or spurious features* that obscure the true confounding patterns; and (ii) the *high-dimensional, sparse, and discrete* nature of text substantially enlarges the covariate space, reducing overlap between treatment groups. As a result, even when positivity holds in the underlying data-generating process, overlap may deteriorate in the observed text. We refer to this mismatch as an *observational-level positivity violation* (illustrated in Figure 2): empirical overlap fails not because positivity is violated at the population level, but owing to high-dimensional and redundant textual proxies that expand the observed

Review of Systems

NEURO: Pt is confused and inappropriate at times. He is capable of following commands but generally chooses not to. He becomes belligerent and agitated with any interventions, despite explanations.

RESP: Sats have been 92–94% on NRB; he refused to keep it on, and was changed over to 6L NC with sats remaining in low 90's. LS very diminished, especially LLL.

C-V: HR 80's, NSR, no ectopy. EKG in EW with ? new ST depressions in V3–V6. Denies CP. CK #1 4006 with negative MB fraction and slightly elevated Troponin. CK #2 pending. BP has been borderline, though pt apparently runs low. Currently, however, BP is in low 80's, and he is receiving a fluid bolus.

GI: Belly soft, NT, ND, active BS. No stool.

ID: Low-grade temp; cx's in EW. ? pnx on CXR. Got Ceftriaxone and Levoquin in EW.

HEME: PT/PTT extremely elevated this AM, but were drawn from Quinton cath (likely contaminated with heparin despite removing it from line).

RENAL: Last dialyzed [**5-1**]. Pt is anuric at baseline.

Social

Pt's spokesperson states he lives with his grandson, and has 3 dtrs.

Information captured by structured variable
Patient behavior and response to interventions
Diagnostic uncertainty or preliminary clinical impressions
Social and family context

Figure 1: Example clinical note with color-coded highlights for information often missing from structured variables. A complete list of structured variables is provided in Table D.1 in Appendix.

covariate space. This leads to extreme estimated propensity scores near 0 or 1, producing unstable inverse probability weighted (IPW) estimates and inflated variance in estimators, as observed in prior work (Belloni et al., 2014; Keller et al., 2015; Feder et al., 2022). Theoretically, Melnychuk et al. (2024) establish bounds on representation-induced confounding bias, showing the risks of unconstrained text encoding for causal estimation. These challenges motivate our focus in this paper:

How can we extract a statistically sufficient, low-dimensional representation of text that preserves unconfoundedness while mitigating observational-level positivity violations?

To address this problem, we propose selecting sparse token subsets that preserve confounding information while reducing redundant textual proxies that contribute to observational-level positivity violations. Our approach is based on a residual dependence criterion derived from the Hilbert–Schmidt Independence Criterion (HSIC), a kernel-based measure widely used in independence testing and causal discovery (Gretton et al., 2007; Zhang et al., 2011; Strobl et al., 2019). The key insight is that if the selected tokens preserve unconfoundedness, then the residuals of the treatment and outcome models should be independent. Residual HSIC therefore provides a *falsification test* that rules out insufficient token subsets and retains essential confounding information to improve overlap. Our main **contributions** are fourfold:

• *Conceptually*, to our knowledge, this is the first work on causal effect estimation with high-dimensional text that emphasizes interpretable text rationalization for robustness. We formalize how high-dimensional text confounding can induce *observational-level positivity violations* and highlight the need for selecting token subsets that are necessary and sufficient for confounder adjustment.

• *Methodologically*, we introduce *Confounding-Aware Token Rationalization (CATR)*, a token-level rationalization framework for causal inference with text data. CATR selects sparse token subsets by minimizing a residual HSIC criterion under a sparsity penalty, preserving the conditional independence structure required for unconfoundedness. The framework provides a principled and computationally efficient approach for mitigating observational-level positivity violations and stabilizing downstream causal estimation.

• *Theoretically*, we establish conditions under which residual HSIC provides a valid falsification test for sufficiency of the selected token subset. We derive nonasymptotic error bounds for the neural nuisance estimators in CATR and show that the resulting augmented IPW estimator achieves $\sqrt{n}$-consistency and asymptotic normality, establishing semiparametric efficiency guarantees for ATE estimation.

• *Empirically*, through semi-synthetic experiments and a real-world analysis of the MIMIC-III sepsis cohort, we demonstrate that CATR consistently improves overlap, reduces weight instability and variance, and yields more stable and interpretable ATE estimates compared to existing text-based causal methods, highlighting the practical importance of observational-level positivity for reliable causal inference with text.

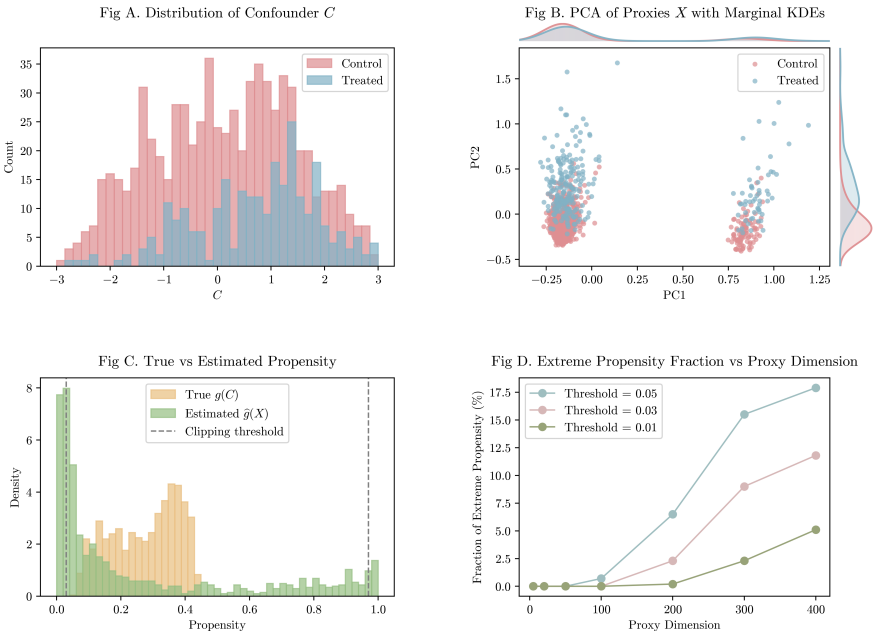

Figure 2: Illustration of observational-level positivity violations induced by high-dimensional proxy covariates. Panel A shows overlap between treated and control units in the latent confounder space. Panel B visualizes the corresponding proxy space using the first two principal components and marginal KDEs. Panel C compares the ground-truth and estimated propensity score distributions, with dashed lines indicating clipping thresholds. Panel D shows that the clipping fraction increases with proxy dimension.

1.1 Related Works

Text-based representation learning aims to construct low-dimensional representations of text that preserve the original causal information. Early work reduced dimensionality through topic models (Roberts et al., 2020; Egami et al., 2022) or bag-of-words (Wood-Doughty et al., 2018). These methods provide coarse summaries that struggle to capture nuanced linguistic features. The advent of transformer-based language models, such as BERT (Devlin et al., 2019), marked a shift toward context-aware text embeddings, which have become widely used in biomedical prediction tasks (Alsentzer et al., 2019; Huang et al., 2019; Lee et al., 2020). These representations have been paired with architectures originally developed for high-dimensional causal inference, including TARNet/CFRNet (Shalit et al., 2017), and DragonNet (Shi et al., 2019). Subsequent work adapts these models more explicitly to the structure of text. One line focuses on representation learning: CausalBERT (Veitch et al., 2020) augments language-modeling objectives while jointly predicting treatment and outcome; Gui & Veitch (2023) study apparent overlap violations when the treatment is itself an attribute of the text, whereas we use text as a high-dimensional confounder. A second line employs latent-variable models, including CEVAE (Louizos et al., 2017), TEDVAE (Zhang et al., 2021), and DIVA (Zhou & He, 2023), which infer latent confounders and decompose text into interpretable factors. While these methods reduce dimensionality, they do *not directly address the observational-level positivity violation* caused by redundant proxies in high-dimensional text. Moreover, latent factors can be *difficult to interpret and sensitive* to model assumptions.

Propensity score stabilization is commonly used to mitigate limited overlap and the resulting extreme inverse-propensity weights for structured covariates. One class of methods operates directly on the propensity score distribution, including trimming or clipping observations near 0 or 1 (Crump et al., 2009; Lee et al., 2011) and rescaling inverse probability weighting to control variance inflation (Cole & Hernán, 2008; Li et al., 2018; van der Laan et al., 2025). Another class, covariate balancing methods (Imai & Ratkovic, 2014), enforces balance conditions directly during estimation to reduce reliance on extreme propensities. However,

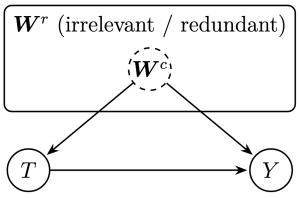

Figure 3: Causal diagram illustrating text-based confounding: $\boldsymbol{W}^c$ (confounding tokens) affects both treatment T and outcome Y, while $\boldsymbol{W}^r$ denotes irrelevant or redundant tokens.

these techniques are *not well suited to high-dimensional text-based confounding*. Post-hoc adjustments such as trimming or clipping discard valid samples without resolving token-level imbalance; traditional balancing methods lack explicit semantic dimensions over which covariate balance can be enforced in text settings.

Selective rationalization methods aim to extract concise token subsets that preserve predictive performance to address the high dimensionality of text (Bao et al., 2018; Jain et al., 2020; Guerreiro & Martins, 2021; Antognini & Faltings, 2021). Early approaches used non-differentiable selectors (Lei et al., 2016; Yu et al., 2019), whereas later work introduced differentiable relaxations such as the Gumbel–Softmax for end-to-end training (Bastings et al., 2019; Geng et al., 2020; Sha et al., 2021; Liu et al., 2022). Recent efforts (Chang et al., 2020; Cai et al., 2023; Zhang et al., 2023) have connected rationalization to causal notions of necessity and sufficiency, though predominantly in *predictive* rather than inference settings and do *not evaluate whether selected tokens preserve valid causal identification*.

2 Problem Setup

2.1 Notations

Let $(\boldsymbol{W}, T, Y)$ denote random variables for the text, treatment, and outcome, respectively. We observe n independent and identically distributed samples $\{(\boldsymbol{w}_i, t_i, y_i)\}_{i=1}^n$, where $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iL_i})$ is the observed text for unit i of length L_i, $t_i \in \{0, 1\}$ is the binary treatment indicator, and $y_i \in \{0, 1\}$ (or $\mathbb{R}$) is the observed outcome. Let $\Lambda(\boldsymbol{W})$ denote the embedding of $\boldsymbol{W}$ produced by a pre-trained encoder $\Lambda(\bullet)$. Following the potential outcome framework (Rubin, 1974), each unit has two potential outcomes, $Y(0)$ and $Y(1)$, under control and treatment, respectively. As discussed in Section 1, although the full text $\boldsymbol{W}$ contains rich contextual information, only a subset $\boldsymbol{W}^c \subset \boldsymbol{W}$ encodes the true confounding signals that jointly affect both treatment T and outcome Y. The remaining tokens, denoted as $\boldsymbol{W}^r$, are irrelevant or redundant. Throughout, we adopt the causal structure illustrated in Figure 3, in which the causal pathways operate through $\boldsymbol{W}^c$ but not through $\boldsymbol{W}^r$.

We define the true propensity score and mean outcome functions as $g(\boldsymbol{w}^c) = \Pr(T = 1 \mid \boldsymbol{W}^c = \boldsymbol{w}^c)$ and $Q_t(\boldsymbol{w}^c) = \mathbb{E}[Y \mid T = t, \boldsymbol{W}^c = \boldsymbol{w}^c]$, respectively, given the true confounding signals. Let $\widehat{g}(\cdot)$ and $\widehat{Q}_t(\cdot)$ denote their corresponding estimators, which may be constructed using either the full text $\boldsymbol{W}$ or a selected subset. We are interested in the estimation of the average treatment effect (ATE), defined as $\tau = \mathbb{E}\{Y(1) - Y(0)\}$. Its identification from observational data requires the following assumptions (Imbens & Rubin, 2015; Farrell et al., 2021):

(A1). Unconfoundedness. The true confounding signals $\boldsymbol{W}^c$ are sufficient for adjustment: $\{Y(0), Y(1)\} \perp T \mid \boldsymbol{W}^c$.

(A2). Positivity on True Signals. Each unit has a nonzero probability of receiving both treatment and control conditioned on $\boldsymbol{W}^c$: $0 < p_1 \leq \Pr(T = 1 \mid \boldsymbol{W}^c) \leq p_2 < 1, \quad \forall \boldsymbol{W}^c$.

(A3). Consistency. The observed outcome corresponds to the potential outcome under the received treatment: $Y = Y(T)$.

Under Assumptions (A1)– (A3), the average treatment effect τ is identifiable. Given an estimated propensity score $\widehat{g}$ and estimated conditional outcomes $\{\widehat{Q}_t\}_{t \in \{0,1\}}$, we consider common estimators for τ, including

outcome regression (OR, Pearl, 2000), inverse propensity weighting (IPW, Rosenbaum & Rubin, 1983), and augmented inverse propensity weighting (AIPW, Bang & Robins, 2005):

$$\widehat{\tau}_{OR} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{Q}_{1,i} - \widehat{Q}_{0,i}\right], \tag{1}$$

$$\widehat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{t_i y_i}{\widehat{g}_i} - \frac{(1-t_i)y_i}{1-\widehat{g}_i}\right], \tag{2}$$

$$\widehat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{t_i(y_i - \widehat{Q}_{1,i})}{\widehat{g}_i} + \widehat{Q}_{1,i} - \frac{(1-t_i)(y_i - \widehat{Q}_{0,i})}{1-\widehat{g}_i} - \widehat{Q}_{0,i}\right], \tag{3}$$

where $\widehat{g}_i$ and $\widehat{Q}_{t,i}$ denote the estimated propensity score and conditional outcome for unit i. While Assumptions (A1)–(A3) are typically imposed when working with structured covariates, their extension to text data introduces additional complexities (Feder et al., 2022). The unconfoundedness assumption is plausibly satisfied when rich text is incorporated, yet the positivity assumption becomes particularly fragile.

2.2 Observational-level Positivity Violation

Since the true confounders $\boldsymbol{W}^c$ are unknown, analysts often condition on the full text $\boldsymbol{W}$, which leads to high-dimensional covariates, diminished empirical overlap, and unstable propensity score estimates. We illustrate this effect through a controlled simulation. Treatment assignment depends on a single latent confounder C, so positivity holds by construction. The observed covariates, however, are high-dimensional sparse proxy features X. Details of the experimental setup are provided in the supplementary material. As shown in Figure 2, treated and control units overlap well in the latent C-space, but their separation becomes more pronounced after mapping to the proxy X-space. This distortion leads the estimated propensities $\widehat{g}(X)$ to place more mass near 0 and 1 than the ground-truth propensities, often crossing clipping thresholds at 0.03 and 0.97. The resulting inverse probability weights become heavy-tailed, inflating the variance of IPW-based causal effect estimators. As the proxy dimension increases, the clipping fraction also rises, mirroring the empirical behavior of high-dimensional text confounders, where irrelevant tokens can degrade overlap without improving confounding adjustment.

We refer to this phenomenon as an *observational-level positivity violation*: although the true treatment mechanism $g(\boldsymbol{W}^c)$ is well balanced, the estimated propensity scores $\widehat{g}(\boldsymbol{W})$ can become unstable when redundant or irrelevant tokens distort the empirical covariate space. In this case, effective overlap deteriorates even though Assumption (A2) holds at the latent confounder level. Formally, we define:

Definition 2.1 (Observational-level positivity violation)**.** Let $g(\boldsymbol{w}) = \Pr(T = 1 \mid \boldsymbol{W} = \boldsymbol{w})$ denote the population-level treatment mechanism, and $\widehat{g}(\boldsymbol{w}) = \Pr_n(T = 1 \mid \boldsymbol{W} = \boldsymbol{w})$ its empirical estimate. An *observational-level positivity violation* occurs when there exists $\boldsymbol{w} \in \text{supp}(\boldsymbol{W})$ such that $\widehat{g}(\boldsymbol{w}) \notin [\epsilon, 1-\epsilon]$ for some $0 < \epsilon < 1$, even though the true treatment assignment mechanism satisfies $g(\boldsymbol{w}^c) \in [\epsilon, 1-\epsilon]$ given the true confounding subset $\boldsymbol{w}^c$. Equivalently, at the empirical level, for some $\delta > 0$, we have

$$\mathbb{P}_n\{\widehat{g}(\boldsymbol{W}) \notin [\epsilon, 1-\epsilon]\} \geq \delta.$$

Importantly, this notion differs from the positivity violation at underlying data-generating process, which leads to non-identifiable causal effects. Observational-level violations, instead, preserve identifiability but complicate estimation by inflating variance, degrading overlap, and producing unstable inverse-propensity weights.

3 Methodology

In this section, we formally introduce Confounding-Aware Token Rationalization (CATR), a framework that integrates HSIC-based residual independence into token-level rationalization, as illustrated in Figure 4 and summarized in Algorithm 1.

3.1 Text Selection

To mitigate the observational-level positivity violation in Definition 2.1, our goal is to reduce the irrelevant or redundant information by learning a sparse text selector

$$S(\cdot) : \boldsymbol{w}_i \longrightarrow \tilde{\boldsymbol{w}}_i = S(\boldsymbol{w}_i),$$

which maps each observed text sequence $\boldsymbol{w}_i$ to a subsequence $\tilde{\boldsymbol{w}}_i$ of length $d_i \ll L_i$, such that $\tilde{\boldsymbol{w}}_i$ approximates the true confounding component $\boldsymbol{w}_i^c$. To ensure interpretability and statistical stability, the selector is subject to a sparsity constraint, such as a token budget or an entropy-based regularization. We require that $\tilde{\boldsymbol{W}}$ inherits the identification assumptions that hold for $\boldsymbol{W}^c$, namely:

$$\underbrace{(Y(0), Y(1)) \perp T \mid \tilde{\boldsymbol{W}}}_{\text{Unconfoundedness}}, \qquad \underbrace{\epsilon \leq g(\tilde{\boldsymbol{W}}) \leq 1 - \epsilon}_{\text{Positivity}}, \qquad \underbrace{Y = Y(T)}_{\text{Consistency}}.$$

3.2 HSIC Measurement for Subset Sufficiency

The sparse text selector faces a fundamental tension: removing irrelevant tokens improves empirical overlap, yet aggressive filtering risks discarding confounding information and violating unconfoundedness. To navigate this trade-off, we introduce a penalty based on the Hilbert–Schmidt Independence Criterion (HSIC, Gretton et al., 2007), which measures statistical dependence. The key idea is that, conditioning on a sufficient selected text subset, the residuals of the treatment and outcome models should be independent.

Specifically, given two random variables U and V, let k_U and k_V be the characteristic kernels with corresponding Gram matrices $K, L \in \mathbb{R}^{n \times n}$ over n samples. Define $H = I_n - (1/n)\mathbf{1}_n\mathbf{1}_n^\top$, where I_n is the $n \times n$ identity matrix and $\mathbf{1}_n$ is an n-dimensional vector of ones. Then the empirical HSIC is given by

$$\widehat{\mathrm{HSIC}}(U, V) = \tfrac{1}{(n-1)^2}\mathrm{tr}(KHLH),$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace. HSIC is nonnegative and equals zero if and only if $U \perp V$. To establish when zero residual HSIC certifies subset sufficiency, we adopt a standard noise structure widely used in causal discovery and nonparametric identification (Hoyer et al., 2009; Peters et al., 2014; Strobl et al., 2019). Formally, we assume:

(A4). Additive independent noise. There exist measurable functions $g^\star : \mathcal{W}^c \to [0, 1]$ and $Q_T^\star : \{0, 1\} \times \mathcal{W}^c \to \mathbb{R}$ such that $T = g^\star(\boldsymbol{W}^c) + \varepsilon_T,\ Y = Q_T^\star(\boldsymbol{W}^c) + \varepsilon_Y$, where ε_T and ε_Y are exogenous noise terms satisfying: (i) $\varepsilon_T \perp \boldsymbol{W}$ and $\varepsilon_Y \perp (T, \boldsymbol{W})$; (ii) $\mathbb{E}[\varepsilon_T] = \mathbb{E}[\varepsilon_Y] = 0$ and $\mathbb{E}[\varepsilon_T^2], \mathbb{E}[\varepsilon_Y^2] < \infty$; (iii) $\varepsilon_T \perp \varepsilon_Y$.

Under Assumption (A4), if the selected token subset $\tilde{\boldsymbol{W}}$ is sufficient for confounding adjustment, then the residual dependence between treatment and outcome vanishes, as detailed in the following proposition with proof provided in Appendix A.

Proposition 3.1. *Let $\tilde{\boldsymbol{W}} \subseteq \boldsymbol{W}$ be any candidate subset. Define $g(\tilde{\boldsymbol{W}}) = \mathbb{E}[T \mid \tilde{\boldsymbol{W}}]$ and $Q_t(\tilde{\boldsymbol{W}}) = \mathbb{E}[Y \mid T = t, \tilde{\boldsymbol{W}}]$, as well as residuals $r_T = T - g(\tilde{\boldsymbol{W}})$ and $r_Y = Y - Q_T(\tilde{\boldsymbol{W}})$. Let $D(\tilde{\boldsymbol{W}}) = \mathrm{HSIC}(r_T, r_Y)$ be the population-level HSIC computed with characteristic and bounded kernels under finite second moments. Suppose Assumptions (A1), (A3), and (A4) hold. Then,*

$$(Y(0), Y(1)) \perp T \mid \tilde{\boldsymbol{W}} \quad \Longrightarrow \quad D(\tilde{\boldsymbol{W}}) = 0.$$

Contrapositively, if the residual HSIC is positive, then $\tilde{\boldsymbol{W}}$ is not sufficient, i.e.,

$$D(\tilde{\boldsymbol{W}}) > 0 \quad \Longrightarrow \quad (Y(0), Y(1)) \not\perp T \mid \tilde{\boldsymbol{W}}.$$

Residual HSIC provides a falsification test: residual dependence indicates that the subset $\tilde{\boldsymbol{W}}$ fails to block all confounding paths. This method follows HSIC-based independence testing (Gretton et al., 2005; 2007) and related independence tests applied in feature selection and causal discovery (Song et al., 2012; Zhang et al., 2011; Strobl et al., 2019). Following the logic of falsification checks in observational studies (e.g., Pearl, 1995; Faller et al., 2024; Karlsson & Krijthe, 2025), the test is necessary but not sufficient for unconfoundedness. We incorporate it as a regularization term in the training objective (Equation 4) and encourage approximate residual independence during optimization, thereby balancing sufficiency and overlap in finite samples.

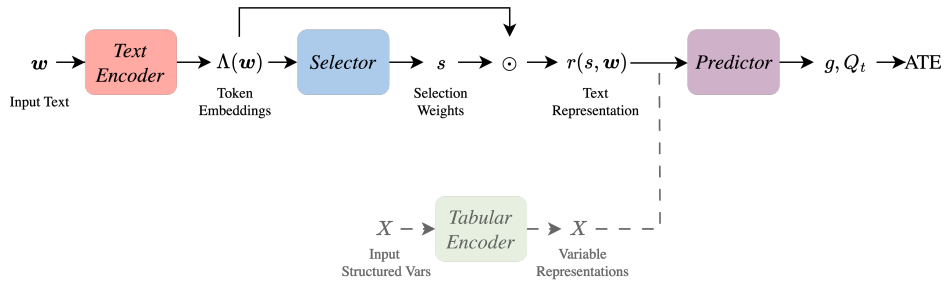

Figure 4: The Confounding-Aware Token Rationalization (CATR) framework (with multimodal extension). The main pipeline (solid blocks) implements selective rationalization for text-only inputs. The semi-transparent dashed components illustrate the extension to tabular covariates, showing that our approach naturally adapts to multi-modal settings.

3.3 Architecture and Training Objective

Building upon the standard rationalization paradigm (see e.g., Jain et al., 2020; Antognini & Faltings, 2021), our framework consists of a selector $S(\cdot)$ and predictors $g(\cdot)$, $\{Q_t(\cdot)\}_{t\in\{0,1\}}$. For each observed text $\boldsymbol{w}_i$, the selector $S_\theta(\cdot)$ produces a token-level mask $s_i = S_\theta(\boldsymbol{w}_i) \in [0,1]^{L_i}$ parameterized by θ. During training, s_i is drawn from a relaxed Bernoulli distribution to introduce stochasticity and encourage exploration of alternative subsets; during evaluation, we instead use deterministic selection. The resulting selected subsequence is formed as $\tilde{\boldsymbol{w}}_i = s_i \odot \boldsymbol{w}_i$, where $\odot$ denotes elementwise multiplication. Given the selected $\tilde{\boldsymbol{w}}_i$, the predictor network (parameterized by ϕ) produces the estimated propensity score $\widehat{g}(\tilde{\boldsymbol{w}}_i)$ and estimated conditional outcomes $\{\widehat{Q}_t(\tilde{\boldsymbol{w}}_i)\}_{t\in\{0,1\}}$. Conceptually, the architecture parallels DragonNet (Shi et al., 2019), except that the input representation $\tilde{\boldsymbol{w}}$ is itself learnable through the selector $S_\theta(\cdot)$ rather than fixed. This design enables token-level control of confounding adjustment and provides interpretable rationales of which linguistic features contribute most.

Given a training dataset $\mathcal{D} = \{(\boldsymbol{w}_i, t_i, y_i)\}_{i=1}^n$, we formalize this architecture through the following joint training objective:

$$\min_{\theta,\phi} \frac{1}{n}\sum_{i=1}^{n}\Big[\mathcal{L}_{\text{sup},i} + \mu\,\mathcal{R}_{\text{sparse}}(s_i) + \gamma\,\text{HSIC}(r_{T,i}, r_{Y,i})\Big], \tag{4}$$

where $s_i = S_\theta(\boldsymbol{w}_i)$ is the selection mask, $\mu, \gamma > 0$ control the trade-off among predictive accuracy, sparsity, and residual independence. The loss consists of three components:

(i) Supervised loss. $\mathcal{L}_{\text{sup},i}$ trains the propensity score and potential outcome predictors, ensuring predictive fidelity. Formally,

$$\mathcal{L}_{\text{sup},i} = \text{CE}\big(t_i, \widehat{g}_\phi(\tilde{\boldsymbol{w}}_i)\big) + \beta\,\text{CE}\big(y_i, \widehat{Q}_{t_i,\phi}(\tilde{\boldsymbol{w}}_i)\big),$$

where $\text{CE}(\cdot,\cdot)$ is the cross-entropy loss and $\beta > 0$ weights the supervised losses.

(ii) Sparsity regularization. $\mathcal{R}_{\text{sparse}}(s_i)$ is a general regularization term that encourages compact token subsets. Depending on the desired sparsity level and behavior, the regularizer can take different forms: a Kullback–Leibler divergence term with a Bernoulli(π) prior to enforce a token budget π; an entropy regularization to promote confident selections; or norm-based penalties for direct sparsity control. In all experiments, we use entropy regularization to avoid additional hyperparameter tuning.

(iii) HSIC penalty. $\text{HSIC}(r_{T,i}, r_{Y,i})$ measures residual dependence between treatment and outcome. In practice, this term is evaluated as a batch-level statistic $\widehat{\text{HSIC}}(\{r_{T,i}\}_{i=1}^m, \{r_{Y,i}\}_{i=1}^m)$ computed over the current minibatch; the per-sample notation in Equation 4 is a presentational simplification. As stated in Proposition 3.1, while minimizing empirical HSIC does not guarantee perfect unconfoundedness, it guides the learned representation to approximate conditional independence. We compute HSIC using characteristic kernels: a linear kernel $k_T(u,v) = u^\top v$ for the binary treatment residual r_T, and a Gaussian RBF kernel $k_Y(u,v) = \exp(-\|u-v\|^2/2\sigma_Y^2)$ for the continuous outcome residual r_Y. The bandwidth σ_Y is selected using

Algorithm 1 Confounding-Aware Token Rationalization (CATR)

Input: dataset $\mathcal{D} = \{(\boldsymbol{w}_i, t_i, y_i)\}_{i=1}^n$; temperature η; prior budget π; loss weights (μ, γ)
Split $\mathcal{D}$ into training and validation sets
Initialize selector $S(\cdot)$, predictor $g(\cdot)$, and $\{Q_t(\cdot)\}_{t\in\{0,1\}}$
while not converged **do**
  Sample minibatch $\{(\boldsymbol{w}_i, t_i, y_i)\}_{i=1}^m \sim \mathcal{D}_{\text{train}}$
  Compute selection score $a_i = S_\theta(\boldsymbol{w}_i)$
  Sample relaxed gates $s_{ij} \sim \text{RelaxedBernoulli}(\eta, a_i)$ to form a differentiable binary mask
  Apply mask to obtain subsequence $\tilde{\boldsymbol{w}}_i = s_i \odot \boldsymbol{w}_i$
  Predictor outputs estimated propensity score $\widehat{g}(\tilde{\boldsymbol{w}}_i)$ and conditional mean outcomes $\{\widehat{Q}_t(\tilde{\boldsymbol{w}}_i)\}_{t\in\{0,1\}}$
  Form residuals $r_{T,i} = t_i - \widehat{g}_i$, $r_{Y,i} = y_i - \widehat{Q}_{t_i,i}$
  Compute loss in Equation 4
  Update model parameters via stochastic gradient descent
end while
Output: trained selector $S(\cdot)$, predictors $g(\cdot)$ and $\{Q_t(\cdot)\}_{t\in\{0,1\}}$

the median heuristic. This choice follows common practice in independence testing and balances bias and variance in the HSIC estimate.

In summary, the objective balances accuracy, sparsity, and sufficiency for valid causal adjustment: the supervised loss ensures predictive fidelity, the sparsity regularizer enforces concise token subsets under user-defined constraints, and the HSIC term promotes residual independence, thereby stabilizing causal effect estimation. At inference time, the trained $\widehat{g}(\tilde{\boldsymbol{w}}_i)$ and $\{\widehat{Q}_t(\tilde{\boldsymbol{w}}_i)\}$ are substituted into Equations (1)–(3) as plug-in estimates, yielding $\widehat{\tau}_{\text{OR}}$, $\widehat{\tau}_{\text{IPW}}$, and $\widehat{\tau}_{\text{AIPW}}$. The hyperparameter selection and implementation details are provided in Section C.

3.4 Extension: Multimodal Data

While our main method is developed for text-only settings, many applications also include well-curated structured covariates. In such cases, text serves as a complementary source of confounding information. We extend our framework to jointly handle textual and structured data by encoding tabular covariates into dense vectors of the same dimension as text embeddings, and then fusing them into a single vector, following the strategy of Ebrahimi et al. (2023). As shown in Figure 4, in this end-to-end differentiable architecture, the token selection mechanism continues to operate on the text encoder output, while the fused representation allows joint adjustment for both text and structured covariates without modifying the core architecture.

4 Theoretical Analysis

This section establishes the statistical properties of the proposed CATR framework. We derive the nonasymptotic error bounds for the neural estimators of the nuisance functions, and establish the $\sqrt{n}$-consistency and asymptotic normality of the resulting AIPW estimator. All proofs are provided in Appendix A.

4.1 Nonasymptotic Error Bound of Neural Network Approximation

We begin by introducing a set of regularity conditions required for our theoretical results.

(A5). Regularity of data and representation. The treatment $T \in \{0, 1\}$ and outcome $Y \in [-M, M]$ for some $M > 0$. The observed text $\boldsymbol{W}$ is encoded by a pretrained model into a bounded, continuous-valued feature vector $\mathbf{x} = \Lambda(\boldsymbol{W}) \in [-1, 1]^d$, whose distribution admits a bounded density on $[-1, 1]^d$, where d is the encoder output dimension.

(A6). Smoothness. For each target function $f \in \{g, Q_0, Q_1\}$, assume that f lies in a Sobolev ball $\mathcal{W}^{\beta,\infty}([-1, 1]^d)$ with smoothness $\beta > 0$, i.e.,

$$f(\mathbf{x}) \in \mathcal{W}^{\beta,\infty}([-1, 1]^d) := \Big\{ f : \max_{|\alpha|\le\beta} \operatorname{ess\,sup}_{\mathbf{x}\in[-1,1]^d} |D^\alpha f(\mathbf{x})| \le 1 \Big\},$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$, $|\alpha| = \sum_j \alpha_j$, and $D^\alpha f$ denotes the weak derivative.

These assumptions are mild in language-based settings: pretrained encoders yield bounded, smoothly varying embeddings, where small semantic changes lead to gradual shifts in propensity or conditional outcome.

Under the above assumptions, we establish the nonasymptotic error bound for deep MLP-ReLU network approximation following the framework of Farrell et al. (2021).

Theorem 4.1. *Suppose Assumptions (A5)–(A6) hold. Let $\widehat{f}$ be the minimizer of the joint objective equation 4, with fixed penalty weights $\mu, \gamma = O(1)$ independent of n. Then, for an MLP-ReLU network class with width $H \asymp n^{\frac{d}{2(\beta+d)}} \log^2 n$ and depth $D \asymp \log n$, there exists $C > 0$ such that, with probability at least $1 - \exp(-n^{\frac{d}{\beta+d}} \log^8 n)$,*

$$\|\widehat{f} - f_*\|^2_{L_2(P_X)} \leq C\Big(n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log\log n}{n}\Big),$$

where f represents either g or Q_t, f_* denotes the true target function.

Theorem 4.1 extends the standard consistency result in Farrell et al. (2021) to the case of jointly estimated models with additional regularization. Since both $\mathcal{R}_{\text{sparse}}$ and HSIC are bounded and weighted by fixed constants, the penalized objective can be viewed as a constrained version of the unpenalized empirical risk minimization; therefore, the convergence rate $n^{-\beta/(\beta+d)}$ remains unchanged. Intuitively, the HSIC term promotes independence between treatment and outcome representations, and the sparsity term encourages concise rationales, but neither alters the statistical order of estimation error. While d corresponds to the embedding size, selective rationalization reduces the effective input complexity by removing spurious or irrelevant components and improves convergence in practice.

4.2 Statistical Inference

We first show the consistency and convergence of nuisance estimators.

Theorem 4.2. *Suppose Assumptions (A1)–(A6) hold, and the nonasymptotic bound in Theorem 4.1 holds, then the estimators $\widehat{g}$ and $\widehat{Q}_t$ satisfy, for $t \in \{0, 1\}$:*

(i) $\mathbb{E}_n[(\widehat{g}(\boldsymbol{w}_i) - g(\boldsymbol{w}_i))^2] = o_P(1), \quad \mathbb{E}_n[(\widehat{Q}_t(\boldsymbol{w}_i) - Q_t(\boldsymbol{w}_i))^2] = o_P(1)$;
(ii) $\mathbb{E}_n[(\widehat{Q}_t(\boldsymbol{w}_i) - Q_t(\boldsymbol{w}_i))^2]\mathbb{E}_n[(\widehat{g}(\boldsymbol{w}_i) - g(\boldsymbol{w}_i))^2] = o_P(n^{-1})$;
(iii) $\mathbb{E}_n[(\widehat{Q}_t(\boldsymbol{w}_i) - Q_t(\boldsymbol{w}_i))(t_i - g(\boldsymbol{w}_i))] = o_P(n^{-1/2})$.

Let $\widehat{\psi}_t$ denote the plug-in orthogonal score, defined for each treatment as

$$\widehat{\psi}_1 = \frac{t_i}{\widehat{g}(\boldsymbol{w}_i)}\{y_i - \widehat{Q}_1(\boldsymbol{w}_i)\} + \widehat{Q}_1(\boldsymbol{w}_i), \qquad \widehat{\psi}_0 = \frac{1 - t_i}{1 - \widehat{g}(\boldsymbol{w}_i)}\{y_i - \widehat{Q}_0(\boldsymbol{w}_i)\} + \widehat{Q}_0(\boldsymbol{w}_i).$$

If $\widehat{g}(\boldsymbol{w}_i)$ is bounded away from 0 and 1, then the empirical mean and second moment of the estimated score converge to their population counterparts:

$$\sqrt{n}\,\mathbb{E}_n[\widehat{\psi}_t - \psi_t] = o_P(1), \quad \frac{\mathbb{E}_n[(\widehat{\psi}_t)^2]}{\mathbb{E}_n[(\psi_t)^2]} - 1 = o_P(1).$$

Theorem 4.2 emphasizes the importance of observational-level positivity for the validity of nuisance function estimation. When the observational-level positivity is violated, the inverse weighting terms $1/\widehat{g}(\boldsymbol{w})$ and $1/\{1 - \widehat{g}(\boldsymbol{w})\}$ may diverge, causing the orthogonal scores to have unbounded variance. Positivity ensures that both the empirical process and the remainder terms in Theorem 4.2 remain well behaved, thereby guaranteeing finite-variance estimation of treatment effects.

Building on these properties, we analyze the convergence and asymptotic normality of the ATE estimators.

Theorem 4.3. *Suppose Assumptions (A1)–(A6) hold. Assume that the convergence rates in Theorem 4.2 hold. Define $\widehat{\tau}_{\text{IPW}}$ and $\widehat{\tau}_{\text{AIPW}}$ as in Equations (2) and (3). Let τ denote the ground truth ATE.*

(i) (Consistency) $\widehat{\tau}_{\text{IPW}}, \widehat{\tau}_{\text{AIPW}} \xrightarrow{P} \tau$.

(ii) (AIPW asymptotic normality) *If we further assume the nuisance estimators $\widehat{g}$ and $\widehat{Q}_t$ satisfy the product-rate condition*

$$\|\widehat{g} - g\|_{L_2(P_{\boldsymbol{w}})} \cdot \|\widehat{Q}_t - Q_t\|_{L_2(P_{\boldsymbol{w}})} = o_P(n^{-1/2}),$$

then

$$\sqrt{n}\,(\widehat{\tau}_{\text{AIPW}} - \tau) \;\rightsquigarrow\; \mathcal{N}(0, V^*), \quad V^* = \text{Var}\big(\psi_1 - \psi_0\big).$$

Theorem 4.3 shows that while both IPW and AIPW estimators are consistent under overlap, only AIPW achieves $\sqrt{n}$-asymptotic normality under the weaker product-rate condition, which aligns with our experimental results in Section 5.

5 Empirical Experiments

In this section, we evaluate the proposed CATR using both controlled semi-synthetic studies and a real-world analysis of the MIMIC-III sepsis cohort. Our empirical objectives are twofold: (i) to assess whether CATR identifies token subsets that preserve essential confounding information; (ii) to examine its ability to mitigate observational-level positivity violations and thereby stabilize causal estimators.

5.1 Semi-Synthetic Experiment

Semi-Synthetic Data Generation. We conduct semi-synthetic experiments with known treatment effects, using the MIMIC-III database (Johnson et al., 2016). We focus on a subset of ICU patients with sepsis. For each patient i, we extract the earliest physician note within eight hours of admission as $\boldsymbol{w}_i$. To allow for benchmarking of causal estimators under a realistic but controlled setting, we simulate treatment assignments and outcomes directly from the clinical notes. We first construct a binary indicator based on the presence of some infection-related keywords (*shock/hypotension, infection/sepsis/pneumonia*). This indicator is transformed through nonlinear mappings to generate both the propensity score and the potential outcome functions:

$$t_i \sim \text{Bernoulli}(g(\boldsymbol{w}_i)), \qquad y_i = Q_{t_i}(\boldsymbol{w}_i) + \varepsilon_i.$$

This semi-synthetic design yields a known ground truth $\tau_{\text{true}} = \mathbb{E}[Y(1) - Y(0)]$. More details on semi-synthetic data construction are provided in Section B in Appendix.

Baselines. We compare our method against four baseline methods: TARNet, CFRNet (Shalit et al., 2017), DragonNet (Shi et al., 2019), and CausalBERT (Veitch et al., 2020). For all models, we use the officially released source code when available, or reimplement them using the optimal hyperparameter settings reported in their respective papers. To ensure comparability, all text-based models use the same encoder backbone, ClinicalBERT (Wang et al., 2023; Liu et al., 2025), to generate embeddings. Implementation uses the Adam optimizer with early stopping based on validation loss. Data are randomly split into training, validation, and test sets in a 25:6:6 ratio, which are used respectively for model fitting, early stopping, and final evaluation. See more details in Section C in Appendix.

Evaluation Protocol and Metrics. Each model produces estimates of the propensity score $g(\boldsymbol{w})$ and the conditional outcomes $Q_t(\boldsymbol{w})$, which are then used to compute the ATE estimates, as defined in Equations (1)-(3). Each model is evaluated over $R = 200$ independent replications, each producing three ATE estimates based on newly sampled treatment and outcome assignments. For each replication–estimator pair, we compute the point estimate $\widehat{\tau}_r$ and its bootstrap standard deviation $\widehat{\text{SE}}_{r,e}$ using $B = 1000$ resamples, and construct the 95% confidence interval $[\widehat{\tau}_r \pm 1.96\,\widehat{\text{SE}}_r]$. Detailed definitions of the evaluation metrics are provided in Appendix C. For each method, the ESS ratio and clipping fraction are reported once, averaged across all replications, while ATE metrics are reported separately for OR, IPW, and AIPW estimators.

Note that OR yields unrealistically small standard errors across all variants, indicating that the outcome model is overconfident and OR estimates are unreliable in this setting; we therefore focus on IPW and AIPW. As shown in Table 1, CATR substantially improves empirical overlap and estimation stability compared to all baselines. By reducing extreme propensity scores and increasing the effective sample size, it mitigates observational-level positivity violations. This leads to lower bias and variance for both IPW and AIPW

Table 1: Comparison of ATE estimation and positivity diagnostics across methods. Results are averaged over 200 replications with bootstrap size $B = 1000$. Confidence intervals use a 95% confidence level, and propensity scores are clipped at 0.03/0.97. TARNet and CFRNet do not explicitly estimate propensity scores, so only OR estimation is reported.

Method	Estimator	ATE Metrics				PS Quality	
		Mean(|Bias|)	SD($\widehat{\text{ATE}}$)	Avg SE	C.I. Coverage	ESS Ratio	Clip. Frac.
TARNet	OR	0.145	0.178	0.022	21.0%	–	–
CFRNet	OR	0.143	0.173	0.022	20.5%	–	–
DragonNet	OR	0.160	0.199	0.020	16.5%	0.272	28.89%
	IPW	1.188	1.044	0.617	35.5%		
	AIPW	0.320	0.438	0.329	85.0%		
CausalBERT	OR	0.194	0.243	0.023	13.5%	0.266	30.96%
	IPW	1.393	1.104	0.649	48.5%		
	AIPW	0.351	0.475	0.349	87.5%		
CATR	OR	0.186	0.231	0.016	10.0%	0.310	10.74%
	IPW	**0.704**	**0.845**	**0.494**	**76.0%**		
	AIPW	**0.261**	**0.317**	**0.265**	**90.0%**		

Table 2: CATR ablation study. Results are averaged over 200 replications with bootstrap size $B = 1000$. Confidence intervals use a 95% confidence level, and propensity scores are clipped at 0.03/0.97.

Method	Estimator	ATE Metrics				PS Quality	
		Mean(|Bias|)	SD($\widehat{\text{ATE}}$)	Avg SE	C.I. Coverage	ESS Ratio	Clip. Frac.
Full CATR	OR	0.186	0.231	0.016	10.0%	0.310	10.74%
	IPW	0.704	0.845	0.494	76.0%		
	AIPW	0.261	0.317	0.265	90.0%		
−HSIC	OR	0.150	0.186	0.015	10.5%	0.290	17.36%
	IPW	0.978	0.947	0.570	67.5%		
	AIPW	0.278	0.335	0.304	90.0%		
−HSIC −Sparsity	OR	0.161	0.199	0.020	17.0%	0.270	29.55%
	IPW	1.241	1.083	0.623	51.0%		
	AIPW	0.321	0.438	0.333	89.5%		

estimators, as well as improved confidence interval coverage, particularly for IPW, which relies entirely on propensity score estimation. In contrast, DragonNet and CausalBERT produce unstable inverse weights and wider confidence intervals due to redundant token dependencies. Overall, these results suggest that HSIC-guided sparse rationalization improves statistical efficiency while preserving essential confounding signals.

Table 2 further demonstrates the effectiveness of CATR. For IPW estimation, removing the HSIC term substantially increases bias, variance, and clipping frequency, highlighting the importance of the residual-independence constraint for stable propensity estimation. Removing both HSIC and sparsity leads to even larger degradation, with severe IPW instability and poor confidence interval behavior, illustrating the role of CATR in mitigating positivity violations. For AIPW estimation, CATR achieves higher coverage rates while simultaneously reducing both standard deviation and standard error, indicating more stable inference and improved uncertainty quantification. Overall, these results show that CATR consistently improves estimation stability for propensity-weighted estimators.

Figure E.2 in Appendix compares the top-20 important words selected with and without HSIC regularization. When a word is tokenized into multiple subword units, its importance score is computed by averaging the corresponding token-level scores. Without HSIC regularization, most selected tokens appear only once and are largely unrelated to the pre-specified confounding words, suggesting that the model captures spurious correlations rather than meaningful causal signals. In contrast, incorporating HSIC successfully recovers all manually defined confounding words (*sepsis, infection, pneumonia, shock, hypotension*) together with related

Table 3: Comparison of ATE estimation and positivity diagnostics on real MIMIC-III datasets. Reported ATE estimates are computed after propensity score clipping at 0.05, and standard deviations are based on 1,000 bootstrap replicates. Clipping fractions are computed using thresholds of 0.03/0.97.

	ATE (Std)			PS Quality	
	OR	IPW	AIPW	ESS Ratio	Clip. Frac.
Structured-only	−0.069 (0.009)	−0.001 (0.105)	−0.043 (0.094)	0.211	8.18%
Multi-modal (w/o CATR)	0.047 (0.005)	0.118 (0.192)	−0.121 (0.154)	0.277	20.44%
Multi-modal (w/ CATR)	−0.114 (0.002)	−0.049 (0.057)	−0.026 (0.041)	0.852	0.00%

variants (e.g., *shocked*, *pneumonic*). By penalizing residual dependence between T and Y, HSIC helps the model avoid coincidental co-occurrences and improves the interpretability of the learned representations.

5.2 Real Data Analysis

To complement the semi-synthetic experiments, we analyze the same cohort from MIMIC-III database as in the simulation, with naturally occurring treatment and outcomes. Treatment is defined as intravenous (IV) fluid therapy and outcome is defined as ICU readmission. For each patient, we extract both physician notes and structured covariates (42 variables including demographics, comorbidities, vitals, and labs, provided in Section D in Appendix). This setting evaluates our multimodal extension (Section 3.4), which integrates structured variables and unstructured text to model confounding.

We compare three adjustment strategies: structured-only, multimodal without CATR, and multimodal with CATR. The structured-only setting uses logistic regression for propensity and outcome modeling, while the multimodal setting follows the architecture in Section 3.4. As shown in Table 3, adding unfiltered text raises the clipping fraction, illustrating the observational-level positivity violations induced by raw text embeddings. The multimodal model with CATR eliminates this entirely and yields the most stable ATE estimates across OR, IPW, and AIPW, together with higher ESS and lower clipping fractions. In addition, the estimated effects become more consistent across estimators, suggesting improved overlap and estimation stability. The AIPW estimate is -2.6% with reduced variance, indicating no significant causal effect of IV treatment on ICU readmission among sepsis patients, consistent with recent studies highlighting heterogeneous subgroup effects and questioning uniform treatment benefits (Lung et al., 2023; Sivapalan et al., 2023).

Beyond quantitative results, Figure E.1 in Appendix shows that CATR selects clinically meaningful tokens, such as terms indicating infections, organ dysfunction, or other severity markers that influence both treatment decisions and outcomes. These include keywords like “sepsis”, “infection”, and “hypotension”, as well as descriptors of patient condition (e.g., “labored breathing”, “unsure”).

6 Conclusion

We propose a framework for causal effect estimation when text is used to adjust for confounding. Our approach explicitly targets *observational-level positivity violations*, which occur when high-dimensional textual confounders induce near-deterministic treatment assignment despite positivity holding at the latent level. To address this, we incorporate selective rationalization that identifies a compact yet sufficient textual subset for confounding adjustment, yielding representations with improved overlap. Empirically, this design reduces propensity score extremity and stabilizes IPW and AIPW estimation, highlighting the practical value of overlap-aware rationalization for causal inference from natural language. Future work could strengthen causal interpretability by integrating counterfactual reasoning to explicitly enforce necessity and sufficiency of selected tokens. In addition, the framework can be naturally combined with post-hoc adjustment methods, such as temperature scaling, beta or histogram calibration, and covariate balancing techniques, to further stabilize propensity estimates and reduce residual bias.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pp. 72–78, 2019.

Diego Antognini and Boi Faltings. Rationalization through concepts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 761–775, 2021.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. Deriving machine attention from human rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1903–1913, 2018.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2963–2977, 2019.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2):29–50, 2014.

Hengrui Cai, Yixin Wang, Michael Jordan, and Rui Song. On learning necessary and sufficient causal graphs. *Advances in Neural Information Processing Systems*, 36:42148–42160, 2023.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1448–1458. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/chang20c.html`.

Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Sayna Ebrahimi, Sercan Ö Arik, Yihe Dong, and Tomas Pfister. LANISTR: Multimodal learning from structured and unstructured data. *CoRR*, 2023.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. How to make causal inferences using texts. *Science Advances*, 8(42):2652, 2022.

Philipp M Faller, Leena C Vankadara, Atalanti A Mastakouri, Francesco Locatello, and Dominik Janzing. Self-compatibility: Evaluating causal discovery without ground truth. In *International Conference on Artificial Intelligence and Statistics*, pp. 4132–4140. PMLR, 2024.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021. ISSN 0012-9682.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.

Xinwei Geng, Longyue Wang, Xing Wang, Bing Qin, Ting Liu, and Zhaopeng Tu. How does selective mechanism improve self-attention networks? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2986–2995, 2020.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in neural information processing systems*, 20, 2007.

Nuno M Guerreiro and André FT Martins. SPECTRA: Sparse structured text rationalization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6534–6550, 2021.

Lin Gui and Victor Veitch. Causal estimation for text data with (apparent) overlap violations. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=Ha2MnQM9Ph`.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396): 945–960, 1986. ISSN 01621459, 1537274X. URL `http://www.jstor.org/stable/2289064`.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Neural Information Processing Systems (NeurIPS)*, 22:689–696, 2009.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263, 2014.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4459–4473, 2020.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Rickard Karlsson and JH Krijthe. Falsification of unconfoundedness by testing independence of causal mechanisms. In *Forty-second International Conference on Machine Learning*. PMLR, 2025.

Bryan Keller, Jee-Seon Kim, and Peter M Steiner. Neural networks for propensity score estimation: Simulation results and recommendations. In *Quantitative Psychology Research: The 79th Annual Meeting of the Psychometric Society, Madison, Wisconsin, 2014*, pp. 279–291. Springer, 2015.

Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Weight trimming and propensity score weighting. *PloS one*, 6(3):e18174, 2011.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117, 2016.

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400, 2018.

Wei Liu, Haozhao Wang, Jun Wang, Ruixuan Li, Chao Yue, and YuanKai Zhang. FR: Folded rationalization with a unified encoder. *Advances in Neural Information Processing Systems*, 35:6954–6966, 2022.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, et al. A generalist medical language model for disease diagnosis assistance. *Nature medicine*, 31(3):932–942, 2025.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

National Heart Lung, Blood Institute Prevention, and Early Treatment of Acute Lung Injury Clinical Trials Network. Early restrictive or liberal fluid management for sepsis-induced hypotension. *New England Journal of Medicine*, 388(6):499–510, 2023.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=d3xKPQVjSc`.

Judea Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 435–443, 1995.

Judea Pearl (ed.). *Causality: models, reasoning and inference, volume 29*. Springer, 2000.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Margaret E. Roberts, Brandon M Stewart, and Richard A. Nielsen. Adjusting for confounding with text matching. *American Journal of Political Science*, 64:887–903, 2020.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Lei Sha, Oana-Maria Camburu, and Thomas Lukasiewicz. Learning from the best: Rationalizing predictions by adversarial information calibration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35 (15):13771–13779, May 2021.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Praleene Sivapalan, Karen L Ellekjaer, Marie K Jessen, Tine S Meyhoff, Maria Cronhjort, Peter B Hjortrup, Jørn Wetterslev, Anders Granholm, Morten H Møller, and Anders Perner. Lower vs higher fluid volumes in adult patients with sepsis: an updated systematic review with meta-analysis and trial sequential analysis. *Chest*, 164(4):892–912, 2023.

Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. Feature selection via dependence maximization. *The Journal of Machine Learning Research*, 13(1):1393–1434, 2012.

Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1):20180017, 2019.

Lars van der Laan, Ziming Lin, Marco Carone, and Alex Luedtke. Stabilized inverse probability weighting via isotonic calibration. In *Causal Learning and Reasoning*, pp. 139–173. PMLR, 2025.

Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In Jonas Peters and David Sontag (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 919–928. PMLR, 2020.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29(10):2633–2642, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=yzkSU5zdwD`. Survey Certification.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. Challenges of using text classifiers for causal inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, pp. 4586, 2018.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4094–4103, 2019.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 804–813, 2011.

Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10923–10930, 2021.

Wenbo Zhang, Tong Wu, Yunlong Wang, Yong Cai, and Hengrui Cai. Towards trustworthy explanation: On causal rationalization. In *International Conference on Machine Learning*, pp. 41715–41736. PMLR, 2023.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Yuxiang Zhou and Yulan He. Causal inference from text: Unveiling interactions between variables. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10559–10571, 2023.

A Proofs and Theoretical Details

A.1 Assumptions

For convenience, we first restate the assumptions following their order in the main paper.

(A1). Confounding Sufficiency. The true confounding signals $\boldsymbol{W}^c$ (contained within the observed text $\boldsymbol{W}$) are sufficient for adjustment:

$$\{Y(0), Y(1)\} \perp T \mid \boldsymbol{W}^c.$$

(A2). Positivity on True Signals. Each unit has a nonzero probability of receiving both treatment and control conditional on the true confounding signals $\boldsymbol{W}^c$:

$$0 < p_1 \leq \Pr(T = 1 \mid \boldsymbol{W}^c) \leq p_2 < 1, \forall \boldsymbol{W}^c.$$

(A3). Consistency. The observed outcome corresponds to the potential outcome under the received treatment: $Y = Y(T)$.

(A4). Additive independent noise. (i) There exists a measurable function $g^\star : \mathcal{W}^c \to [0, 1]$ and a noise term ε_T such that $T = g^\star(\boldsymbol{W}^c) + \varepsilon_T$, where ε_T is independent of $\boldsymbol{W}$, $\mathbb{E}[\varepsilon_T] = 0$, and $\mathbb{E}[\varepsilon_T^2] < \infty$. (ii) There exists a measurable function $Q_T^\star : \{0, 1\} \times \mathcal{W}^c \to \mathbb{R}$ and a noise term ε_Y such that $Y = Q_T^\star(\boldsymbol{W}^c) + \varepsilon_Y$, where ε_Y is independent of $(T, \boldsymbol{W})$, $\mathbb{E}[\varepsilon_Y] = 0$, and $\mathbb{E}[\varepsilon_Y^2] < \infty$.

(A5). Regularity of data and representation. The treatment $T \in \{0, 1\}$ and outcome $Y \in [-M, M]$ for some $M > 0$. The observed text $\boldsymbol{W}$ is encoded by a pretrained model into a bounded continuous feature vector $\mathbf{x} = \Lambda(\boldsymbol{W}) \in [-1, 1]^d$, whose distribution admits a bounded density on $[-1, 1]^d$, where d is the encoder output dimension.

(A6). Smoothness. For each target function $f \in \{g, Q_0, Q_1\}$, assume that f lies in a Sobolev ball $\mathcal{W}^{\beta,\infty}([-1, 1]^d)$ with smoothness $\beta > 0$, i.e.,

$$f(\mathbf{x}) \in \mathcal{W}^{\beta,\infty}([-1, 1]^d) := \Big\{ f : \max_{|\alpha| \leq \beta} \operatorname{ess\,sup}_{\mathbf{x} \in [-1,1]^d} |D^\alpha f(\mathbf{x})| \leq 1 \Big\},$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$, $|\alpha| = \sum_j \alpha_j$, and $D^\alpha f$ denotes the weak derivative.

A.2 Proof of Proposition 3.1

We restate the result for convenience. For any candidate subset $\tilde{\boldsymbol{W}} \subseteq \boldsymbol{W}$, define

$$g(\tilde{\boldsymbol{W}}) = \mathbb{E}[T \mid \tilde{\boldsymbol{W}}], \qquad Q_t(\tilde{\boldsymbol{W}}) = \mathbb{E}[Y \mid T = t, \tilde{\boldsymbol{W}}],$$

and residuals

$$r_T = T - g(\tilde{\boldsymbol{W}}), \qquad r_Y = Y - Q_T(\tilde{\boldsymbol{W}}).$$

Let

$$D(\tilde{\boldsymbol{W}}) = \mathrm{HSIC}(r_T, r_Y)$$

denote the population-level HSIC between the treatment and outcome residuals, computed with characteristic and bounded kernels, under finite second moments. Under Assumptions (A1), (A3), and (A4), we show that

$$(Y(0), Y(1)) \perp T \mid \tilde{\boldsymbol{W}} \quad \Longrightarrow \quad D(\tilde{\boldsymbol{W}}) = 0.$$

Step 1. Residuals recover exogenous noise under sufficiency By the consistency assumption, $Y = Y(T)$. Assumption (A1) states that $(Y(0), Y(1)) \perp T \mid \boldsymbol{W}^c$, and we further suppose that $\tilde{\boldsymbol{W}}$ is sufficient in the sense that

$$(Y(0), Y(1)) \perp T \mid \tilde{\boldsymbol{W}}.$$

In particular, $\tilde{\boldsymbol{W}}$ contains all information in $\boldsymbol{W}^c$ that is relevant for the treatment and outcome mechanisms.

Under the additive-noise treatment model in Assumption (A4),

$$T = g^\star(\boldsymbol{W}^c) + \varepsilon_T, \qquad \varepsilon_T \perp \boldsymbol{W}, \qquad \mathbb{E}[\varepsilon_T] = 0,\ \mathbb{E}[\varepsilon_T^2] < \infty.$$

Since $\tilde{\boldsymbol{W}}$ is sufficient for the treatment mechanism, we have

$$g(\tilde{\boldsymbol{W}}) = \mathbb{E}[T \mid \tilde{\boldsymbol{W}}] = \mathbb{E}[g^\star(\boldsymbol{W}^c) + \varepsilon_T \mid \tilde{\boldsymbol{W}}] = g^\star(\boldsymbol{W}^c),$$

and therefore

$$r_T = T - g(\tilde{\boldsymbol{W}}) = \big(g^\star(\boldsymbol{W}^c) + \varepsilon_T\big) - g^\star(\boldsymbol{W}^c) = \varepsilon_T.$$

Similarly, under the additive-noise outcome model in Assumption (A4),

$$Y = Q_T^\star(\boldsymbol{W}^c) + \varepsilon_Y, \qquad \varepsilon_Y \perp (T, \boldsymbol{W}), \qquad \mathbb{E}[\varepsilon_Y] = 0,\ \mathbb{E}[\varepsilon_Y^2] < \infty.$$

Sufficiency of $\tilde{\boldsymbol{W}}$ for the outcome mechanism implies

$$Q_t(\tilde{\boldsymbol{W}}) = \mathbb{E}[Y \mid T = t, \tilde{\boldsymbol{W}}] = \mathbb{E}[Q_T^\star(\boldsymbol{W}^c) + \varepsilon_Y \mid T = t, \tilde{\boldsymbol{W}}] = Q_t^\star(\boldsymbol{W}^c),$$

and hence

$$r_Y = Y - Q_T(\tilde{\boldsymbol{W}}) = \big(Q_T^\star(\boldsymbol{W}^c) + \varepsilon_Y\big) - Q_T^\star(\boldsymbol{W}^c) = \varepsilon_Y.$$

Thus, under sufficiency of $\tilde{\boldsymbol{W}}$, the residuals (r_T, r_Y) coincide with the exogenous noise terms $(\varepsilon_T, \varepsilon_Y)$.

Step 2. Independence of residuals By Assumption (A4), the exogenous noises are independent:

$$\varepsilon_T \perp \varepsilon_Y.$$

Since we have shown $r_T = \varepsilon_T$ and $r_Y = \varepsilon_Y$, it follows that

$$r_T \perp r_Y.$$

Step 3. HSIC characterization of independence For bounded characteristic kernels $k_T(\cdot,\cdot)$ and $k_Y(\cdot,\cdot)$, the Hilbert–Schmidt independence criterion satisfies

$$\mathrm{HSIC}(U, V) = 0 \quad \Longleftrightarrow \quad U \perp V,$$

for any random variables U, V with finite second moments (Gretton et al., 2005; 2007). Applying this to $U = r_T$ and $V = r_Y$ yields

$$r_T \perp r_Y \quad \Longrightarrow \quad \mathrm{HSIC}(r_T, r_Y) = 0.$$

Step 4. Conclusion Combining Steps 1–3, we conclude that if $\tilde{\boldsymbol{W}}$ is sufficient for unconfoundedness, then the treatment and outcome residuals are independent and their HSIC equals zero:

$$(Y(0), Y(1)) \perp T \mid \tilde{\boldsymbol{W}} \quad \Longrightarrow \quad D(\tilde{\boldsymbol{W}}) = \mathrm{HSIC}(r_T, r_Y) = 0.$$

The contrapositive follows immediately under the structural additive-noise model: if $D(\tilde{\boldsymbol{W}}) > 0$, then r_T and r_Y are dependent, and $\tilde{\boldsymbol{W}}$ cannot be sufficient for unconfoundedness. □

A.3 Proof of Theorems 4.1, 4.2, and 4.3

We restate the result here for completeness. Under Assumptions (A5)–(A6), consider the joint penalized objective

$$\min_{\theta,\phi} \frac{1}{n}\sum_{i=1}^{n}\Big[\mathcal{L}_{\text{sup},i} + \mu\,\mathcal{R}_{\text{sparse}}(s_i) + \gamma\,\text{HSIC}(r_{T,i}, r_{Y,i})\Big],$$

where $\mathcal{L}_{\text{sup},i}$ is the supervised prediction loss, $\mathcal{R}_{\text{sparse}}$ is the sparsity regularizer, and $\text{HSIC}(r_{T,i}, r_{Y,i})$ measures residual dependence. Let $(\hat{\theta}, \hat{\phi})$ be the minimizer, and denote $\hat{f}$ as either $\hat{g}$ or $\hat{Q}_t$. If $\mu, \gamma = O(1)$ are fixed constants independent of n, then for an MLP-ReLU network with width $H \asymp n^{\frac{d}{2(\beta+d)}} \log^2 n$ and depth $D \asymp \log n$, there exists $C > 0$ such that, with probability at least $1 - \exp(-n^{\frac{d}{\beta+d}} \log^8 n)$,

$$\|\hat{f} - f\|^2_{L_2(P_X)} \le C\Big(n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log\log n}{n}\Big).$$

The joint objective can be viewed as the Lagrangian form of a constrained empirical risk minimization:

$$\begin{aligned}
&\min_{\theta,\phi} \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{\text{sup},i},\\
&\text{s.t.}\ \ \frac{1}{n}\sum_{i=1}^{n}\mathcal{R}_{\text{sparse}}(s_i) \le \Lambda_1, \quad \frac{1}{n}\sum_{i=1}^{n}\text{HSIC}(r_{T,i}, r_{Y,i}) \le \Lambda_2,
\end{aligned}$$

for some (Λ_1, Λ_2) determined by the Lagrange multipliers (μ, γ).

By standard Lagrangian duality, when $\mu, \gamma > 0$ are fixed and independent of n, this penalized and constrained formulations are equivalent. Hence, optimization in Equation equation 4 is equivalent to minimizing the supervised loss over the restricted function class

$$\mathcal{F}(\Lambda) = \big\{f \in \mathcal{F} : \mathcal{R}_{\text{sparse}}(f) \le \Lambda_1,\ \text{HSIC}(f) \le \Lambda_2\big\}.$$

The regularizers therefore act as soft constraints that shrink the feasible hypothesis space without changing the population minimizer f_*, as long as $\mathcal{R}_{\text{sparse}}(f_*)$ and $\text{HSIC}(f_*)$ are finite.

Following Theorem 1 in Farrell et al. (2021), let $\mathcal{F}$ denote the class of MLP-ReLU networks, width H and depth D. Under Assumptions (A5)–(A6), for an MLP-ReLU network with width $H \asymp n^{\frac{d}{2(\beta+d)}} \log^2 n$ and depth $D \asymp \log n$, there exists $C > 0$ such that, with probability at least $1 - \exp(-n^{\frac{d}{\beta+d}} \log^8 n)$,

$$\|\hat{f} - f\|^2_{L_2(P_X)} \le C\Big(n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log\log n}{n}\Big),$$

which gives the error bound for unrestricted function class.

Since $\mathcal{R}_{\text{sparse}}$ is bounded (e.g., the entropy penalty lies in $[0, \log 2]$) and HSIC is bounded by the product of kernel norms, and $\mu, \gamma = O(1)$, the penalized loss in Equation equation 4 differs from the unpenalized version by at most

$$\Big|\frac{1}{n}\sum_{i=1}^{n}\big[\mu\,\mathcal{R}_{\text{sparse}}(s_i) + \gamma\,\text{HSIC}(r_{T,i}, r_{Y,i})\big]\Big| = O_p\Big(\tfrac{1}{n}\Big),$$

which is asymptotically negligible. Therefore, the Lagrange penalties modify only the constants and polylog factors, leaving the main convergence rate unchanged.

Given the nonasymptotic L_2 bounds of the MLP approximation error, the convergence rate of nuisance estimators, IPW and AIPW estimators of ATE, and the nonasymptotic normality of AIPW estimator directly follow from classical semiparametric theory, as Theorem 3 in (Farrell et al., 2021). Therefore, we omit the proof for brevity.

B Semi-Synthetic Data Construction

We synthesize treatments and outcomes from text-derived binary confounders following a standard potential-outcomes framework.

Confounders from text For each clinical note, we extract two binary indicators using simple keyword matching:

$$C_i = (\texttt{has_shock}_i, \texttt{has_infection}_i) \in \{0,1\}^2,$$

where $\texttt{has_shock}_i = 1$ if shock-related terms appear in the text, and $\texttt{has_infection}_i = 1$ if infection- or sepsis-related terms appear.

Propensity scores and treatment assignment To produce the propensity score g_i, each combination of $(\texttt{has_shock}, \texttt{has_infection})$ corresponds to a pre-specified cell-wise propensity. Treatments are independently drawn as $T_i \sim \text{Bernoulli}(g_i)$. This design introduces confounding through $C \to T$, while maintaining overlap by bounding g_i away from 0 and 1.

Outcome model and potential outcomes Control potential outcomes are generated via a linear model with additive Gaussian noise:

$$\begin{aligned} Y_i(0) &= \beta_0 + \beta_1 \cdot \texttt{has_shock}_i + \beta_2 \cdot \texttt{has_infection}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \\ Y_i(1) &= Y_i(0) + \tau_{\text{True}}. \end{aligned}$$

Our simulation is carefully designed to reflect key challenges in real-world text-based causal inference. The text inputs are actual MIMIC-III clinical notes, preserving the semantic and distributional complexity of real clinical language. We introduce nonlinear relationships between confounders, treatment, and outcome through tanh and sigmoid-based transformations, as well as confounder-treatment interaction terms, which make the estimation problem more realistic and challenging. Together, these elements create a testbed that is both analytically tractable and empirically demanding, making it well-suited for evaluating causal inference methods under realistic yet controlled conditions.

C Implementation Details

Selectors and sparsity penalty. We implement $\boldsymbol{s} \in [0,1]^d$ with a temperature-controlled concrete/Gumbel–Softmax gate. For selective rationalization, we use a temperature parameter $\eta = 1.0$ to control the sharpness of the gating distribution. A lower temperature leads to sparser token selection, promoting interpretability and improving stability in high-dimensional settings. The entropy penalty is

$$\mathcal{R}_{\text{ent}}(\boldsymbol{s}) = \frac{1}{d} \sum_{j=1}^{d} \big\{ -s_j \log s_j - (1 - s_j) \log(1 - s_j) \big\}.$$

HSIC penalty. Given mini-batch residuals $\{e_T^i, e_Y^i\}_{i=1}^B$ computed from current nuisance networks, we estimate HSIC using the unbiased V-statistic with Linear/Gaussian kernels of bandwidth fixed by the median heuristic per batch.

Network sizes. The CATR model uses a pretrained transformer encoder, ClinicalBERT (Wang et al., 2023; Liu et al., 2025), with hidden size 768, followed by three task-specific heads:

- **Propensity head:** one hidden layer of dimension 256, ReLU activation, outputting a single logit.
- **Outcome heads (Q_0, Q_1):** two hidden layers of width 256 with ReLU activation, each outputting one scalar.
- **Token-importance scorer:** a linear layer.

The total loss is a weighted sum of (i) binary cross-entropy loss for propensity, (ii) binary cross-entropy loss for outcomes, (iii) a sparsity regularizer (L1 or entropy) on token weights, and (iv) an HSIC penalty on residuals:

$$\mathcal{L} = \mathcal{L}_{\text{prop}} + \beta\mathcal{L}_{\text{out}} + \mu\mathcal{L}_{\text{sp}} + \gamma\mathcal{L}_{\text{HSIC}}.$$

Here β weights the supervised losses, μ controls sparsity, and γ controls the HSIC penalty, consistent with the notation in Equation equation 4.

Hyperparameter. In practice, we balance the four loss components (treatment supervision, outcome supervision, sparsity regularization, and HSIC residual decorrelation) at roughly a 9:9:1:1 ratio in magnitude. To achieve this balance, the corresponding weights (β, μ, γ) are set to (0.4, 0.02, 2) throughout all experiments.

Optimization

We train with Adam using separate learning rates for the encoder and heads, 5e-5 and 2e-5, respectively, and batch size 32. Mixed-precision training (AMP) is used on GPU. Early stopping is based on validation loss with patience 3.

Evaluation Metrics

Across replications, we report:

- Mean(|Bias|): $\frac{1}{R}\sum_{r=1}^{R} |\widehat{\tau}_r - \tau_{\text{true}}|$;
- SD($\widehat{\text{ATE}}$): empirical standard deviation of $\widehat{\tau}_r$ over replications;
- Avg SE: mean of the bootstrap standard deviations, $\frac{1}{R}\sum_{r=1}^{R} \widehat{\text{SE}}_r$;
- C.I. Coverage (95%): proportion of confidence intervals that cover the true ATE, $\frac{1}{R}\sum_{r=1}^{R} \mathbb{I}\{\tau_{\text{true}} \in [\widehat{\tau}_r \pm 1.96\,\widehat{\text{SE}}_r]\}$;
- ESS Ratio: ratio of the estimated effective sample size to its oracle counterpart, $\text{ESS}_{\text{ratio}} = \text{ESS}(\widehat{w})/\text{ESS}(w^*)$, where $\text{ESS}(w) = (\sum_i w_i)^2/\sum_i w_i^2$ and w^* denotes the true inverse-propensity weights. Values closer to 1 indicate better recovery of the effective sample size and stronger empirical overlap;
- Clipping Frac.: fraction of units with estimated propensity scores $\widehat{g}_i \notin [0.03, 0.97]$, indicating observational-level positivity violations.

Environment and computation resources

PyTorch 2.7.1; CUDA 12.4; Python 3.11. All experiments are conducted on a server with 4 NVIDIA RTX A6000 GPUs.

Open access to code

All code required to reproduce our experiments is provided in the supplementary materials.

D Structured Variable List of Real Data

The following table lists the structured variables used for the MIMIC-III case study. All these variables are available in MIMIC-III but are joined from different tables.

E Supplementary Experimental Results

Table D.1: Structured variables from MIMIC-III.

Category	**Variable**	**Description**
Demographics & Baseline	age	Patient age (truncated at 89)
	gender	Biological sex
	elixhauser	Elixhauser comorbidity score
	SOFA	Sequential Organ Failure Assessment score
	SIRS	Systemic Inflammatory Response Syndrome score
Vital Signs	GCS	Glasgow Coma Scale (level of consciousness)
	HR	Heart rate (beats per minute)
	SysBP	Systolic blood pressure
	MeanBP	Mean arterial pressure
	RR	Respiratory rate
	SpO2	Oxygen saturation (%)
	Temp_C	Body temperature (°C)
	Shock_Index	Heart rate divided by systolic blood pressure
Respiratory Status	FiO2_1	Fraction of inspired oxygen
	PaO2_FiO2	Oxygenation index (PaO2 / FiO2)
	mechvent	Mechanical ventilation applied (baseline proxy)
Laboratory Measurements	Hb	Hemoglobin concentration
	WBC_count	White blood cell count
	Platelets_count	Platelet count
	Glucose	Blood glucose level
	BUN	Blood urea nitrogen
	Creatinine	Kidney function indicator
	CO2_mEqL	Carbon dioxide (acid-base status)
	Sodium	Serum sodium
	Potassium	Serum potassium
	Magnesium	Serum magnesium
	Calcium	Total serum calcium
	Ionised_Ca	Ionized calcium
	SGOT	AST enzyme (liver function)
	SGPT	ALT enzyme (liver function)
	Total_bili	Total bilirubin
	Albumin	Serum albumin
	PT	Prothrombin time
	PTT	Partial thromboplastin time
	INR	International normalized ratio
	Arterial_pH	Arterial blood pH
	paCO2	Arterial partial pressure of CO2
	Arterial_BE	Base excess
	Arterial_lactate	Lactate concentration
	HCO3	Bicarbonate

Fig A. Top-k selected tokens

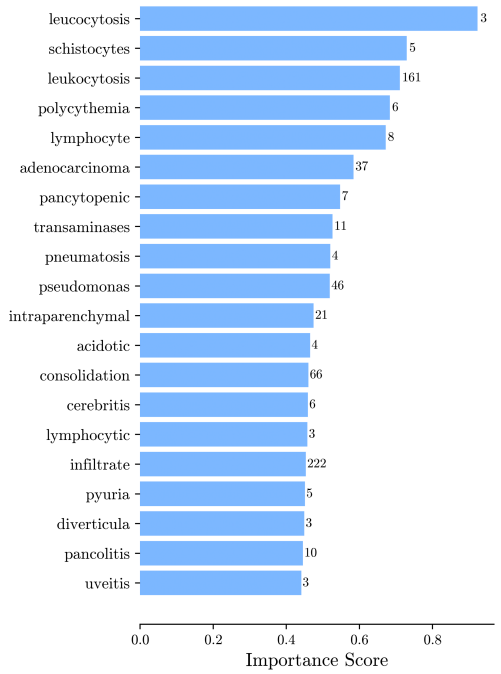

Fig B. Selective rationalization sample note

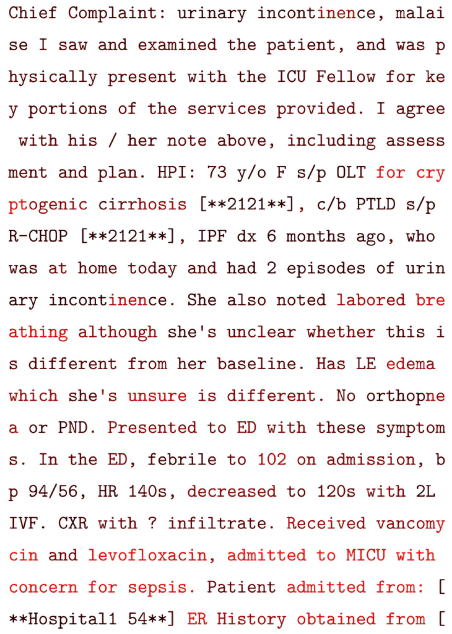

Figure E.1: Fig. A shows top-k selected tokens that are medically informative. Fig. B visualizes the selective rationalization output with color-encoded selection importance scores.

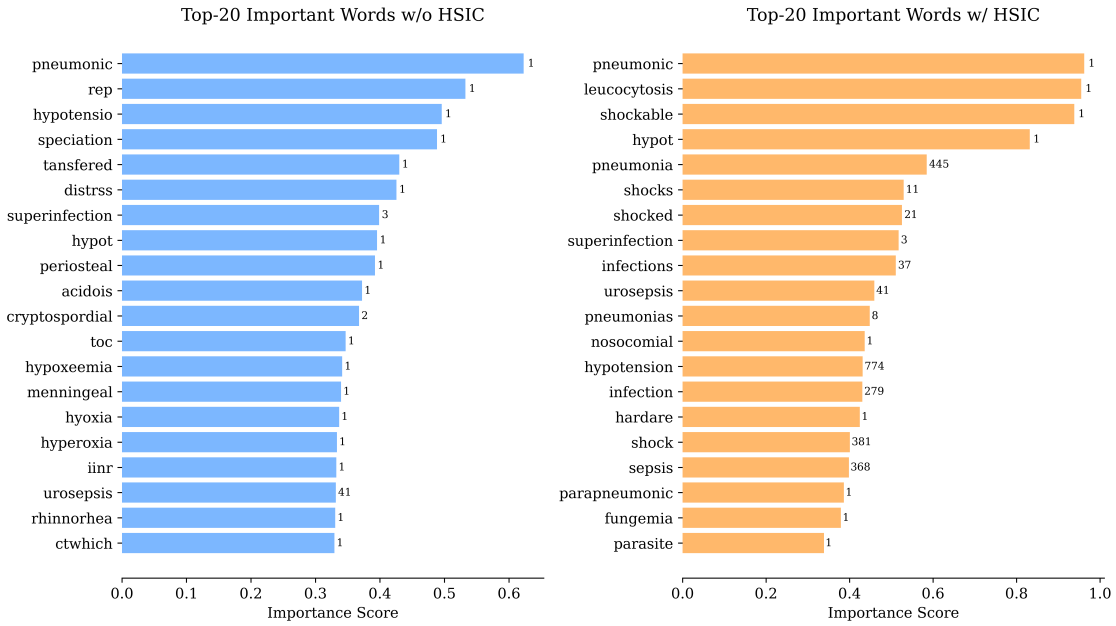

Figure E.2: Comparison of importance score distributions with and without HSIC regularization.