One Concept at a Time: Subspace-Constrained Causal Inference for High-Dimensional Treatments

Gaurav Sood

December 7, 2025

Abstract

Social scientists increasingly wish to reason causally about high-dimensional treatments (texts, images, prompts) while isolating the effect of a single latent concept (respectful tone, partisan framing, a policy clause) and holding all else fixed. We propose using pretrained models as measurement devices and activation-steering methods as treatment generators. The framework addresses two questions: first, how to detect and estimate a low-dimensional concept-tangent subspace in activation space from minimal counterfactual pairs; second, how to generate interventions confined to this subspace and obtain causal estimates of outcomes with respect to movement within it. The pipeline has two stages. Stage 1 uses minimal edit pairs and a pretrained encoder to estimate a concept-tangent subspace, with diagnostics (low-rank structure, stability, semantic alignment, entanglement with nuisance attributes) that can explicitly reject the existence of a clean subspace for a given encoder and concept. When Stage 1 supports a low-dimensional approximation, activations decompose into a concept coordinate and an orthogonal residual. Stage 2 uses this decomposition to (a) constrain activation-steering maps to the concept subspace, yielding approximate minimal edits, and (b) treat concept coordinates as treatments in a double/debiased ML estimator that controls for non-concept variation via the residual. The resulting estimand is a local, representation-dependent linear effect in the learned concept coordinate, not a universal causal effect of the underlying human concept. We discuss identification assumptions, limitations, and connections to matching and synthetic control.

1 Introduction

Many causal questions in the social sciences involve complex treatments: long texts, images, or combinations of modalities. A campaign might send voters SMS messages

whose tone, length, and policy emphasis vary. A bureaucracy might respond to citizen requests in more or less respectful language. A platform might show users different textual or visual framings of the same information. Researchers often want to know how changes in a specific latent concept—say, respectful tone—affect outcomes, while keeping other aspects of the treatment fixed.

Formally, let Y be an outcome, X covariates, and T a complex treatment object (e.g. text). The researcher is interested in how Y changes when a latent concept C embedded in T is varied, holding other properties of T constant. Two difficulties make this hard.

First, any particular representation of T (bag-of-words, topics, neural embeddings) typically entangles many semantic and stylistic properties. Changing a coordinate correlated with C may also change topic, sentiment, or length. Second, we rarely observe *minimal counterfactuals*: for a given unit, we do not see both T and a version T' that differs only in C. In observational data, units who choose high-C treatments may differ systematically in other ways that affect Y.

Recent work on text as data and causal inference with text has made both points explicit: representation choices are central to identification, and re-using data to both learn representations and estimate effects can induce serious biases [Egami et al., 2022, Feder et al., 2022, Veitch et al., 2020]. At the same time, work on activation steering in large models shows that meaningful behavioral dimensions can often be controlled by moving along particular directions in activation space [Turner et al., 2023, Zou et al., 2023, Li et al., 2023, Wu et al., 2024, Rodriguez et al., 2025].

This paper proposes a framework that connects these threads. We treat a large pretrained model as a *measurement system* that maps complex treatments T to internal activation states $h(T)$. Using a small set of *minimal edit pairs* (T^-, T^+) that differ primarily in a concept C, we attempt to estimate a low-dimensional *concept-tangent subspace* U_C in activation space. Crucially, we also provide diagnostics that sometimes tell us that a clean such subspace *does not* exist, at least for the given encoder and minimal pairs. When the diagnostics are favorable, we decompose each activation as

$$h(T) = U_C z(T) + r(T),$$

where $z(T)$ is a low-dimensional concept coordinate and $r(T)$ is an orthogonal residual. We use this decomposition in two ways.

First, we constrain activation-steering maps—for example, the affine optimal-transport maps of LinEAS [Rodriguez et al., 2025]—to operate only inside the span of U_C. This yields a representation-level notion of "minimal edits": interventions preserve the residual $r(T)$ while shifting $z(T)$. Second, we treat $z(T)$ as a treatment and $(X, r(T))$ as controls in a partially linear causal model estimated via double/debiased machine learning (DML)

[Chernozhukov et al., 2018]. This yields an estimator for the effect of moving along the concept coordinate.

The key interpretive move is to be explicit about what this estimand is. It is not a universal effect of the human concept C in the world. Instead, it is a *local, encoder- and subspace-dependent average effect* of changes in the learned concept coordinate $Z = z(T)$ on Y, conditional on the orthogonal residual representation. We regard this as a useful, well-defined target for representation-level causal analysis.

Contributions

We summarize our contributions as follows.

- We formalize the use of minimal counterfactual pairs to estimate *concept-tangent subspaces* in activation space and, equally importantly, to compute diagnostics that can accept or reject the claim that variation in a concept is well-approximated by a low-dimensional linear subspace for a given encoder.
- We propose constraining activation-steering maps, such as those learned by LinEAS, to act only in concept-tangent subspaces, thereby encoding a representation-level notion of minimal edits.
- We define a *representation-level causal estimand*: a local linear effect of the concept coordinate on the outcome in a partially linear model, conditional on the residual representation and covariates. We articulate assumptions under which double machine learning recovers this estimand.
- We discuss identification challenges, including the quality of minimal pairs, the geometry and stability of concept subspaces, the use of post-treatment residuals as controls, and domain shift between minimal pairs and observational data. We outline diagnostics and connections to matching and synthetic control in concept space.

The framework is not a turnkey solution; it rests on substantive assumptions about representation geometry and confounding. Its value lies in translating informal talk of "changing only one thing in a model" into explicit objects, diagnostics, and estimands that can be scrutinized.

2 Background and related work

2.1 Text as treatment and representation choices

A growing literature uses text as treatment, outcome, or confounder in causal analysis. Egami et al. [2022] formalize how to construct text-based measures for causal inference and emphasize the dangers of re-using data for both discovery and estimation. Feder et al. [2022] survey methods for causal inference with text, covering settings where text is a treatment, outcome, mediator, or control. They stress that latent representations can introduce new identification problems if they are not aligned with causal structure.

Veitch et al. [2020] propose adapting text embeddings for causal inference by learning representations that satisfy conditional independence properties implied by a causal graph. Their work treats the embedding as part of the modeling choice subject to causal constraints, which resonates with our treatment of a pretrained encoder as a measurement system whose geometry matters.

2.2 Representation learning for causal inference

Several works learn representations explicitly for causal effect estimation. Li and Fu [2017] learn nonlinear representations on which treated and control units can be matched; the representation is trained to both balance covariate distributions across treatment groups and predict outcomes. Yao et al. [2021] decompose learned representations into shared, treatment-specific, and control-specific components to estimate individual treatment effects. Parbhoo et al. [2021] extend such ideas to combinations of treatments, learning neural representations that capture interactions between components of multi-valued treatments.

These methods learn representations from observational outcome data and treatment labels, rather than from minimal pairs or tangent information in activation space. They focus on balancing and prediction rather than on explicitly aligning a subspace with a particular concept.

2.3 Activation steering and LinEAS

Activation-steering methods modify internal representations of pretrained models to change behavior while leaving weights fixed. Simple methods include *activation addition*, which adds a steering vector derived from differences between two sets of activations [Turner et al., 2023], and other linear editing schemes that rely on paired prompts or classifier hyperplanes [Zou et al., 2023, Li et al., 2023]. Representation finetuning (ReFT) learns low-rank interventions on hidden states to improve model behavior with

paired supervision [Wu et al., 2024].

LinEAS (*linear end-to-end activation steering*) is a recent method that learns affine maps on activations across multiple layers using a global distributional loss grounded in optimal transport [Rodriguez et al., 2025]. Given unpaired sets of source and target prompts, LinEAS interlaces the frozen network with diagonal affine maps at selected layers and trains these maps to jointly minimize the sum of one-dimensional Wasserstein distances between transported source activations and target activations at each layer, with group-sparse regularization to select a small set of neurons. LinEAS can steer language and text-to-image models with few unpaired samples and modest compute, achieving competitive performance on toxicity mitigation and concept editing tasks. We will treat such maps as representation-level transport operators that can be constrained to act in concept-tangent subspaces.

2.4 Optimal transport, matching, and synthetic control

Optimal transport (OT) provides a natural framework for matching treated and control distributions. OT-based methods have been applied to treatment effect estimation by constructing transport plans that reweight control units to approximate the treated distribution in covariate space [Wang et al., 2023]. More broadly, OT has been proposed as a unifying lens on causal identification and distributional adjustment [Gunsilius, 2025].

Synthetic control constructs a weighted combination of control units whose pre-treatment outcomes match those of a treated unit, yielding a counterfactual trajectory [Abadie et al., 2010]. Deep synthetic control methods extend this idea using learned representations and regularization to handle high-dimensional outcomes and covariates [Ramachandra, 2025].

We interpret activation-steering maps, and especially globally trained maps such as LinEAS, as OT-like transformations in representation space. Our concept-tangent subspaces define a restricted geometry for such transformations.

2.5 Double/debiased machine learning

Double/debiased machine learning (DML) provides general tools for estimating low-dimensional causal parameters in the presence of high-dimensional nuisance functions using flexible machine learning [Chernozhukov et al., 2018]. In the partially linear regres-

sion model

$$Y = D^\top \theta_0 + g_0(X) + \zeta, \quad \mathbb{E}[\zeta \mid D, X] = 0,$$
$$D = m_0(X) + V, \quad \mathbb{E}[V \mid X] = 0,$$

DML constructs orthogonal moment functions and uses sample splitting and cross-fitting to obtain $\sqrt{n}$-consistent, asymptotically normal estimators of θ_0 even when g_0 and m_0 are estimated with regularized or nonparametric methods. Extensions cover multi-valued and continuous treatments and heterogeneous effects [Chernozhukov et al., 2018, Wager and Athey, 2018].

We will use DML with treatments given by concept coordinates in activation space and controls that include covariates and residual representations.

3 Setup and notation

We distinguish between observational data, used for causal estimation, and minimal edit data, used to learn concept-tangent subspaces and their diagnostics. We also fix a pretrained encoder whose activations define the representation space.

3.1 Observational data

For $i = 1, \ldots, n$, we observe:

- covariates $X_i \in \mathcal{X}$,
- a complex treatment $T_i \in \mathcal{T}$ (e.g. a text or prompt),
- an outcome $Y_i \in \mathbb{R}$.

We make no structural assumptions yet about how T_i is chosen, other than that it may depend on X_i and unobserved factors that also influence Y_i.

3.2 Minimal edit pairs

Separately, we have access to m *minimal edit pairs* $\{(T_j^-, T_j^+)\}_{j=1}^m$. Each pair is intended to differ primarily in a single concept C: T_j^+ is a version of T_j^- with concept C increased or toggled, while other aspects remain as similar as possible. Minimal pairs can be obtained by human editing, templated transformations, or cautious use of steering methods.

We will use these pairs to infer directions in activation space associated with C, and to assess whether a low-dimensional linear approximation is defensible.

3.3 Encoder and activations

Let f_θ be a pretrained encoder with parameters θ (e.g. the hidden state function of a language model at a given layer). For any treatment T, we write

$$h(T) \in \mathbb{R}^d$$

for a fixed-dimensional activation vector derived from f_θ at chosen layers and positions, reshaped as needed. For example, $h(T)$ might be the concatenation of layer-normalized residual stream activations averaged over tokens across several layers.

Throughout, we treat f_θ as fixed: all randomness arises from the data-generating process for (X, T, Y) and from the construction of minimal pairs.

4 Concept-tangent subspaces from minimal pairs

We now describe how to estimate a low-dimensional concept-tangent subspace from minimal edit pairs, and how to diagnose when such a subspace is not a good approximation.

4.1 Minimal-edit assumption

For each pair (T_j^-, T_j^+), we hope that:

1. T_j^+ differs from T_j^- primarily in the concept C of interest (e.g. level of respect), and
2. other aspects of T_j, such as topic, length, and style, are as similar as feasible.

In practice, this assumption is only approximate: modifying a concept may inadvertently change correlated features. Our estimation procedure therefore assumes that, across many pairs, the dominant systematic change in activations corresponds to C, while incidental changes average out.

4.2 Activation differences and low-rank structure

For each minimal pair, define the activation difference

$$\Delta h_j = h(T_j^+) - h(T_j^-) \in \mathbb{R}^d.$$

Stack these as columns of a matrix $D \in \mathbb{R}^{d \times m}$. We posit that the Δh_j lie *approximately* in a low-rank subspace associated with concept C, so that D has rapidly decaying singular values.

A simple estimator proceeds via principal component analysis. Compute the empirical covariance

$$S = \frac{1}{m} D D^\top,$$

and let $U_k \in \mathbb{R}^{d \times k}$ contain the top k eigenvectors of S, with $k \ll d$ and $U_k^\top U_k = I_k$. We call $\text{span}(U_k)$ the *concept-tangent subspace* for C.

The choice of k can be guided by the eigenvalue spectrum, cross-validation on downstream tasks, or prior beliefs about the complexity of C. In practice, k will be small (e.g. 1–10).

4.3 Concept coordinates and residuals

Given U_k, we can decompose any activation $h \in \mathbb{R}^d$ as

$$h = U_k z + r, \quad z = U_k^\top h \in \mathbb{R}^k, \quad r = (I_d - U_k U_k^\top) h \in \mathbb{R}^d. \tag{1}$$

We interpret z as a vector of *concept coordinates* and r as an *orthogonal residual* capturing all other variation.

When $k = 1$, z is a scalar measuring the strength or sign of C in representation space, relative to the orientation implied by the minimal pairs. When $k > 1$, z may capture multiple facets or contexts of C.

This decomposition depends on the encoder, the minimal pairs, and the eigenvector choice; it is not unique. Any orthogonal rotation of U_k yields an equivalent subspace. We will treat U_k as fixed once estimated, recognizing that θ and U_k are part of the model specification.

4.4 Stage-1 diagnostics and possible rejection

Stage 1 should not only estimate U_k but also determine whether it is reasonable to treat $\text{span}(U_k)$ as a low-dimensional, concept-aligned tangent subspace that is approximately

orthogonal to other treatment properties. We therefore compute diagnostics along four axes.

Low-rank structure. We quantify how well a small number of principal components explain the variation in Δh_j.

- *Explained variance.* For a candidate k, define

$$R_k^2 = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{d} \lambda_i},$$

  where λ_i are the eigenvalues of S. If R_k^2 is small for all modest k, then concept variation is not well-concentrated.

- *Reconstruction error.* For each j, compute

$$\hat{\Delta h_j} = U_k U_k^\top \Delta h_j, \quad \text{err}_j = \frac{\|\Delta h_j - \hat{\Delta h_j}\|}{\|\Delta h_j\|}.$$

  If typical err_j is large, the low-rank approximation is poor.

Stability across samples and contexts. We assess whether the estimated subspace is stable across different subsets of minimal pairs.

- *Split-half stability.* Randomly split the minimal pairs into two halves, estimate $U_k^{(a)}$ and $U_k^{(b)}$, and compute principal angles between the two subspaces. Repeat over many splits. Large and highly variable angles indicate instability.

- *Context stability.* When minimal pairs can be grouped by template or domain, estimate U_k separately within each group and compare the resulting subspaces. Large differences suggest that the concept direction is highly context-specific.

Semantic alignment with the intended concept. We test whether movement along the leading principal directions actually tracks changes in C.

Assume we have concept labels or scores for minimal pairs—either human ratings or model-based scores. For each pair, project Δh_j onto the first principal component u_1 and compute $s_j = u_1^\top \Delta h_j$. We assess:

- whether the sign of s_j matches the direction of change in C (e.g. T_j^+ has higher respect than T_j^-);

- the correlation between s_j and the change in the concept score;
- the performance of a simple classifier that predicts which element of a pair has higher C using only z-coordinates.

If these alignment measures are weak or unstable across subsets, the principal directions are not reliably tracking the intended concept.

Entanglement with nuisance attributes and other concepts. Finally, we investigate whether the putative concept direction is heavily entangled with other features.

- *Nuisance leakage.* For each pair, compute changes in basic nuisance measures such as length, sentiment, perplexity, or topic classifier scores. Check whether these changes are strongly correlated with projections onto U_k. Strong correlations suggest that U_k is partly encoding nuisances.
- *Cross-concept interference.* When minimal pairs are available for multiple concepts C and C', estimate subspaces U_C and $U_{C'}$ separately and compute principal angles, as well as how well z_C coordinates predict labels for C', and vice versa. Near-collinearity or high cross-predictability indicates entanglement.

Stage-1 quality report. We aggregate these diagnostics into a qualitative assessment of the Stage 1 output for concept C. For example, we may define:

- a *low-rank score* L_C based on R_k^2 and reconstruction errors;
- a *stability score* S_C based on average cosines of principal angles across splits;
- an *alignment score* A_C based on correlations between projections and concept labels;
- an *entanglement score* E_C based on correlations with nuisance measures and interference with other concepts.

Stage 1 then assigns a regime:

- *Clean subspace:* high L_C, S_C, A_C, low E_C. Proceed to Stage 2 and interpret results as representation-level effects of C.
- *Noisy subspace:* moderate scores. Proceed with Stage 2 but flag estimates as exploratory.

- *No subspace:* low L_C or S_C, or very low A_C. In this case, the method explicitly reports that a low-dimensional, orthogonalizable concept subspace is not supported by the data for this encoder and concept, and Stage 2 should not claim to isolate C.

This rejection option is an important part of the framework: not all concepts need to be orthogonalizable in a given representation.

4.5 Remarks on layers and context

In transformer models, activations vary by layer and token. Our description has abstracted away from this by using a flattened h. In practice, one can:

- estimate a single U_k on concatenated activations across selected layers and positions;
- or estimate layer-specific subspaces $U_{k,\ell}$ using layer-specific $\Delta h_{j,\ell}$, then decompose h_ℓ at each layer.

Stability diagnostics can be run at each layer or for concatenated representations. Our causal framework applies to any fixed mapping $T \mapsto (Z, R)$ derived from minimal pairs, conditional on the Stage 1 quality report.

5 Steering within concept-tangent subspaces

The concept-tangent subspace can be used not just to measure concept coordinates but also to constrain interventions in activation space.

5.1 Activation steering and LinEAS

Activation-steering methods modify internal activations of frozen models to alter behavior. In LinEAS [Rodriguez et al., 2025], a generative model is viewed as a composition of frozen layer maps $f_1, \ldots, f_{L+1}$, and learned affine maps $T_1, \ldots, T_L$ are interleaved:

$$o = f_{L+1} \circ T_L \circ f_L \circ \cdots \circ T_1 \circ f_1(x).$$

Each T_ℓ acts on activations at layer ℓ via coordinate-wise affine transformations

$$T_\ell(z) = \omega_\ell \odot z + b_\ell,$$

with parameters ω_ℓ, b_ℓ. The maps are trained jointly to minimize a sum of one-dimensional Wasserstein distances between transported source activations and target activations at each layer, with sparse-group lasso regularization selecting a small subset of neurons. This yields low-data, modality-agnostic steering with a continuous strength parameter.

We can view LinEAS and related methods as learning a representation-level transport operator S that pushes the activation distribution associated with one behavior toward that of another.

5.2 Concept-constrained steering maps

Given a concept-tangent basis U_k, we can constrain steering maps to act only in span(U_k). At a single layer, this can be expressed as

$$S(h) = h + U_k B U_k^\top (h - \mu), \tag{2}$$

where $B \in \mathbb{R}^{k\times k}$ and μ is a reference mean activation (e.g. over source prompts). Then $S(h) - h \in$ span(U_k) for all h, and the residual component in (1) is preserved up to numerical error.

A simpler variant is an additive shift

$$S_\gamma(h) = h + U_k\gamma, \quad \gamma \in \mathbb{R}^k, \tag{3}$$

which replaces z by $z + \gamma$ while leaving r unchanged.

In a LinEAS-style setup, one can restrict each layer's affine map to have the form in (2) in layer-specific subspaces $U_{k,\ell}$. The global OT loss is then minimized over B_ℓ given fixed $U_{k,\ell}$, yielding a concept-constrained transport operator.

5.3 Synthetic treatments and approximate minimal edits

Concept-constrained steering provides a mechanism for generating approximate minimal pairs. Given a base treatment T:

1. Compute $h(T)$ and decompose $h(T) = U_k z(T) + r(T)$.
2. Choose a target concept coordinate $z^\star$ (or a steering strength λ) that represents a desired change in C.
3. Apply a steering map S (e.g. (3)) so that the transformed activation has concept coordinate close to $z^\star$ and residual approximately $r(T)$.

4. Decode the output representation into a synthetic treatment $T^\star$ using the pre-trained decoder.

Because S is constrained to span(U_k), changes in h occur only in concept directions. Nonlinearities and the decoding process mean that $T^\star$ may nonetheless differ from T in other ways, but the constraint is a principled way to reduce interference.

One can use such synthetic treatments in designed experiments (by actually showing $T^\star$ to subjects) or in simulation studies that probe the model's internal behavior.

6 Representation-level causal estimand

We now clarify what parameter we aim to estimate from observational data, given a fixed encoder and concept-tangent subspace that has passed Stage 1 diagnostics.

6.1 Representation-level potential outcomes

Thought experiment: fix a family of steering maps $\{S_\lambda\}_{\lambda\in\Lambda}$ that act only in span(U_k) and are indexed by a scalar strength parameter λ (e.g. linearly interpolating between identity and a full concept shift). For a given unit with base treatment T, consider the activation path

$$h_\lambda(T) = S_\lambda(h(T)),$$

and the corresponding decoded treatment T_λ. If we could expose the unit to T_λ instead of T, there would be a well-defined potential outcome $Y(\lambda)$.

Our observational data, however, do not contain such steering-based counterfactuals. Units choose T (and hence $h(T)$ and $Z = z(T)$) in ways we do not control, and we observe only Y corresponding to the realized choice. We therefore work with *representation-induced* variables in the observed data and treat the steering-based potential outcomes as a conceptual device guiding interpretation.

6.2 Induced variables and partially linear approximation

Given the encoder f_θ and concept-tangent basis U_k, define for each observational unit:

$$\begin{aligned} h_i &= h(T_i), \\ Z_i &= U_k^\top h_i \in \mathbb{R}^k, \\ R_i &= (I_d - U_k U_k^\top) h_i \in \mathbb{R}^d, \\ W_i &= (X_i, R_i). \end{aligned}$$

We posit that the conditional mean of Y given (Z, W) can be approximated by a partially linear model:

$$\mathbb{E}[Y \mid Z, W] \approx Z^\top \theta_0 + g_0(W), \tag{4}$$

for some parameter vector $\theta_0 \in \mathbb{R}^k$ and nuisance function g_0. The approximation is intended to be local in Z: θ_0 should be interpreted as an average derivative of the conditional mean with respect to the concept coordinate, conditional on W.

The parameter θ_0 is thus a *representation-level causal estimand*: it is defined conditional on the choice of encoder and subspace, and it describes how Y responds to variation in Z within that representation.

7 Assumptions and double machine learning

We now state assumptions under which a DML estimator recovers θ_0 in (4) and explain how the estimator works.

7.1 Assumptions

We fix an encoder f_θ and a concept-tangent basis U_k that Stage 1 has not rejected. We treat the induced variables (Z, R) as given.

Assumption A1 (Representation-level ignorability). For all z in the support of Z,

$$Y(z) \perp Z \mid W = (X, R),$$

where $Y(z)$ is the potential outcome associated with setting the concept coordinate to z while holding W fixed. Intuitively, after controlling for covariates and the residual representation, residual variation in Z is as good as random.

Assumption A2 (Local linearity in Z). There exists a parameter vector θ_0 and function g_0 such that, for typical (Z, W) in the observed support,

$$\mathbb{E}[Y \mid Z, W] = Z^\top \theta_0 + g_0(W) + \varepsilon, \quad \mathbb{E}[\varepsilon^2] \text{ small.}$$

That is, $Z^\top \theta_0 + g_0(W)$ is a good local approximation to the conditional mean.

Assumption A3 (Concept stability). The way Y responds to small representation-level perturbations that change Z along the steering path $\lambda \mapsto S_\lambda(h(T))$ is similar to the way it responds to naturally occurring differences in Z in the observational data. This links the observational estimand θ_0 to the effects one would estimate by experimentally randomizing λ.

Assumption A4 (Overlap). For the region of (Z, W) over which we interpret θ_0, the conditional distribution of Z given W has sufficient variation: there are no strata of W where Z is (almost) deterministic.

Assumptions A1–A4 are strong and model-dependent. They are analogous to standard selection-on-observables and overlap assumptions, but expressed in representation space with R acting as a high-dimensional proxy for non-concept features of the treatment. They do not assert that θ_0 is a universal effect of the underlying human concept, only that it is a well-defined representation-conditional parameter.

7.2 Double/debiased machine learning estimator

Under A1–A4, the partially linear model

$$Y_i = Z_i^\top \theta_0 + g_0(W_i) + \zeta_i, \quad \mathbb{E}[\zeta_i \mid Z_i, W_i] = 0, \tag{5}$$

$$Z_i = m_0(W_i) + V_i, \quad \mathbb{E}[V_i \mid W_i] = 0, \tag{6}$$

is compatible with the data-generating process. Here g_0 and m_0 are nuisance functions capturing, respectively, baseline outcome variation and the dependence of Z on W.

Following Chernozhukov et al. [2018], we construct an orthogonal estimating equation based on residuals. Let g and m be candidate functions and define the score

$$\psi(Y, Z, W; \theta, g, m) = (Z - m(W))^\top \big(Y - g(W) - \theta^\top (Z - m(W))\big).$$

Under (5)–(6), $\mathbb{E}[\psi(Y, Z, W; \theta_0, g_0, m_0)] = 0$. Moreover, the derivative of $\mathbb{E}[\psi]$ with respect to (g, m) vanishes at (g_0, m_0), which makes the moment condition *orthogonal* to small

first-stage errors.

The DML estimator proceeds as follows:

1. Split the sample into K folds. For each fold k, fit nuisance estimators $\hat{g}^{(-k)}$ and $\hat{m}^{(-k)}$ on the other $K-1$ folds using flexible machine learning methods.

2. For each observation i in fold k, compute residuals

$$\tilde{Y}_i = Y_i - \hat{g}^{(-k)}(W_i), \quad \tilde{Z}_i = Z_i - \hat{m}^{(-k)}(W_i).$$

3. Solve the second-stage regression

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \left(\tilde{Y}_i - \tilde{Z}_i^\top \theta\right)^2 = \left(\sum_i \tilde{Z}_i \tilde{Z}_i^\top\right)^{-1} \sum_i \tilde{Z}_i \tilde{Y}_i,$$

   assuming $\sum_i \tilde{Z}_i \tilde{Z}_i^\top$ is invertible.

Under regularity conditions on the complexity and estimation error of $\hat{g}$ and $\hat{m}$, $\hat{\theta}$ is $\sqrt{n}$-consistent and asymptotically normal, with variance that can be estimated from sample analogues of the score [Chernozhukov et al., 2018].

7.3 Multi-concept and multi-treatment settings

When $k > 1$, $Z_i \in \mathbb{R}^k$ contains multiple concept coordinates. The model (5) then interprets θ_0 as a vector of marginal effects of each coordinate, conditional on the others and on W. The same DML procedure applies, with multivariate regression in the second stage.

This provides a route to multi-treatment estimation where different concept dimensions act as treatments. For example, one might estimate the joint effects of tone, policy emphasis, and identity framing, each corresponding to a dimension of Z, while controlling for residual representation and covariates.

7.4 Interpreting R as a control

The residual representation R is a function of the treatment object T. Including R in W is therefore a form of controlling for a post-treatment variable with respect to the underlying concept intensity. This requires careful interpretation.

In our representation-level view, Z and R are simply coordinates of $h(T)$, and controlling for R means estimating the direct effect of movement in the concept coordinate

holding the residual representation fixed. If some of the effect of Z on Y flows through changes in R, conditioning on R will block those mediated paths and produce a direct-effect-type estimand. If unobserved factors jointly influence Z and R, conditioning on R could introduce bias via collider paths.

Thus, the meaning of θ_0 depends on how we regard R: as a proxy for confounders, as a mediator, or both. We view R as a high-dimensional proxy for non-concept features of the treatment that may be correlated with Z and with Y. Whether controlling for R is desirable depends on the substantive question.

8 Matching, optimal transport, and synthetic control in concept space

The concept-tangent decomposition suggests modified geometries for nonparametric causal estimators.

8.1 Matching in (Z, W) space

Matching methods aim to compare units with similar covariates across treatment levels. Rather than matching on raw text or arbitrary embeddings, we can match on (Z, W), where Z summarizes concept intensity and W bundles covariates and residual representation. This is analogous to matching in learned representation spaces [Li and Fu, 2017], but with an explicit decomposition into concept and non-concept components and with Stage 1 diagnostics guiding whether Z is meaningful.

8.2 Synthetic control with representation trajectories

In panel settings, treatments and representations evolve over time. One can compute (Z_{it}, W_{it}) over periods t and apply synthetic control to match pre-treatment trajectories of both concept coordinates and residual representations. For a unit whose concept coordinate changes at time t_0, we seek donor units whose pre-t_0 (Z_{it}, W_{it}) trajectories resemble those of the treated unit, and whose post-t_0 trajectories provide a counterfactual path for (Z_{it}, W_{it}) and hence for Y_{it} [Abadie et al., 2010, Ramachandra, 2025].

8.3 Optimal transport in concept space

Optimal transport can be used to reweight or map distributions between treated and control groups based on costs in representation space [Wang et al., 2023, Gunsilius, 2025].

The concept-tangent subspace provides a natural structure for defining such costs: one can emphasize discrepancies in Z while controlling changes in W, or parameterize transport maps that act primarily along concept directions. LinEAS itself can be seen as implementing affine OT in activation space [Rodriguez et al., 2025].

9 Identification challenges and diagnostics

We collect and discuss key challenges for identification and interpretation, and how Stage 1 diagnostics fit into them.

9.1 Quality of minimal pairs

Estimation of U_k assumes that minimal pairs isolate the concept C. In practice, edits that increase C may systematically affect other properties, such as length or hedging. If these shifts are correlated across pairs, the top principal components of D may mix C with other concepts.

Stage 1 alignment and entanglement diagnostics are designed to detect such problems: weak correlation between projections and concept labels, or strong correlation with nuisance measures, suggests that U_k is not capturing C cleanly. Mitigations include careful human design of minimal pairs, diversity in base texts to average out incidental changes, and iterative refinement of U_k informed by these diagnostics.

9.2 Geometry and stability of concept-tangent subspaces

We assume that there exists a low-dimensional linear subspace in which C varies while other properties are relatively stable. In highly nonlinear and context-dependent models, this may fail: the same conceptual change can be realized via different activation patterns across contexts and layers.

Stability diagnostics—split-half principal angles, context-specific subspaces, and cross-concept interference—provide empirical evidence on whether such a subspace exists and is shared across the minimal pair distribution. If these diagnostics fail, the framework should report that no robust global concept-tangent subspace is available for the specified encoder and concept, and causal analysis should not be interpreted as isolating C.

9.3 Representation-level ignorability and R as a control

Assumption A1 is strong: it requires that, conditional on (X, R), the remaining variation in Z is as good as random. Since R is a high-dimensional representation, it is plausible that it absorbs many aspects of T that might confound Z and Y, but this is not guaranteed.

Moreover, R may lie on causal paths from Z to Y, in which case controlling for it estimates a direct effect rather than a total effect. Researchers should be explicit about which effect they seek, and recognize that R-based control trades off bias from omitted confounding against potential over-control.

Sensitivity analysis can probe robustness of $\hat{\theta}$ to inclusion/exclusion of R, to alternative encoders, and to different concept subspaces.

9.4 Domain shift between minimal pairs and observational data

Minimal pairs may come from a different domain than the observational treatments: short synthetic emails versus long real bureaucratic messages, for example. The concept-tangent subspace estimated from (T_j^-, T_j^+) may then poorly capture variation in C in the observational domain.

Diagnostics include estimating U_k on different domains and comparing subspaces, and checking whether movements along U_k have similar semantic effects in both domains. If Stage 1 alignment scores drop sharply when evaluated on observational treatments, downstream estimates should be treated with skepticism.

9.5 Finite-sample uncertainty in $\mathrm{U_k}$

When m is small and d large, U_k is estimated with noise. Treating U_k as fixed in DML ignores this uncertainty. Stability diagnostics partially address this by revealing sampling variability. Formal inference that accounts for both stages is challenging; bootstrap procedures that re-estimate U_k and re-run DML may provide approximations but are computationally heavy.

9.6 Diagnostics for steering and causal estimates

Beyond Stage 1, at least three types of diagnostics are useful:

- *Subspace diagnostics:* eigenvalue spectra, stability of U_k, and sparsity patterns across layers.

- *Steering diagnostics:* human or auxiliary-model ratings of concept and non-concept attributes for T and steered $T^\star$, as functions of steering strength.
- *Causal diagnostics:* sensitivity of $\hat{\theta}$ to k, to the choice of layers and encoder, to inclusion/exclusion of R, and to different nuisance estimators.

These diagnostics cannot prove identification, but they can reveal gross violations and guide model refinement.

10 Discussion and future directions

We have sketched a framework for using minimal counterfactual pairs and activation steering to define and estimate representation-level causal effects of concepts embedded in complex treatments. Several directions appear particularly promising.

First, an empirical program in political science and economics could build curated minimal-pair datasets targeting specific concepts (respect, identity appeals, fairness in bureaucratic replies) and run Stage 1 diagnostics across languages and domains. This would test whether concept-tangent subspaces exist and are robust, and whether steering-based synthetic treatments behave as intended.

Second, formal identification results could characterize conditions under which concept-tangent subspaces approximate structural concept directions, and under which representation-level ignorability is plausible. This would likely require assumptions about encoder training and about how human editing policies map to representation changes.

Third, multi-modal extensions would apply the same logic to image or audio treatments, where pretrained encoders are standard. Panel extensions would use concept coordinates over time in difference-in-differences or synthetic control designs.

Finally, normative and fairness considerations are central. If the encoder encodes biased viewpoints—e.g. associating politeness with certain social groups—then concept coordinates may inadvertently reflect such biases. Steering-based interventions may have heterogeneous effects across subgroups. Connecting concept-tangent causal inference with fairness-aware causal methods is an important avenue for future work.

11 Conclusion

Estimating the effect of "one concept" in a complex treatment, while holding everything else fixed, is an appealing but ambitious goal. In realistic settings, we cannot hope to

achieve perfect isolation. What we can do is make explicit how we use models and data to approximate this goal, and when the approximation itself fails.

We proposed to treat a pretrained encoder as a measurement device, to estimate concept-tangent subspaces from minimal pairs, to subject those subspaces to diagnostics that can explicitly reject the existence of a clean low-dimensional approximation, to constrain activation-steering maps to act within accepted subspaces, and to use the resulting concept coordinates as treatments in a double machine learning framework. The resulting estimand is a local, representation-dependent linear effect of movement in the concept coordinate, conditional on residual representation and covariates.

This is not the last word on causal inference with structured treatments, but it is a concrete way to tie together representation geometry, intervention design, diagnostics, and econometric estimation in a single framework that can be analyzed, critiqued, and improved.

References

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010. doi: 10.1198/jasa.2009.ap08746.

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

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):eabg2652, 2022. doi: 10.1126/sciadv.abg2652.

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.

Florian F. Gunsilius. A primer on optimal transport for causal inference with observational data. *arXiv preprint*, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Sheng Li and Yun Fu. Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 929–939, 2017.

Sonali Parbhoo, Stefan Bauer, and Patrick Schwab. NCoRE: Neural counterfactual representation learning for combinations of treatments. *arXiv preprint*, 2021.

Vikas Ramachandra. Deep synthetic controls: Penalized, representation-learned, sparsity-aware counterfactual estimation. *Authorea Preprints*, 2025. doi: 10.22541/au.176072431.11742213.

Pau Rodriguez, Michal Klein, Eleonora Gualdoni, Valentino Maiorca, Arno Blaas, Luca Zappella, Marco Cuturi, and Xavier Suau. LinEAS: End-to-end learning of activation steering with a distributional loss. *NeurIPS*, 2025. To appear; arXiv:2503.10679.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint*, 2023.

Victor Veitch, Dhanya Sridhar, and David M. Blei. Adapting text embeddings for causal inference. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Hao Wang, Zhichao Chen, Jiajun Fan, Haoxuan Li, Tianqiao Liu, Weiming Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. Optimal transport for treatment effect estimation. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Liuyi Yao, Yaliang Li, Sheng Li, Mengdi Huai, Jing Gao, and Aidong Zhang. SCI: Subspace learning based counterfactual inference for individual treatment effect estimation. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 3583–3587, 2021. doi: 10.1145/3459637.3482175.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint*, 2023.