EXPLORATORY CAUSAL INFERENCE IN SAENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized Controlled Trials are one of the pillars of science; nevertheless, they rely on hand-crafted hypotheses and expensive analysis. Such constraints prevent causal effect estimation at scale, potentially anchoring on popular yet incomplete hypotheses. We propose to discover the unknown effects of a treatment directly from data. For this, we turn unstructured data from a trial into meaningful representations via pretrained foundation models and interpret them via a sparse autoencoder. However, discovering significant causal effects at the neural level is not trivial due to multiple-testing issues and effects entanglement. To address these challenges, we introduce *Neural Effect Search*, a novel recursive procedure solving both issues by progressive stratification. After assessing the robustness of our algorithm on semi-synthetic experiments, we showcase, in the context of experimental ecology, the first successful unsupervised causal effect identification on a real-world scientific trial.

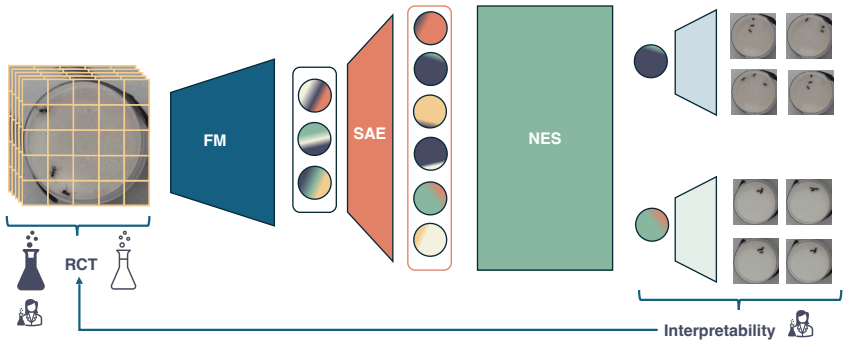

Figure 1: Pipeline for Exploratory Causal Inference: (i) Collect data from a Randomized Controlled Trial, (ii) Extract representations via a *Foundation Model* and *Sparse Autoencoder*, (iii) Apply *Neural Effect Search*, and (iv) domain experts interpret the causal findings.

1 INTRODUCTION

In science, data has been historically collected to answer specific questions (Popper, 2005). In this *rational* view, scientists formulate a hypothesis, often as a causal association, and collect data to falsify it. For example, an experimental ecologist may suspect that exposure to some substance may affect how ants behave, or more in general, "a *treatment* T has a causal effect on an *outcome* Y". They then perform a controlled experiment, administering T or a placebo to a number of individuals and check whether there is a significant difference in the correlation between the treatment assignment and the outcome. While this paradigm has dominated science for centuries, modern science started embracing the creation of *atlases*: vast, comprehensive maps of natural phenomena, collected for general purposes. Today, we have planetary-scale maps of life genomes (Chikhi et al., 2024), sequencing of 33 different types of cancer (Weinstein et al., 2013), imaging of cells under thousands of perturbations (Sypetkowski et al., 2023) to name a few. Different than the classical paradigm, these datasets call for an *empiricist* view, starting with exploratory data-driven investigations. The new challenge is that the immense size of these datasets prohibits scientists from just "looking at the data and finding out what is interesting". Even beyond atlases, consider the specific example of experimental ecology, where fine-grained social interactions between many individuals are critical to understanding the spread of disease (Finn et al., 2019). Clearly, this can

be dramatically accelerated with computer vision, using the predictions of a model as input for causal inference pipelines (Cadei et al., 2025). Still, scientists need to decide what to annotate a priori before they can meaningfully look at and understand the data. This introduces a biasing effect, known as the "Matthew effect" (Merton, 1968) or informally as "rich-get-richer": scientists are biased by prior successful investigations, e.g., behaviors that they have already studied.

In this paper, we characterize differences and synergies between the classical *rationalist* view and the emerging *empiricist* one and propose a method to identify statistically interesting outcomes in exploratory experiments, formally grounding it with the language of statistical causality, see Figure 1. We formulate this problem as analyzing a randomized controlled trial, where one or multiple treatments are administered randomly and the effect is measured indirectly, e.g., via imaging or other raw observations. Instead of scientists formulating only a priori hypotheses on the effect, label some data, and train a model to extend labels to the whole dataset (i.e., the rationalist view (Cadei et al., 2024; 2025)), we propose to train sparse autoencoders (SAEs) on the representation of foundation models in a purely empiricist view. With these, we identify several statistically significant differences across the treated and control groups with proper corrections, and present them to the scientists for interpretation. The main challenge is that, if the SAE is not *perfectly* disentangled (Elhage et al., 2022), any neuron minimally entangled with the effect can be found as statistically significant by vanilla statistical tests, which makes the interpretation very difficult. Instead, we propose a novel recursive stratification technique to iteratively correct the correlation between entangled neurons one effect after the other.

Looking at the data before formulating a hypothesis, we overcome the Matthew effect, enriching the rationalist view in a data-driven way. We propose to work with pretrained foundation models, training SAEs directly on the target experimental data. This is important because pretrained foundation models can be biased as well, which is problematic for drawing scientific conclusions (Cadei et al., 2024). Instead of committing to a single hypothesis, our approach is to look for multiple hypotheses with proper statistical corrections in a semantically expressive latent space. While some may be due to model biases or finite sample spurious correlations, scientists can judge and interpret them a posteriori. This is in stark contrast with existing approaches in causality like "causal feature learning" (Chalupka et al., 2017), which only commits to a single hypothesis on pixel correlations. Our contributions are:

- Using the formalism of statistical causality, we theoretically differentiate the problems of rationalist and empiricist approaches to causal inference, highlighting their different strengths.
- We propose a purely empiricist methodology building on foundation models and sparse autoencoders. We characterize the statistical challenges in multiple hypothesis testing to discover treatment effect with neural representations in our *paradox of exploratory causal inference*. Then, we propose an iterative hypothesis testing procedure that avoids such challenges.
- We showcase in both semi-synthetic (real images but synthetic causal relations) and a real-world trial in experimental ecology that our approach is capable of disentangling and discovering the treatment effect in an experiment. To the best of our knowledge, this is the first successful application of sparse autoencoders to causal inference, which we also test in a real-world scientific dataset.

2 Treatment Effect Estimation in Randomized Controlled Trials

Notation. *In the paper, we refer to random variables as capital letters and their realizations as lowercase letters. Matrices are referred to as upper-case, boldface letters.*

Causal Inference. The central goal of causal inference is to quantify the effects that an intervention on a *treatment* variable has on other variables (often called *effect* or *outcome* variables), see Figure 2 (left). For simplicity, we consider binary treatments $T = \{0, 1\}$ (e.g., taking a drug or a placebo), and an outcome variable $Y \in \{0, 1\}^r$ (e.g., whether the conditions of a sick patient improve). While continuous extensions would be interesting, we focus on discrete outcomes since continuous concepts in SAEs are less well understood. Our goal is to estimate the *Average Treatment Effect* (ATE):

$$\tau = \mathbb{E}[Y(T = 1) - Y(T = 0)], \tag{1}$$

where $Y(T = 1)$ (or $Y(1)$ for short) and $Y(0)$ denote the potential outcomes under treatment and control (Rubin, 1974), equivalently $Y|do(T) = 1$ and $Y|do(T) = 0$ according to Pearl's

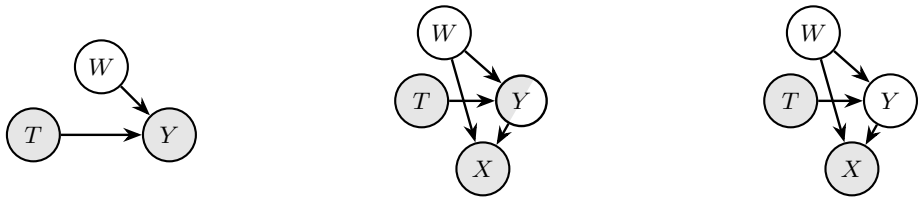

Figure 2: Exemplary graphical models for randomized controlled trials (i.e., no edge from W to T). In **Causal Inference (left)**, both T and Y are observed, and W does not influence T as we are assuming a randomized controlled trial. In **Prediction-Powered Causal Inference (center)**, Y is not observed directly but is known and can be partially labeled. The missing Y is predicted by a neural network from high-dimensional X that is trained either on the same trials if labels are available (Cadei et al., 2024) or on other trials with the same label space (Cadei et al., 2025). In **Exploratory Causal Inference (right)**, Y is unknown and unobserved and is discovered by a neural network from high-dimensional X in a purely data-driven way.

do-calculus (Pearl, 2009). This is challenging to estimate, because, in practice, only one of the two can be observed (fundamental problem of causal inference (Holland, 1986)): for any individual, we can only observe the outcome of whatever treatment they received and not the other. This problem is mitigated in the sciences by performing, whenever possible, a *Randomized Controlled Trial* (RCT). By randomly assigning the treatment, i.e., T has no causes, this prevents spurious correlation between the treatment and any other cause $W \in \mathbb{R}^q$ of the outcome (no confounders), allowing to statistically identify the ATE with the associational difference:

$$\tau = \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0], \tag{2}$$

under standard causal assumptions (Rubin, 1986) of consistency (observing $T = t$, then $Y = Y(t)$), and no interference across individuals (i.e., all individuals are independent samples from the population, and the treatment assignment to the individual i does not affect individual j). It follows that the difference between the treated and control groups' sample means is already an unbiased estimator of the ATE. Nonetheless, more sophisticated estimators such as Augmented Inverse Propensity Weighting (AIPW (Robins et al., 1994)) can achieve lower variance and thus greater efficiency.

Prediction-Powered Causal Inference and the *rationalist approach*. Assume that Y is not observed directly. Instead, we observe high-dimensional measurements $X \in \mathcal{X} \subseteq \mathbb{R}^p$ of the system, capturing the affected outcome information, i.e., $H(Y|X) = 0$, mixed with the other attributes of the individual $W \in \mathbb{R}^q$. Prior work by Cadei et al. (2024; 2025) showed how to train a model on partially labeled data or similar experiments to predict factual outcomes $\hat{Y}$ from X that approximate Y and then use them for causal inference. For simplicity, we assume that T is not directly visible in X, a common practice in double-blind randomized trials (e.g., neither the patient nor the doctor analyzing the results knows which treatment was assigned). A summary can be seen in Figure 2 (center). To simplify the notation, we ignore that some covariates W may only influence X and not Y. If such covariates exist, we group them into W and assume the causal mechanism from W to Y is invariant to those. For example, in the trial by Cadei et al. (2024), ants are treated with an invisible substance, which affects their grooming behaviors. Ecologists do not record the behaviors directly but rather take videos of the ant interactions, which they then analyze.

Exploratory Causal Inference and the *empiricist approach*. The rationalist view requires knowing what the treatment will affect *a priori*, which is also prone to the Matthew effect (Merton, 1968) in exploratory experiments (hypotheses are often informed by the outcome of prior successful trials). In this paper, we consider the setting where experiments are *exploratory*, which we informally model as the scientists having no *a priori* knowledge of what Y may be. This is shown in Figure 2 (right), with Y being unobserved and unknown (only measured through X). We remark that this problem is related to causal abstraction (Rubenstein et al., 2017). In principle, one may consider the pixels themselves as influenced by the treatment. We instead consider the ground truth Y to be the *coarsest* possible abstraction of the effect of T. In other words, we have that $T \perp\!\!\!\perp W|Y$ and the mutual information $I(Y, X)$ is as small as possible (Achille and Soatto, 2018; Fumero et al., 2023). With a slight abuse of notation, we do not need to assume that such Y exists, so r can be zero if the treatment has no effect at all. Our goal is to propose candidate effects Y to the scientists in a purely data-driven way, discovering significant statistics that differentiate the treated and control populations. It is important to remark that we do not interpret these statistics as

necessarily scientifically relevant. The reason is that, when working with high-dimensional data, there can be many correlations. Our approach is to identify *all* significant statistics and leave the interpretation to the domain experts. The empiricist view should not replace the rationalist one, but enrich it with additional data-driven hypotheses.

3 Exploratory Causal Inference via Neural Representations

To detect treatment effects when only high–dimensional measurements X are available, we turn these raw observations into analyzable measurements. We first pass samples x through a pretrained foundation model (FM) (Bommasani et al., 2022), obtaining representations $h = \phi(x) \in \mathbb{R}^d$ whose geometry captures semantically meaningful regularities (Amir et al., 2022; Valeriani et al., 2023). Throughout, we assume the FM is *sufficient for the outcome information* (Achille and Soatto, 2018) (i.e., $I(X, Y) = I\big(\phi(X), Y\big)$,) so working in h preserves exactly the information about the (unknown) outcome factors Y that is present in the raw data. Under sufficiency, any arm difference that exists in X is detectable in representation space, making h a principled surrogate for measurement.

From FM features to a measurement dictionary. While FM features are semantically structured, individual coordinates in h need not align with human–readable concepts (Bricken et al., 2023). We therefore *reparameterize* the representation into a sparse, interpretable measurement dictionary using a sparse autoencoder (SAE) (Bricken et al., 2023; Huben et al., 2024). Intuitively, the SAE expresses each h as a sparse linear combination of atoms that behave like measurable channels; sparsity biases solutions toward localized, approximately monosemantic features that scientists can inspect a posteriori. Given foundation model's features $h \in \mathbb{R}^d$, the SAE computes a high–dimensional but *sparse* code $z \in \mathbb{R}^d$ and reconstructs h linearly:

$$z \;=\; f(h) \;=\; g\big(\mathbf{E}^\top h + b_e\big), \qquad \hat{h} \;=\; \mathbf{D}z, \tag{3}$$

where $\mathbf{E}, \mathbf{D} \in \mathbb{R}^{d\times m}$ are respectively the encoder, and decoder linear maps, $b_e \in \mathbb{R}^m$ is the encoder learnable bias, and $g : \mathbb{R}^m \to \mathbb{R}^m$ is the encoder nonlinearity (Bricken et al., 2023). Training minimizes a reconstruction loss with a sparsity: $\min_{D,z\geq 0} \; \mathbb{E}\big[\|h - \mathbf{D}z\|_2^2\big] \;+\; \lambda\mathcal{S}(z)$. With $\mathbf{D} = [d_1, \ldots, d_m]$, each input is summarized as $\hat{h} \approx \sum_j z_j d_j$, where $\ell_0(z) \ll d$ (Bricken et al., 2023). This turns the FM representation into a dictionary of measurable channels: each coordinate z_j serves as a putative detector of a simple attribute, with some inevitable leakage (Huben et al., 2024).

Monosemanticity, leakage, and entanglement. In exploratory experiments, we would like each SAE code coordinate to behave like a single, human–readable measurement channel for a simple outcome factor. When this happens, a scientist can read off "what changed" from the few activated codes. In practice, however, codes often *leak* across factors: one neuron can respond weakly to several distinct attributes, creating *entanglement* (Locatello et al., 2019). We need a minimal language to talk about (i) the direction in code space associated with a factor and (ii) how widely those directions spill across neurons. Let $Z \in \mathbb{R}^m$ be SAE codes and $Y = (Y_1, \ldots, Y_r)$ the (unknown) binary outcome factors with $m \gg r$. To define the leakage set and index, we first define the *code–mean map* and the *effect vector* of factor Y_k respectively as:

$$\mu(y) \;:=\; \mathbb{E}[\,Z \mid Y = y\,] \in \mathbb{R}^m, \qquad v_k \;:=\; \mu(Y_k{=}1) - \mu(Y_k{=}0) \in \mathbb{R}^m, \tag{4}$$

Definition 3.1 (Leakage set and index)**.** *Fix a threshold $\varepsilon > 0$. We say neuron j is* activated *by factor Y_k if $|(v_k)_j| \geq \varepsilon$, and define the* leakage set *and* leakage index*, respectively, as*

$$\mathcal{A}_\varepsilon \;=\; \bigcup_{k=1}^{r} \{\, j : |(v_k)_j| \geq \varepsilon \,\}, \qquad \rho_\varepsilon \;:=\; \frac{|\mathcal{A}_\varepsilon|}{m}. \tag{5}$$

When the v_k are *sparse* and largely *disjoint* across coordinates, each factor "lights up" only a few neurons, and different factors use different neurons. Large ρ_ε indicates that many neurons respond to multiple factors (high entanglement), ruling out a monosemantic regime, whereas monosemanticity with respect to Y implies $\rho_\varepsilon \approx \frac{r}{m}$.

Codes as statistical measurement channels. Under FM sufficiency and an (approximately) monosemantic SAE, it becomes natural to pose causal questions at the level of individual codes. If the true affected outcomes Y are perfectly localized in disjoint subsets of coordinates of Z, then one can test each coordinate for a treatment–control mean shift using a two–sample t–test, applying

Bonferroni adjustment (Bonferroni, 1936) to control the family–wise error rate at α regardless of the number of tests m. This provides an idealized measurement interface: we can scan Z for treatment–responsive channels and later interpret significant coordinates via the dictionary atoms d_j.

A paradox: multiplicity meets entanglement. The above picture breaks down when leakage occurs, as any neuron entangled with the true affected outcome will eventually be identified as significantly activated. Intuitively, entangled neurons that are primarily assigned to other concepts still activate differently depending on Y, so with more powerful tests (larger sample sizes or strong causal effects), they would be deemed significant. Thereafter, classical multiplicity correction does not rescue interpretability here, leading to the paradox of Exploratory Causal Inference:

Paradox of Exploratory Causal Inference

As the sample size n or the effect magnitude τ grows, multiple testing, even with Bonferroni adjustment, selects all ε leakage neurons from Y in the SAE as independent and statistically significant effects.

We formalize these two phenomena below. Let τ_j denote the treatment effect on code j.

Theorem 3.1 (Significance level collapse with sample size)**.** *Suppose at least $\rho_\varepsilon m$ neurons have nonzero effect $|\tau_j| \geq \varepsilon > 0$. Via multiple testing, regardless of the Bonferroni adjustment,*

$$\Pr\Big[\{\textit{all } j \in \mathcal{A}_\varepsilon \textit{ are rejected}\}\Big] \;\to\; 1 \quad \textit{as } n \to \infty,$$

and the number of rejections converges to $\rho_\varepsilon m$ in probability.

Proof sketch. For each j, the t–statistic is asymptotically normal with noncentrality $\lambda_j = \sqrt{n}\,\tau_j/\sigma$. The Bonferroni cutoff, which determines the significance of τ_j, grows like $\sqrt{2\log m}$; this cutoff is dominated by the growth in expectation of τ_j ($\sqrt{n}$ as $n \to \infty$). Hence, any j with $\tau_j \neq 0$ is eventually rejected. Without Bonferroni correction, the significance cutoff is constant. □

Theorem 3.2 (Significance collapse with effect magnitude)**.** *Fix $n < \infty$ and let $\tau_j(s) = s\,\gamma_j$ with $s > 0$. Via multiple testing, regardless of the Bonferroni adjustment,*

$$\Pr\Big[\{\textit{all } j \in \mathcal{A}_\varepsilon \textit{ are rejected}\}\Big] \;\to\; 1 \quad \textit{as } s \to \infty,$$

and the number of rejections converges to $\rho_\varepsilon m$ in probability.

Proof sketch. The noncentrality $\lambda_j(s) = \sqrt{n}\,s\gamma_j/\sigma$ grows linearly in s, while the cutoff, even with Bonferroni correction, is fixed for fixed m, n; every $\gamma_j \neq 0$ is eventually rejected. □

Numerical illustration. Let $T \sim \text{Bernoulli}(1/2)$, $Y \mid T = t \sim \mathcal{N}(\tau t, 1)$ (single effect), and $Z = [Z_A, Z_B] \in \mathbb{R}^m$ where $Z_A = Y$ (a "true" channel) and $Z_B \mid Y = y \sim \mathcal{N}\big(0.01\,y\,\mathbf{1}_m,\, I_m\big)$ (entangled channels). As shown in Figure 3 for 10 seeds, increasing either n or τ leads Bonferroni to flag essentially all weakly entangled Z_B coordinates as significant, despite their negligible semantic relevance. This motivates the disentangling, stratified testing procedure introduced next (Section 4).

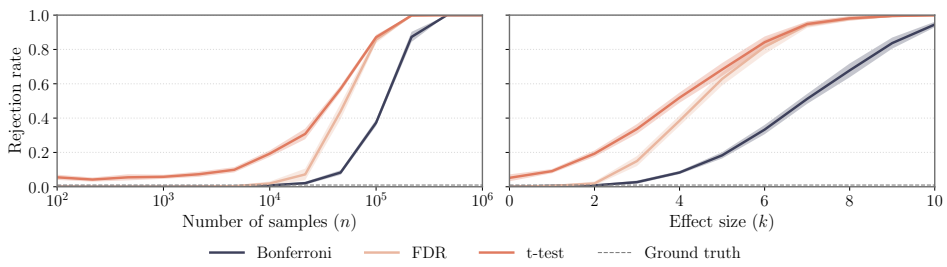

Figure 3: **The Paradox of Exploratory Causal Inference**: Increasing the power of the test (n and τ), the effect on any entangled code becomes significant, regardless of the interpretation.

4 NEURAL EFFECT SEARCH

To mitigate the multi-test significance collapse with entangled representation, we propose a novel causally principled algorithm that disentangles the leaked effects by recursive stratification.

Algorithm 1 Neural Effect Search (NES)

function NEURALEFFECTSEARCH($T, Z, \alpha, \mathtt{S} = \varnothing$)
  $m \leftarrow \#\{\, j : j \notin \mathtt{S} \,\}$ ▷ number of hypotheses to test
  for each neuron $j \notin \mathtt{S}$ **do**
    $(\hat{\tau}_j, p_j) \leftarrow$ NEURALEFFECTTEST($T, Z, j, \mathtt{S}$) ▷ p_j tests $H_0 : \tau_j = 0$
  end for
  $\mathtt{R} \leftarrow \{\, j \notin \mathtt{S} : p_j < \alpha/m \,\}$, ordered by $|\hat{\tau}_j|$ (desc) ▷ Bonferroni adjustment
  if $\mathtt{R} = \varnothing$ **then**
    return $\mathtt{S}$
  else
    return NEURALEFFECTSEARCH($T, Z, \alpha,\ \mathtt{S} \cup \mathtt{R}_1$)
  end if
end function

NEURALEFFECTTEST (Algorithm 1) is a procedure for multi-hypothesis testing on all the neurons j, by arm-wise residual stratification over the already retrieved effects $\mathtt{S}$. See full description in Appendix B. The key idea is that if we test all neurons simultaneously, Bonferroni cannot distinguish whether a neuron carries its own causal effect or merely leaks information about another concept. By contrast, NES first recovers the most prominent effect, then *conditions* on it in subsequent tests. Stratification ensures that once the leading effect is controlled for, spurious correlations induced by leakage vanish, operating over the arm-wise residuals guarantees no bias leakage from other outcome causes W, being Y a collider between W and T.

Theorem 4.1 (Consistency of Neural Effect Search)**.** *Suppose the SAE codes $Z \in \mathbb{R}^m$ contain exactly r true causal effect directions $\{v_1, \ldots, v_r\}$ for the outcome concepts $Y_1, \ldots, Y_r$, possibly entangled across coordinates. Then, as $n \to \infty$, the output of NES satisfies*

$$\Pr\big(S_{\textit{final}} = \{j_1, \ldots, j_r\}\big) \ \to \ 1,$$

where each j_ℓ is a coordinate aligned with a distinct causal effect vector v_ℓ.

Proof Sketch. At the first iteration, entanglement makes many neurons appear significant, but the one most aligned with some v_k maximizes $|\hat{\tau}_j|$ in expectation and is consistently selected under Bonferroni as $n \to \infty$. Residual stratification then regresses out this effect arm by arm, so that (i) leakage of v_k into other coordinates vanishes in expectation, and (ii) collider bias from other causes W is blocked. Consequently, the remaining test statistics are centered at zero. Inductively, each round peels away one true direction until all r are recovered, at which point no further neurons show nonzero effect and the recursion halts. Thus $\Pr(\mathtt{S}_{\text{final}} = \{j_1, \ldots, j_r\}) \to 1$ and $\mathbb{E}[|\mathtt{S}_{\text{final}}|] \to r$. □

Discussion. NES recovers the r effect concepts in probability and terminates, in sharp contrast with the paradox described earlier. While standard multi-hypothesis tests collapse with increasing power, i.e., n and τ, proposing *all* entangled neurons with Y as significant effects, NES avoids this pitfall by recursively stratifying. Each iteration removes the spurious signal caused by leakage and collider bias, so that only the direct causal direction remains detectable. In this sense, NES does not merely test for effects: it *disentangles* the representation, peeling away one true causal factor at a time until the entire effect subspace is recovered. Thus, NES can be interpreted both as a multiple-testing correction method robust to entanglement and as a principled effect disentanglement algorithm.

5 RELATED WORKS

Interpretable Heterogeneous Treatment Effect Estimation. A closely related line of work is the *empirical* discovery of treatment effect heterogeneity across covariates W. Methods such as causal trees, forests, and decision rules ensembles (Athey and Imbens, 2016; Athey et al., 2019; Bargagli-Stoffi et al., 2020) identify subpopulations with different responses, recognizing that pointwise es-

timation of the Conditional Average Treatment Effect (CATE) is almost impossible to test, and still difficult and risky to interpret. Since W is lower-dimensional, interpretability of these partitions or rules is crucial, and the field has developed around making this empirical exploration scientifically meaningful. Our work is analogous in spirit: instead of asking *who* is affected (heterogeneity over W), we ask *what* is affected (discovering affected outcomes Y) when the outcome space itself is high-dimensional and initially unknown.

Causal abstractions and representations. In the line of work of causal abstractions, Visual Causal Feature Learning (VCFL, Chalupka et al., 2014) was introduced to discover interventions in data rather than outcomes. In scientific trials, however, treatments are fixed by design, and the challenge is to recover their effects from complex outcome measurements. Causal Feature Learning (CFL, Chalupka et al., 2017) extends this to outcome clustering by grouping micro- to macro-variables using equivalence classes of $P(X \mid do(T))$. This requires density estimation in high-dimensional spaces, which is generally infeasible. While clustering other metrics may be possible, causal feature learning commits to a single grouping rule, while we find all statistically significant ones. Another line of work tackles the discovery of causal variables from high-dimensional observations (Schölkopf et al., 2021). Closest in spirit to our setting are interventional approaches (Varici et al., 2023; 2024; Zhang et al., 2023; Yao et al., 2025), which, even with all the necessary extra assumptions, would only offer identification results for the experimental settings W and not the outcome (i.e., the component invariant to the intervention (Yao et al., 2025)). Therefore, they can not be applied to exploratory causal inference because they cannot discover outcome variables.

Scientific discovery via SAEs. A related line of work uses SAEs to decompose *polysemantic* hidden representations in foundation models into more *monosemantic* units that align with single concepts (Bricken et al., 2023; Templeton et al., 2024; Huben et al., 2024; Papadimitriou et al., 2025). Although SAEs were initially proposed as an interpretability tool (Bricken et al., 2023), a growing body of negative results, including spurious interpretability on random networks (Heap et al., 2025), failures to isolate atomic concepts (Leask et al., 2025; Chanin et al., 2025), and limited downstream benefits (Wu et al., 2025), casts doubt on whether SAE features faithfully reflect underlying mechanisms rather than post-hoc artifacts. Despite these interpretability concerns, recent work shows that SAEs can still be useful for generating scientific hypotheses from high-dimensional data (Peng et al., 2025). For example, *HypotheSAEs* (Movva et al., 2025) leverage SAEs to surface human-understandable patterns correlated with target outcomes (e.g., engagement levels), which researchers can then treat as hypotheses for follow-up study. Our setting is related but distinct: whereas these approaches focus on correlations and do not provide statistical procedures to test the significance of the unsupervised discoveries, we target *causal* effects and develop inference to assess which high-dimensional outcomes Y are affected, offering principled support for exploratory causal claims.

6 Experiments

We validate our analyses (significance collapse paradox, and NES consistency) in two complementary settings: a semi-synthetic benchmark where ground-truth causal effects are known, and a real-world randomized trial from experimental ecology.

6.1 Semi-Synthetic Benchmark

We simulated a family of RCTs $\{T_i, W_i, Y_i\}_{i=1}^n$, relating both the individual covariates and outcomes one-to-one with specific attributes in the CelebA (Liu et al., 2018) dataset, e.g., `wearing_hat` and `eyeglasses`, and then assigned a random image X_i from the dataset perfectly matching such attributes. Given the corresponding random sample $\{T_i, X_i\}_{i=1}^n$ we (i) trained a SAE over the image representations encoded by SigLIP (Zhai et al., 2023), and (ii) tested NES for effect discovery against vanilla statistical tests (t-test, FDR (Schweder and Spjøtvoll, 1982), Bonferroni) and top-k effects selection. For quantitative evaluation, we first assessed SAE monosemanticity with respect to the considered CelebA attributes (see Figure 7), and extracted the ground truth neurons referring to Y. Then, for each effect discovery, we computed Recall, Precision, and Intersection over Union (IoU) with respect to them. Full details about the data generating procedure, training, and evaluation with additional assessment on interpretability and SAE entanglement, together with extensive ablations on method variants, i.e., estimator and test, and hedge cases, i.e., no-effect, are reported in Appendix C-D. The main results ($r = 2$) are summarized in Figure 4.

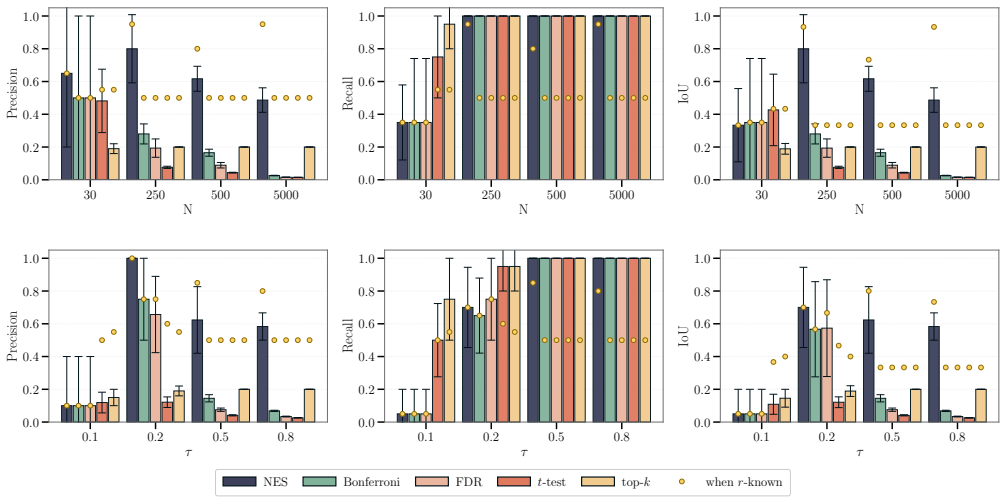

Figure 4: **Semi-synthetic benchmark.** Precision, Recall, and IoU of different testing procedures across sample size N (top) and effect size τ (bottom). NES consistently achieves the best trade-off, avoiding the significance collapse of standard corrections.

Results. Increasing the power of the tests (increasing the sample size n or effect magnitude τ), all the methods eventually retrieve the true significant effects, i.e., Recall $\to 1$. However, while all the baselines drop the Precision and corresponding IoU (Paradox of Exploratory Causal Inference), NES is the only method that mitigates such entanglement biases. As expected, the Paradox doesn't emerge with very small sample ($n = 30$) and effect regime ($\tau = 0.1$), and more explorative approaches, as vanilla t-test or top-k selection, could be preferred, at the price of potentially more false significant hypotheses, i.e., Precision $\ll 1$. With a yellow dot, we report the performance of each method assuming the number of affected outcomes r is known. NES still manages to find both effects most of the time. Instead, all the baseline methods fail to reach Precision and Recall above 0.5: they succeed in retrieving the most significant effect (equivalent to the first step in NES), but then get confounded by the entanglement and miss the second one. While this is clearly a toy experiment, this is undesirable. For example, if in real trials there are multiple effects with different magnitudes (e.g., the positive effect of a drug on the health metric of interest and rare side effects) the leakage of strong effects may prevail over the weaker ones, which would then be missed.

6.2 Real-World Randomized Trial from Experimental Ecology

ISTAnt (Cadei et al., 2024) is an ecological experiment where ants from the same colony are randomly exposed to a treatment or a control substance and continuously filmed in triplets in a closed environment to study the concept of Social Immunity. The biologists are interested in identifying which latent behaviors are significantly affected by treatment. According to previous analysis, we first encoded each frame in the trial with DINOv2 (Oquab et al., 2023), and then we trained a SAE directly on the trial data. NES is then applied without Bonferroni adjustment due to the small sample size ($n = 44$ videos) to discover treatment-sensitive codes, and only two neurons are returned.

Results. Figure 5 qualitatively summarizes the interpretations of such neurons, visualizing their corresponding most and least activated clips in the dataset. In agreement with the previous analysis on the dataset (Cadei et al., 2024; 2025), the first neuron retrieved (`code 394`) represents the grooming event, already measured as significantly affected by the treatment in any previous rationalist approach to the experiment, i.e., actually manually annotating and testing for it. Quantitatively, such a neuron is exactly the most predictive neuron for grooming event (F1-*score*=0.398) out of all the 4608 SAE's codes, confirming the consistent results of our pipeline. We remark that our focus is on the identification of the effect as statistically significant. The imperfect F1-*score* means that one should not compute treatment effects directly on the neural activation, e.g., without further labeling. The second neuron activated (`code 550`) represents the palette background (top right black color mark in the top left position in the first 4 batches of videos), which strongly correlates with the treatment due to the small experiment size (as discussed in the annotation bias by Cadei et al. (2024)).

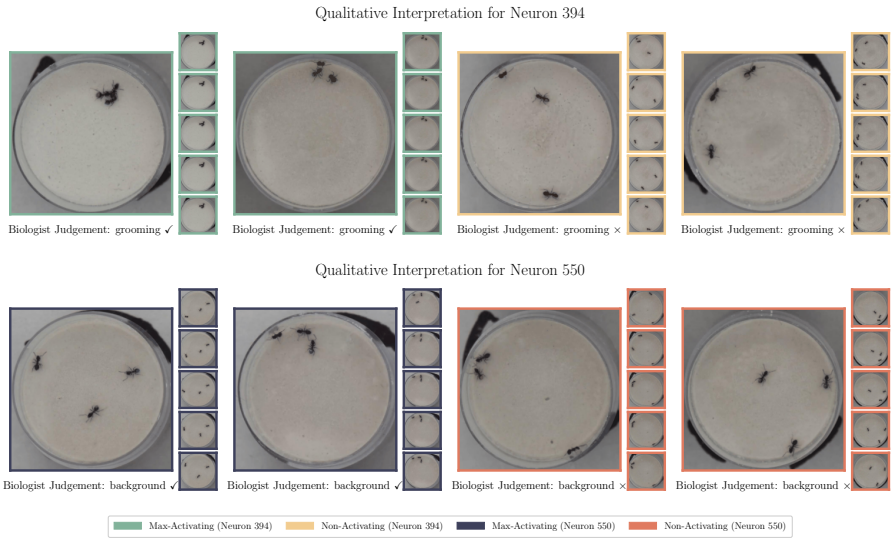

Figure 5: **Exploratory Causal Inference for Experimental Ecology.** Without any knowledge of the behaviors of interest, our procedure retrieves two significant treatment effects, i.e., grooming and background, in agreement with previous literature.

The fact that the model also identifies the effect of the treatment on the background due to the small sample size is a strength of the method: it is a statistically significant signal, and we want to retrieve it *in addition* to the behaviors since it is present in the dataset. Domain experts can select which signal is scientifically relevant and even use this information to improve their experimental settings.

7 Conclusion

In this paper, we have discussed how foundation models and SAEs can address the challenges of exploratory causal inference, serving as learned measurement devices. A key challenge is that SAE neurons may not map one-to-one onto high-level concepts, and even weak or mixed associations propagate the dependency on the treatment. This means that many neurons can be activated, making the interpretation difficult as they do not encode a single concept, and they activate more with larger sample sizes or stronger effects. We address this issue with Neural Effect Search, a statistical hypothesis testing procedure that iteratively controls for the biased dependency between neurons after they have been selected. Our experiments on semi-synthetic and real-world randomized trials are encouraging: our method uncovers both scientifically relevant effects and, when present, interpretable associations like background effects due to finite samples that experts can readily dismiss. Overall, we view this work as a first step toward AI-driven efficiency gains in exploratory data science, where foundation models can "look at massive amounts of data first" and then domain experts can identify which patterns have scientific value.

Our approach has several limitations. First, we assume that the observed variables X adequately capture information of the unknown Y, which can be a strong assumption. At the same time, our method enables more complex measurement processes for X, so a natural extension would be to incorporate multi-modal data, potentially with the effect Y visible in different modalities depending on its realization. It would also be useful to extend our methodology for continuous concepts. The biggest limitation is that we assume foundation models encode concepts linearly and that SAEs can approximately recover them. We believe the first assumption is mild: even if current foundation models are imperfect, future iterations are likely to improve. The second assumption is our strongest, but could be mitigated by advances in identifiability results for SAEs. Promising early work already exists (Cui et al., 2025; Hindupur et al., 2025), but the identifiability theory of SAEs is not currently as well understood as that of causal representations (Yao et al., 2025). In our paper, we took a more empirical and future-looking stance on improvements in SAEs, focusing on finite samples and leveraging pretrained foundation models. Lacking identifiability then means that domain experts can today only use our method "as a rescue system for hypotheses they may have missed", before properly annotating the data and following the rationalist approach. We hope that our work will serve as a practical motivation for future work on identifiability in foundation model representations and SAE.

Ethics Statement

All datasets used in this work are publicly available. In particular, the ISTAnt dataset (Cadei et al., 2024) was annotated and pre-processed by domain experts. While our model is capable of detecting statistically significant signals in randomized trials, the conclusions should not be interpreted as scientifically relevant unless domain experts interpret them. Since we cannot guarantee identifiability, it should only be used as a rescue system for hypotheses that may have been missed before committing to the rationalist approach, which is still necessary.

Reproducibility Statement

Together with the paper, we submitted the code for NES, which can be used on top of any library for SAEs. Standalone code to reproduce all experiments will be released with the final version of the paper. All the datasets we use are publicly available and experiment details are thoroughly detailed in Appendix C.

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors, 2022. URL https://arxiv.org/abs/2112.05814.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. 2019.

Falco J Bargagli-Stoffi, Riccardo Cadei, Kwonsang Lee, and Francesca Dominici. Causal rule ensemble: Interpretable discovery and inference of heterogeneous treatment effects. *arXiv preprint arXiv:2009.09036*, 2020.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze*, 8:3–62, 1936.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec,

Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Riccardo Cadei, Lukas Lindorfer, Sylvia Cremer, Cordelia Schmid, and Francesco Locatello. Smoke and mirrors in causal downstream tasks. *arXiv preprint arXiv:2405.17151*, 2024.

Riccardo Cadei, Ilker Demirel, Piersilvio De Bartolomeis, Lukas Lindorfer, Sylvia Cremer, Cordelia Schmid, and Francesco Locatello. Causal lifting of neural representations: Zero-shot generalization for causal inferences. *arXiv preprint arXiv:2502.06343*, 2025.

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. *arXiv preprint arXiv:1412.2309*, 2014.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Isaac Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders, 2025. URL https://openreview.net/forum?id=LC2KxRwC3n.

Rayan Chikhi, Téo Lemane, Raphaël Loll-Krippleber, Mercè Montoliu-Nerin, Brice Raffestin, Antonio Pedro Camargo, Carson J Miller, Mateus Bernabe Fiamenghi, Daniel Paiva Agustinho, Sina Majidian, et al. Logan: planetary-scale genome assembly surveys life's diversity. *bioRxiv*, pages 2024–07, 2024.

Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the theoretical understanding of identifiable sparse autoencoders and beyond. *arXiv preprint arXiv:2506.15963*, 2025.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Kelly R Finn, Matthew J Silk, Mason A Porter, and Noa Pinter-Wollman. The use of multilayer network analysis in animal behaviour. *Animal behaviour*, 149:7–22, 2019.

Marco Fumero, Florian Wenzel, Luca Zancato, Alessandro Achille, Emanuele Rodolà, Stefano Soatto, Bernhard Schölkopf, and Francesco Locatello. Leveraging sparse and shared feature activations for disentangled representation learning. *Advances in Neural Information Processing Systems*, 36:27682–27698, 2023.

Thomas Heap, Tim Lawson, Lucy Farnik, and Laurence Aitchison. Sparse autoencoders can interpret randomly initialized transformers, 2025. URL https://arxiv.org/abs/2501.17727.

Sai Sumedh R Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Demba Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. *arXiv preprint arXiv:2503.01822*, 2025.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81 (396):945–960, 1986.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Patrick Leask, Bart Bussmann, Michael T Pearce, Joseph Isaac Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9ca9eHNrdH.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations, 2019. URL https://arxiv.org/abs/1811.12359.

Robert K Merton. The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63, 1968.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse autoencoders for hypothesis generation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=4R0pugRyN5.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham Kakade, and Stephanie Gil. Interpreting the linear structure of vision-language model embedding spaces, 2025. URL https://arxiv.org/abs/2504.11695.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Kenny Peng, Rajiv Movva, Jon Kleinberg, Emma Pierson, and Nikhil Garg. Use sparse autoencoders to discover unknown concepts, not to act on known concepts, 2025. URL https://arxiv.org/abs/2506.23845.

Karl Popper. *The logic of scientific discovery*. Routledge, 2005.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

P Rubenstein, S Weichwald, S Bongers, J Mooij, D Janzing, M Grosse-Wentrup, and B Schölkopf. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 808–817, 2017.

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

Donald B Rubin. Comment: Which ifs have causal answers. *Journal of the American statistical association*, 81(396):961–962, 1986.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Tore Schweder and Eil Spjøtvoll. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69(3):493–502, 1982.

Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, et al. Rxrx1: A dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4285–4294, 2023.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models, 2023. URL https://arxiv.org/abs/2302.00294.

Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 2023.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2024.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=K2CckZjNy0.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36:50254–50292, 2023.

Appendix

A PROOFS

A.1 SIGNIFICANCE LEVEL COLLAPSE WITH SAMPLE SIZE (THEOREM 3.1)

Theorem A.1 (Significance level collapse with sample size)**.** *Let $Z \in \mathbb{R}^m$ be SAE codes and τ_j the treatment effect on neuron j. By definition*

$$\mathcal{A}_\varepsilon := \{\, j : |\tau_j| \geq \varepsilon \,\}, \qquad |\mathcal{A}_\varepsilon| = \rho_\varepsilon m. \tag{6}$$

In multiple testing at level α, regardless of Bonferroni correction

$$\Pr\Big(\textit{all } j \in \mathcal{A}_\varepsilon \textit{ are rejected}\Big) \;\to\; 1 \quad \textit{as } n \to \infty, \tag{7}$$

and the number of rejections R_n satisfies

$$R_n \;\to\; \rho_\varepsilon m \quad \textit{in probability}. \tag{8}$$

In words: as the sample size grows, all entangled neurons with the (true) affected outcomes are declared significantly affected by the treatment, regardless of being principally related to other concepts.

Proof. For each neuron j, let $\hat{\tau}_j$ be the estimated treatment effect and t_j its t-statistic. Under standard randomization, we have the asymptotic distribution

$$t_j \;\xrightarrow{d}\; \mathcal{N}(\lambda_j, 1), \qquad \lambda_j = \tfrac{\sqrt{n}}{\sigma}\,\tau_j, \tag{9}$$

where σ^2 is the asymptotic variance of $\hat{\tau}_j$.

Multiple testing with Bonferroni adjustment rejects $H_{0j} : \tau_j = 0$ if $|t_j| > z_{\alpha/(2m)}$, where $z_{\alpha/(2m)}$ is the $(1 - \alpha/(2m))$ quantile of $\mathcal{N}(0, 1)$. As $m \to \infty$, the threshold satisfies

$$z_{\alpha/(2m)} \;\asymp\; \sqrt{2 \log m}. \tag{10}$$

For any $j \in \mathcal{A}_\varepsilon$, we have $\tau_j \neq 0$, hence λ_j diverges at rate $\sqrt{n}$ as $n \to \infty$. Since $\sqrt{n}$ grows faster than $\sqrt{\log m}$, it follows that

$$\Pr\big(|t_j| > z_{\alpha/(2m)}\big) \;\to\; 1. \tag{11}$$

Therefore, for all $j \in \mathcal{A}_\varepsilon$, the null is rejected with probability tending to one, and analogously

$$\Pr\big(|t_j| > z_{\alpha/2}\big) \;\to\; 1. \tag{12}$$

without Bonferroni adjustment. By the union bound,

$$\Pr\Big(\text{all } j \in \mathcal{A}_\varepsilon \text{ are rejected}\Big) \;\to\; 1. \tag{13}$$

Hence, the number of rejections converges in probability to $|\mathcal{A}_\varepsilon| = \rho_\varepsilon m$, proving the claim. □

A.2 SIGNIFICANCE COLLAPSE WITH EFFECT MAGNITUDE (COROLLARY 3.2)

Corollary A.1 (Significance collapse with effect magnitude)**.** *Fix a finite sample size n. Suppose the treatment effects scale as*

$$\tau_j(s) = s\,\gamma_j, \qquad j = 1, \ldots, m, \tag{14}$$

where $s > 0$ is a scaling parameter and γ_j are fixed coefficients. By definition

$$\mathcal{A}_\varepsilon := \{\, j : |\gamma_j| > \frac{\varepsilon}{s} \,\}, \qquad |\mathcal{A}_\varepsilon| = \rho_\varepsilon m. \tag{15}$$

In multiple testing at level α regardless of the Bonferroni correction,

$$\Pr\Big(\textit{all } j \in \mathcal{A}_\varepsilon \textit{ are rejected}\Big) \;\to\; 1 \quad \textit{as } s \to \infty, \tag{16}$$

and the number of rejections R_s satisfies

$$R_s \;\to\; \rho_\varepsilon m. \quad \textit{in probability}. \tag{17}$$

In words: even at a fixed sample size, amplifying the effect magnitude all the entangled neurons with the (true) affected outcomes are declared significantly affected by the treatment, regardless of being principally related to other concepts.

Proof. For neuron j, the noncentrality parameter of the t-statistic under effect scaling s is

$$\lambda_j(s) = \tfrac{\sqrt{n}}{\sigma}\,\tau_j(s) = \tfrac{\sqrt{n}}{\sigma}\,s\gamma_j. \tag{18}$$

If $\gamma_j = 0$, then $\lambda_j(s) = 0$ for all s and the rejection probability remains bounded by α/m.

If $\gamma_j \neq 0$, then $\lambda_j(s) \to \infty$ linearly in s, while the Bonferroni threshold $z_{\alpha/(2m)}$ is fixed (since n, m are fixed). Therefore,

$$\Pr\big(|t_j| > z_{\alpha/(2m)}\big) \;\to\; 1 \quad \text{as } s \to \infty. \tag{19}$$

Analogously, without Bonferroni $\frac{1}{m}$ significance correction. Thus, for every $j \in \mathcal{A}_\varepsilon$, the null is eventually rejected with probability tending to one. By independence of limits,

$$\Pr\Big(\text{all } j \in \mathcal{A}_\varepsilon \text{ are rejected}\Big) \;\to\; 1, \tag{20}$$

and $R_s \to \rho_\varepsilon m$ in probability, completing the proof. □

A.3 Consistency of Neural Effect Search (Theorem 4.1)

Let $T \in \{0,1\}$ be a randomized treatment, $W \in \mathbb{R}^q$ exogenous causes, $Y \in \mathbb{R}^r$ the (unknown) causal outcome factors, and $Z \in \mathbb{R}^m$ SAE codes. Assume the following structural causal model:

$$T \sim \text{Bernoulli}(p), \qquad T \;\perp\!\!\!\perp\; W, \tag{21}$$

$$Y := f_Y(T, W, \eta_Y), \tag{22}$$

$$Z := VY + BW + \varepsilon, \qquad V \in \mathbb{R}^{m\times r},\; B \in \mathbb{R}^{m\times q}, \tag{23}$$

with mutually independent noises (η_Y, ε) and finite second moments. Write the k-th column of V as $v_k \in \mathbb{R}^m$ (the *effect vectors*), so

$$\mathbb{E}[\,Z \mid do(T{=}t)\,] \;=\; V\,\mu_Y(t) + B\,\mu_W, \quad \text{where } \mu_Y(t) = \mathbb{E}[\,Y \mid do(T{=}t)\,],\; \mu_W = \mathbb{E}[W].$$

Define the interventional contrast on codes:

$$\tau^Z \;:=\; \mathbb{E}[\,Z \mid do(T{=}1)\,] - \mathbb{E}[\,Z \mid do(T{=}0)\,] \;=\; V\,\tau^Y, \qquad \tau^Y := \mu_Y(1) - \mu_Y(0). \tag{24}$$

At round ℓ, let $\mathcal{S}_{\ell-1} \subseteq [m]$ be the set of already-selected neurons and let $\mathcal{E}_{\ell-1} = \text{span}\{v_k :$ some selected coordinate aligns with $v_k\}$ be the recovered effect subspace. For a candidate coordinate $j \notin \mathcal{S}_{\ell-1}$, the Neural Effect Test constructs *arm-wise residuals* by regressing $Z_{\cdot j}$ on $Z_{\cdot \mathcal{S}_{\ell-1}}$ *within each treatment arm*:

$$\beta_t \in \arg\min_\beta\; \mathbb{E}\Big[\big(Z_j - \beta^\top Z_{\mathcal{S}_{\ell-1}}\big)^2 \,\Big|\, T{=}t\Big], \qquad R_j \;:=\; Z_j - \beta_T^\top Z_{\mathcal{S}_{\ell-1}}. \tag{25}$$

Neural Effect Test then tests $H_0:\; \tau_j^R = 0$, where

$$\tau_j^R \;:=\; \mathbb{E}[\,R_j \mid do(T{=}1)\,] - \mathbb{E}[\,R_j \mid do(T{=}0)\,]. \tag{26}$$

The test statistic is a post-stratified difference-in-means built on discretized $R_{\mathcal{S}_{\ell-1}}$ (Appendix B); Bonferroni is applied across $j \notin \mathcal{S}_{\ell-1}$ and Neural Effect Search adds the top-$|\hat{\tau}_j|$ rejection, updating $\mathcal{S}_\ell$.

We now show three facts:

i. identification of τ^Z via *do*-calculus (Proposition A.1)

ii. arm-wise residualization removes leakage from already-discovered effects in expectation (Lemma A.1);

iii. stratified diff-in-means over arm-wise residuals is unbiased for τ_j^R and preserves the remaining (undiscovered) causal contrast (Lemma A.2),

and then we use them to inductively show Neural Effect Search's consistency.

Proposition A.1 (do-identification of code-level contrasts)**.** *Under Equations 21–23, for any (measurable) function h of Z,*

$$\mathbb{E}[h(Z) \,|\, do(T{=}t)] \;=\; \mathbb{E}[h(Z) \,|\, T{=}t]\,, \tag{27}$$

and in particular:

$$\tau^Z = \mathbb{E}[Z \mid T{=}1] - \mathbb{E}[Z \mid T{=}0] = V\,\tau^Y. \tag{28}$$

Proof. By randomization $T \perp\!\!\!\perp W$ and exogeneity of noises, the post-intervention distribution $P(Z \mid do(T{=}t))$ equals the observational $P(Z \mid T{=}t)$ (Rule 2 of do-calculus, or the truncated factorization). Thus expectations coincide. Plugging Equation 23 and taking expectations yields Equation 24. □

Lemma A.1 (Arm-wise residualization cancels discovered effects in expectation)**.** *Fix a round with discovered subspace $\mathcal{E}_{\ell-1}$ and suppose Z follows Equation 23 with finite covariance matrices in each arm. Let Π_t be the $L^2(P(\cdot \mid T{=}t))$-projection of Z_j onto $\sigma(Z_{\mathcal{S}_{\ell-1}})$, whose coefficient is β_t in Equation 25. Then*

$$\mathbb{E}[\,R_j \mid do(T{=}t)\,] \;=\; \Big(\mathrm{Id} - \mathsf{P}_t\Big)\,\mathbb{E}[\,Z_j \mid do(T{=}t)\,], \tag{29}$$

where P_t is the linear map induced by $\beta_t^\top Z_{\mathcal{S}_{\ell-1}}$, and hence

$$\tau_j^R \;=\; \Big(\mathrm{Id} - \mathsf{P}_1\Big)\,\mathbb{E}[Z_j \mid do(1)] - \Big(\mathrm{Id} - \mathsf{P}_0\Big)\,\mathbb{E}[Z_j \mid do(0)]. \tag{30}$$

If the selected coordinates span $\mathcal{E}_{\ell-1}$ (i.e., $Z_{\mathcal{S}_{\ell-1}}$ contains $V_{\ell-1}Y$ with $V_{\ell-1}$ a basis of $\mathcal{E}_{\ell-1}$), then the contribution of all $v_k \in \mathcal{E}_{\ell-1}$ cancels in τ_j^R.

Proof. By definition of the arm-wise projection, $\mathbb{E}[\,Z_j - \beta_t^\top Z_{\mathcal{S}_{\ell-1}} \mid T{=}t\,] = \mathbb{E}[Z_j \mid T{=}t] - \beta_t^\top \mathbb{E}[Z_{\mathcal{S}_{\ell-1}} \mid T{=}t]$. By Proposition A.1, replacing $T{=}t$ with $do(T{=}t)$ preserves expectations. Subtract the two arms to get the display. If $Z_{\mathcal{S}_{\ell-1}}$ spans $\mathcal{E}_{\ell-1}$, then $Z_{\mathcal{S}_{\ell-1}}$ contains the mean-shift components $V_{\ell-1}\mu_Y(t)$. The linear projection P_t exactly removes any mean shift lying in that span, so the contribution of already-discovered v_k (those in $\mathcal{E}_{\ell-1}$) is annihilated in τ_j^R. □

Lemma A.2 (Stratified diff-in-means on arm-wise residuals is unbiased)**.** *Let $\mathcal{G}$ be strata formed by deterministic binarizations/quantizations of $R_{\mathcal{S}_{\ell-1}}$ computed via Equation 25. Define the post-stratified estimator*

$$\widehat{\tau}_j^R \;=\; \sum_{g\in\mathcal{G}} w_g\Big(\overline{R}_{j,1g} - \overline{R}_{j,0g}\Big), \qquad w_g \propto n_{1g} + n_{0g}. \tag{31}$$

Under randomization and SUTVA, $\mathbb{E}[\widehat{\tau}_j^R] = \tau_j^R$.

Proof. Because T is randomized, within each g the treated and control samples are iid draws from $P(\cdot \mid G{=}g, do(T{=}1))$ and $P(\cdot \mid G{=}g, do(T{=}0))$ respectively (by Rule 2 of do-calculus, randomization implies $P(\cdot\,|G, do(T{=}t)) \;=\; P(\cdot\,|G, T{=}t)$ even if G is post-treatment, provided G

is a deterministic function of *arm-wise* statistics that do not mix arms). Thus $\mathbb{E}[\overline{R}_{j,tg}] = \mathbb{E}[R_j \mid G{=}g, do(T{=}t)]$ and

$$\mathbb{E}[\widehat{\tau}_j^R] \;=\; \sum_g \omega_g \left(\mathbb{E}[R_j \mid G{=}g, do(1)] - \mathbb{E}[R_j \mid G{=}g, do(0)]\right), \tag{32}$$

with $\omega_g = \Pr(G{=}g)$ in the limit of large samples under $w_g \propto n_{1g} + n_{0g}$. Law of total expectation gives $\sum_g \omega_g \, \mathbb{E}[R_j \mid G{=}g, do(t)] = \mathbb{E}[R_j \mid do(t)]$, proving $\mathbb{E}[\widehat{\tau}_j^R] = \tau_j^R$. □

Proposition A.2 (One-step correctness)**.** *Suppose at round ℓ the discovered subspace equals $\mathcal{E}_{\ell-1}$. For any $j \notin \mathcal{S}_{\ell-1}$,*

$$\tau_j^R \;=\; \sum_{k:\, v_k \notin \mathcal{E}_{\ell-1}} \langle e_j, \widetilde{v}_k \rangle \, \tau_k^Y, \tag{33}$$

where $\widetilde{v}_k$ is the residual (w.r.t. arm-wise projection) of v_k onto $\mathrm{span}\{Z_{\mathcal{S}_{\ell-1}}\}$. In particular, $\tau_j^R = 0$ if and only if Z_j carries no remaining component from any undiscovered v_k. Moreover, $\widehat{\tau}_j^R$ is unbiased for τ_j^R and its t-statistic is asymptotically normal.

Proof. By Lemma A.1, discovered directions vanish from the do-contrast of residuals; only undiscovered v_k contribute, and only via their residual components $\widetilde{v}_k$. Lemma A.2 gives unbiasedness. Standard Lindeberg–Feller CLT for stratified diff-in-means with finite variances yields asymptotic normality of the t-statistic (the Satterthwaite df in Alg. B gives a finite-sample correction). □

Theorem A.2 (Consistency of Neural Effect Search)**.** *Assume Equations 21–equation 23, consistency, no interference, i.e., SUTVA, finite fourth moments, and that at each round the top-$|\widehat{\tau}_j^R|$ rejection is selected with Bonferroni level α/m. Suppose V has full column rank r and every causal direction v_k has at least one coordinate j whose residual component $\widetilde{v}_{k,j} \neq 0$ when earlier effects are removed. Then, as $n \to \infty$, Neural Effect Search selects one new direction per round and stops in r rounds with probability $\to 1$:*

$$\Pr\big(S_{\mathrm{final}} = \{j_1, \ldots, j_r\}\big) \to 1, \qquad \mathbb{E}\big[|S_{\mathrm{final}}|\big] \to r. \tag{34}$$

Proof. (Induction over rounds.) At $\ell = 1$, $\tau^Z = V\tau^Y$ by Prop. A.1. Among all coordinates, at least one j aligned with some nonzero v_k has nonzero τ_j^Z; Bonferroni with $n \to \infty$ rejects it with probability $\to 1$ (noncentrality grows as $\sqrt{n}$). Assume at round ℓ the subspace $\mathcal{E}_{\ell-1}$ of discovered effects is correct. By Prop. A.2, for any $j \notin \mathcal{S}_{\ell-1}$, the target contrast equals $\tau_j^R = \sum_{k:\, v_k \notin \mathcal{E}_{\ell-1}} \langle e_j, \widetilde{v}_k \rangle \tau_k^Y$. By the rank and nondegeneracy assumptions, there exists at least one undiscovered k and one j with $\langle e_j, \widetilde{v}_k \rangle \tau_k^Y \neq 0$. Unbiasedness (Lemma A.2) and asymptotic normality then imply its test rejects with probability $\to 1$ under Bonferroni. Conversely, for any j orthogonal (in residual sense) to all remaining directions, $\tau_j^R = 0$ and the test does not reject with probability $\to 1$. Thus, the selected coordinate introduces a new direction, enlarging $\mathcal{E}_{\ell-1}$. After at most r rounds, all directions are discovered and, by Prop. A.2, $\tau_j^R = 0$ for all remaining j, so no further rejections occur. □

The arm-wise projection in Equation 25 kills two spurious sources at once: (i) it cancels *leakage* from already-discovered causal directions (Lemma A.1); and (ii) by projecting *within arms*, it avoids opening the collider $T \to Y \leftarrow W$ (no cross-arm conditioning on Y-functions), so confounding through W cannot re-enter. Therefore, Neural Effect Search's stratified estimator targets the *residual* do-contrast of Z_j, which equals the undiscovered causal contribution only (Prop. A.2). Unlike plain Bonferroni—which inflates discoveries across all entangled coordinates as n or effect size grows—Neural Effect Search peels one causal direction per round and then *stops*, acting as a principled disentanglement-by-testing procedure.

B ALGORITHM DETAILS

Algorithm 2 Neural Effect Test (NET) with arm-wise residual stratification

function NEURALEFFECTTEST($T,\ Z,\ j,\ \mathrm{S}$)
  // A) Arm-wise residualize only the tested neuron j
  if $\mathrm{S} = \varnothing$ **then**
    set $r_j \leftarrow Z_{\cdot j}$ *(first round: no conditioning)*
  else
    for $t \in \{0, 1\}$ **do**
      regress $Z_{\cdot j}$ on $Z_{\cdot,\,\mathrm{S}}$ using only samples with $T = t$
      for each i with $T_i = t$: $r_{j,i} \leftarrow Z_{ij} - \hat{\beta}_t^{(j)\top} Z_{i,\,\mathrm{S}}$
    end for
  end if

  // B) Stratification from raw Z_{S}
  if $\mathrm{S} = \varnothing$ **then**
    put all units in a single stratum: $\mathcal{G} = \{\text{all}\}$
  else
    compute pooled (ignore T) medians/quantiles of each $Z_{\cdot s}$, $s \in \mathrm{S}$
    assign each unit i to a cell $g(i)$ by binning $Z_{i,\mathrm{S}}$ via those cutpoints
    drop any stratum g with $n_{1g} = 0$ or $n_{0g} = 0$
  end if

  for each stratum $g \in \mathcal{G}$ **do**
    $n_{1g}, n_{0g} \leftarrow$ treated/control counts in g
    $\mu_{1g}, \mu_{0g} \leftarrow$ treated/control means of r_j in g
    $\sigma^2_{1g}, \sigma^2_{0g} \leftarrow$ treated/control variances of r_j in g
    $w_g \leftarrow \dfrac{n_{1g} + n_{0g}}{\sum_h (n_{1h} + n_{0h})}$
  end for
  $\hat{\tau}_j \leftarrow \sum_g w_g \left(\mu_{1g} - \mu_{0g}\right)$
  $V \leftarrow \sum_g w_g^2 \left(\dfrac{\sigma^2_{1g}}{n_{1g}} + \dfrac{\sigma^2_{0g}}{n_{0g}}\right)$
  $t \leftarrow \dfrac{\hat{\tau}_j}{\sqrt{V}}; \qquad \nu \leftarrow \dfrac{V^2}{\sum_{g \in \mathcal{G}} \left(\dfrac{\left(w_g^2\, \sigma^2_{1g}/n_{1g}\right)^2}{\max(n_{1g} - 1, 1)} + \dfrac{\left(w_g^2\, \sigma^2_{0g}/n_{0g}\right)^2}{\max(n_{0g} - 1, 1)} \right)}$ ▷ Satterthwaite df
  $p \leftarrow 2 \cdot \Pr(|T_\nu| \geq |t|)$ ▷ tests $H_0 : \tau_j^R = 0$
  return $(\hat{\tau}_j,\ p)$
end function

The algorithm tests whether neuron j still carries a *residual* causal contrast after accounting for already–discovered effects S. We first compute an *arm-wise* residual $r_j := Z_j - \hat{\beta}_T^{(j)\top} Z_{\mathrm{S}}$, where $\hat{\beta}_t^{(j)}$ is fit using only units with $T = t$. Arm-wise fitting avoids pooled "bad control" on post-treatment codes and cancels leakage from previously found directions as they manifest within each arm.

We then form treatment-agnostic strata $\mathcal{G}$ by coarsening the raw Z_{S} (e.g., medians/quantiles computed *pooled* over T) and drop cells lacking both arms. Within each $g \in \mathcal{G}$ we take the treated–control mean difference of r_j and aggregate with weights $w_g \propto n_{1g} + n_{0g}$. This is standardization (g-computation):

$$\widehat{\tau}_j \;=\; \sum_g w_g \big(\overline{r}_{j,1g} - \overline{r}_{j,0g}\big) \;\xrightarrow{\;\mathbb{E}\;}\; \sum_g \Pr(G{=}g)\big(\mathbb{E}[r_j \mid G, g, do(1)] - \mathbb{E}[r_j \mid G, g, do(0)]\big) = \tau_j^R,$$

so the estimator is unbiased under randomization/SUTVA. The reported variance and Satterthwaite df are the usual stratified formulas.

C Experiments Details

C.1 CelebA semi-synthetic RCTs

Dataset. We use CelebA (Liu et al., 2018), a face attributes dataset with $> 200\text{k}$ images and 40 binary attributes per image [1]. Furthermore, for implementation details, labels have been doubled (we pass from `Beard` to `Has_Beard` and `Has_notBeard`). We follow the authors' official *train/val/test* split, and we employ the validation data for training SAEs and the test data to interpret them. Attributes are treated as ground-truth binary labels. From this source, we simulate several RCTs following the data generating process (DGP) described below, varying the sample size ($n \ll 200k$) and treatment effect (τ), reflecting realistic randomized controlled trial characteristics.

Data Generating Processes To evaluate discovery accuracy with known ground truth, we simulate RCTs by reusing real images but stochastically sampling treatment and outcomes from CelebA attributes:

- *Treatment:* $T \sim \text{Bernoulli}(0.5)$.
- *Outcome factors:* we designate two binary effects $Y{=}(Y_1, Y_2)$ using CelebA attributes: $Y_1{=}$`Eyeglasses`, $Y_2{=}$`Wearing_Hat`.
- *Exogenous Cause:* $W{=}$`Smiling`.

We implement a "co-effect" model in which T shifts both Y_1 and Y_2 with arm-specific probabilities and W modifies only Y_1:

$$\Pr(Y_2{=}1 \mid T{=}1) = p_1^{(Z)}, \quad \Pr(Y_2{=}1 \mid T{=}0) = p_0^{(Z)},$$

$$\Pr(Y_1{=}1 \mid T{=}t, W{=}w) = \begin{cases} p_{11}^{(Y)} & (t{=}1, w{=}1) \\ p_{10}^{(Y)} & (t{=}1, w{=}0) \\ p_{01}^{(Y)} & (t{=}0, w{=}1) \\ p_{00}^{(Y)} & (t{=}0, w{=}0) \end{cases}$$

with $W \sim \text{Bernoulli}(0.5)$. We vary effect magnitude via an ATE grid $\text{ATE} \in \{0, 0.1, \ldots, 0.8\}$ (9 values). Concretely, starting from a base rate 0.5, we set:

$$p_1^{(Z)} = 0.5 + \tfrac{\text{ATE}}{2}, \quad p_0^{(Z)} = 0.5 - \tfrac{\text{ATE}}{2},$$

and analogously for Y_1 in the $W{=}1$ arm:

$$p_{11}^{(Y)} = 0.5 + \tfrac{\text{ATE}}{2}, \quad p_{01}^{(Y)} = 0.5 - \tfrac{\text{ATE}}{2}, \quad p_{10}^{(Y)} = 0.2 + \text{ATE}, \quad p_{00}^{(Y)} = 0.2.$$

For each simulated unit, we draw (T, W, Y_1, Y_2), then assign an *actual image* whose CelebA attributes match the realized (Y_1, Y_2, W).

FM features. Each image x is encoded with SigLIP (Zhai et al., 2023) into a patch-level representation; we use the final-layer token features (dim $d{=}768$, 196 patches/token positions). Unless noted otherwise, we do not use any task-specific fine-tuning.

SAE Details. We train a SAE on SigLIP features to obtain interpretable codes $Z \in \mathbb{R}^m$ that serve as hypotheses for treatment effect estimation. Thereafter, the details for the SAE in Table 1. Lastly, to turn hidden representation into hypotheses, aggregate patchwise by *mean pooling* to a single $Z \in \mathbb{R}^{9216}$ per image. These per-image codes are the units we test in NES and baseline procedures.

Evaluation. We evaluate discoveries against concept–aligned SAE codes extracted from CelebA. Let $m{=}9216$ be the number of codes and $Z_j(X) \in \mathbb{R}$ the activation of code $j \in [m]$ on image X; a code is *active* when $Z_j(X) > 0$. For true effect $Y_k \in \{0, 1\}$ (here $k \in \{1, 2\}$) and each code j,

[1] It can be downloaded from flwrlabs/celeba

Component	Setting
Encoder nonlinearity	Top-k with k=5 active codes
Input dimension	768
Code / decoder dimension (m)	9216
Optimizer / LR / batch	Adam / 5×10^{-4} / 20
Epochs / grad clipping	20 / 1.0

Table 1: Training details for the SAE employed in semi-synthetic experiments.

we induce predictions $\hat{y}_{ik}^{(j)} := \mathbb{I}\{Z_j(X_i) > 0\}$ and compute the F1-score of $\{\hat{y}_{ik}^{(j)}\}_{i=1}^n$ against the ground-truth labels $\{y_{ik}\}_{i=1}^n$; the *best* neuron for the concept is then

$$g_k \;:=\; \arg\max_{j\in[m]} \; \mathrm{F1}\Big(\{\hat{y}_{ik}^{(j)}\}_{i=1}^n,\; \{y_{ik}\}_{i=1}^n\Big)\,.$$

The resulting ground-truth set of affected codes is $\mathcal{G} := \{g_1, g_2\}$ (in general $|\mathcal{G}| = r$). Each method (NES or a baseline) returns a set of discovered codes $\mathcal{S} \subseteq [m]$, which we compare to $\mathcal{G}$ via set metrics. Defining $\mathrm{TP} := |\mathcal{S} \cap \mathcal{G}|$, $\mathrm{FP} := |\mathcal{S} \setminus \mathcal{G}|$, and $\mathrm{FN} := |\mathcal{G} \setminus \mathcal{S}|$, we report

$$\mathrm{Precision} \;=\; \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \qquad \mathrm{Recall} \;=\; \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \qquad \mathrm{F1} \;=\; \frac{2\,\mathrm{Precision}\cdot\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}},$$

and the set Intersection-over-Union (IoU)

$$\mathrm{IoU} \;=\; \frac{|\mathcal{S}\cap\mathcal{G}|}{|\mathcal{S}\cup\mathcal{G}|} \;=\; \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}\,.$$

C.2 ISTANT

Data and RCT. We considered the randomized controlled trial introduced by Cadei et al. (2024). Videos of ant triplets were collected under randomized treatment/control assignment. Throughout our unsupervised pipeline, *domain annotations from biologists were used only a posteriori for interpretation/evaluation of discovered codes, never for training*, as discussed in the main text.

FM features. Each frame X is encoded with DINOv2 (Oquab et al., 2023) into a patch-level representation; we use the final-layer token features (dim d=384, 256 patches/token positions). Unless noted otherwise, we do not use any task-specific fine-tuning.

SAE Details. We train a SAE on the DINOv2 features to obtain interpretable codes $Z \in \mathbb{R}^m$ that serve as hypotheses for treatment effect estimation. Thereafter, the details for the SAE are in Table 2. Lastly, to turn hidden representation into hypotheses, we aggregate patchwise by *mean pooling* to a single $Z \in \mathbb{R}^{4608}$ per frame. These per-frame codes are the units we test in NES and baseline procedures.

Component	Setting (ISTAnt)
Encoder nonlinearity	Top-K with K=20 active codes
Input dimension	384
Code / decoder dimension (m)	4608
Optimizer / LR / batch	Adam / 5×10^{-4} / 128
Epochs / grad clipping	10 / 1.0

Table 2: Training details for the SAE employed on ISTAnt.

Evaluation. Evaluation follows exactly the CELEBA protocol: we score discovered codes against ground-truth concepts via code–induced predictions and compute Precision/Recall/F1 and IoU for the set of returned codes (with domain annotations used only for interpreting and quantifying performance, not for training).

D ADDITIONAL EXPERIMENTS

D.1 EVALUATION ON CELEBA: WHAT DOES OUR GROUND TRUTH MODEL?

We assess how well SAE codes behave as measurement channels on CELEBA by aligning individual neurons with ground–truth attributes (see Section C.1). For each code j, we treat the event $Z_j > 0$ as a binary predictor and compute its F1–score against the attribute label. The two most predictive neurons for the two affected factors are: (i) neuron `38` for `Wearing_Hat` with $\mathrm{F1} = 0.841$, and (ii) neuron `6051` for `Eyeglasses` with $\mathrm{F1} = 0.748$. Qualitative inspection of the top–activated images (Figure 6) confirms that these codes fire on the intended visual concept, supporting their use for exploratory causal inference.

Figure 6: **Qualitative neurons' interpretations.** Each panel shows the 12 most–activated test images for the most predictive neuron of each affected outcome concept (activation = highest code value).

At the same time, the F1–score spectra over *all* neurons reveal a familiar pattern: a single, dominant "monosemantic" code per concept, accompanied by a long tail of weaker yet clearly non–zero correlations (Figure 7). This tail is stronger for `Eyeglasses`, where several neurons reach moderate F1, indicating broader leakage/entanglement. As discussed in the main text (see Section 3), such low–amplitude but widespread correlations are precisely what trigger the *Paradox of Exploratory Causal Inference*: with enough power, standard multi–testing will flag all of these leakage neurons as "significant." Our NES counters this by retrieving the leading effect first and then recursively stratifying on previously discovered codes, so that subsequent tests target the *residual* causal signal rather than its leakage.

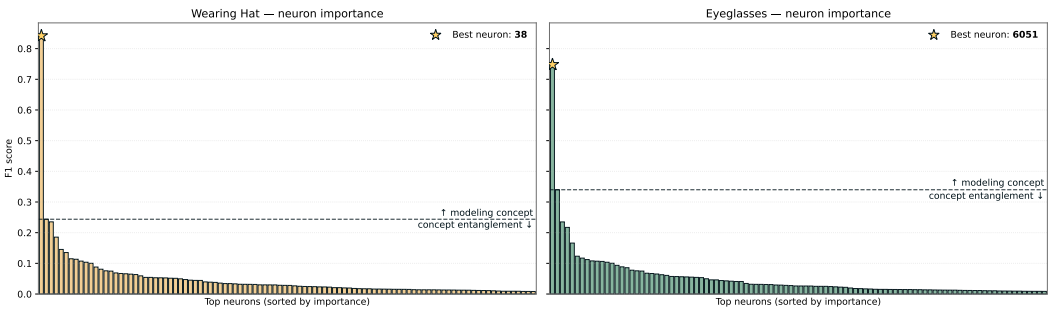

Figure 7: **Monosemantic peaks with entanglement tails.** For each attribute, we rank SAE codes by F1 against the CELEBA label and visualize the top performers in order.

D.2 In–depth analysis: full semi–synthetic results

This subsection expands the quantitative picture in Figure 4 by showing the *more complete grid* of results across sample size and effect magnitude, for two evaluation regimes:

1. **Unknown number of effects (r).** Each method returns its *own* set of significant codes at level α or simply the `Top-K`. We then report Precision, Recall, and IoU against the ground–truth affected codes (Section C.1).

2. **Known number of effects (r).** We assume to know the true number of effects, and we just look at $r-$ highest effect among each method. We again compute Precision, Recall, and IoU (namely, we apply `Top-2` selection on top of other methods).

As detailed in Appendix C.1, we vary (*i*) the **sample size** $n \in \{30, 50, 100, 250, 500, 1000\}$ and (*ii*) the **ATE magnitude** $\tau \in \{0.1, \ldots, 0.8\}$, holding the semi–synthetic DGP and SAE training protocol fixed. Each cell aggregates 10 random seeds (RCT re–draws and SAE initializations).

Main takeaways. Across both regimes and over the entire grid, NES maintains high Precision and IoU while matching the best Recall of baselines. When the experiment power increases (larger n or τ), vanilla t–tests and classical multiplicity corrections (FDR/Bonferroni) exhibit the *significance–collapse* behavior: Recall saturates but Precision drops sharply as leakage neurons become significant, driving IoU toward zero. Enforcing the correct cardinality (r known) mitigates over–selection but does *not* resolve entanglement: baselines still replace a true effect with a leakage surrogate in later picks, keeping Precision < 0.5 in the high–power regime. In contrast, NES's residual stratification peels one principal effect component per round and then *stops*, preserving interpretability.

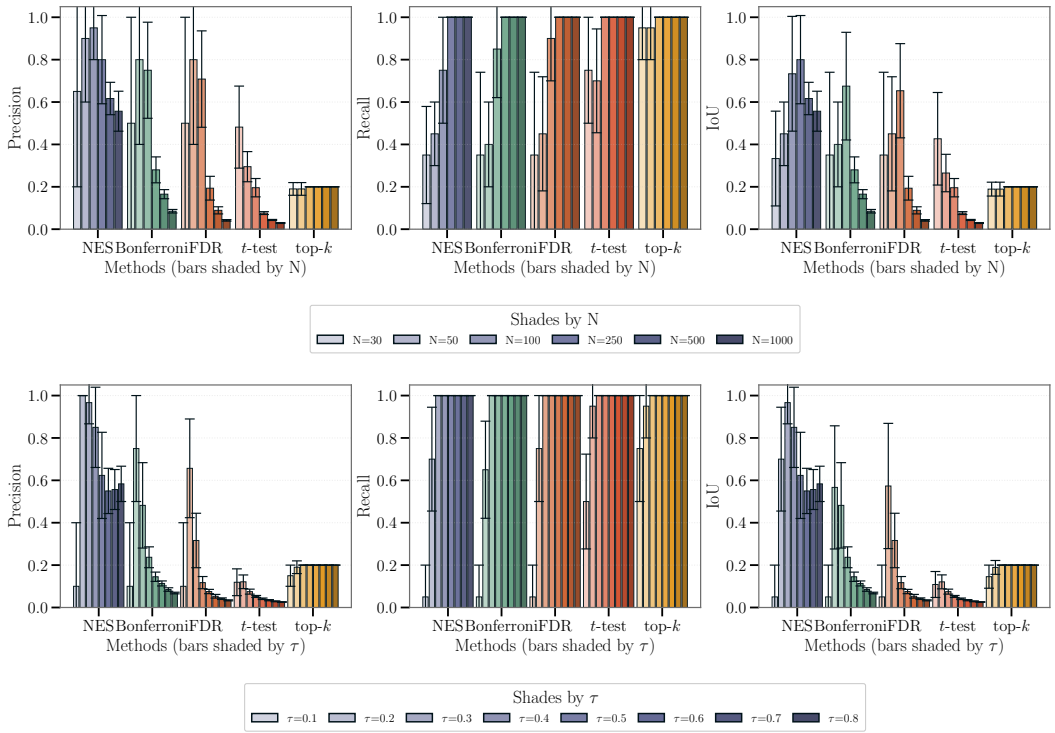

Figure 8: **Full results with r unknown.** Precision, Recall, and IoU for all methods when each returns its own set of significant codes at level $\alpha = 0.05$.

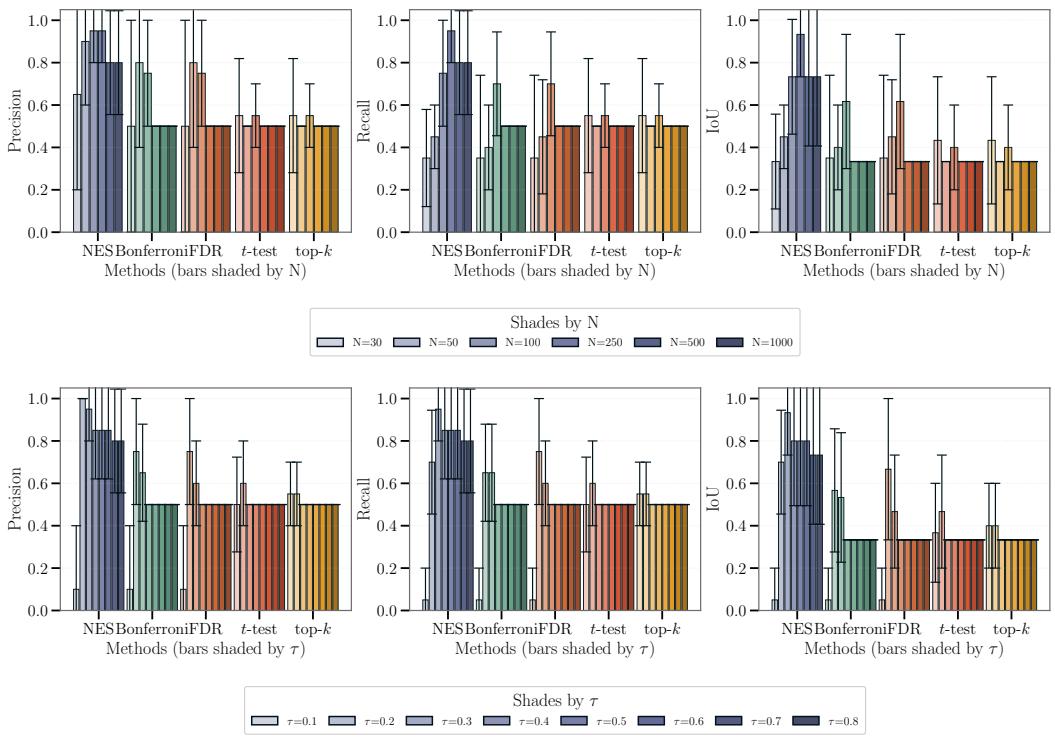

Figure 9: **Full results with r known (top–r selection).** Precision, Recall, and IoU when every method is forced to return exactly r codes (the true number of effects).

D.3 Ablation I: No Causal Effect

We repeat the semi-synthetic evaluation of Section D.2 but set the true ATE to zero, namely $\tau = 0$ factors. In this regime, a well-calibrated discovery procedure should return *no* significant neurons.

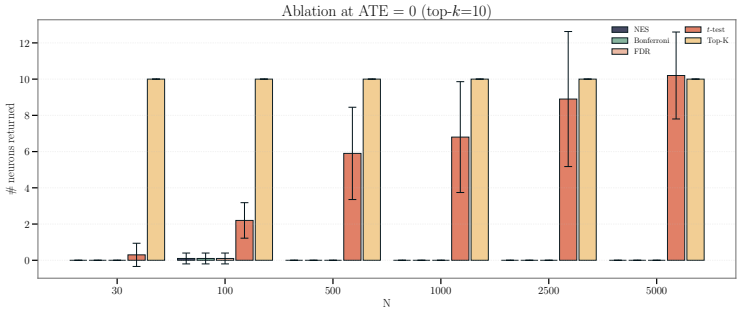

Figure 10: **Zero-effect ablation .** Number of discovered neurons by method when ATE is 0.

We keep the data-generating process, foundation model, SAE training, and testing grid over sample sizes n identical to Section D.2, changing only the interventional contrast to ATE $= 0$. For each method, we record the number of discoveries per run. Across all sample sizes, NES returns an empty set: in the first iteration, no neuron survives Bonferroni at level α/m, and the recursion halts. Furthermore, both Bonferroni and FDR also yield essentially zero discoveries. In contrast, the uncorrected t-test produces spurious positives (false discoveries), and Top-k necessarily reports k indices by design, labeling pure noise as significant. This behavior matches our theoretical intuition: with $\tau = 0$ there is no effect vector to leak into entangled coordinates, so the paradox of Sec. 3 does not arise; procedures that control multiplicity (NES via its first-step Bonferroni gate, Bonferroni,

and FDR) appropriately abstain, whereas selection rules that ignore multiplicity (Top-k, plain t-tests) over-discover.

D.4 Ablation II: Testing in NES

We compare three per-round gates inside NES (Alg. 1): Bonferroni, FDR, and t-test. Same setup as Sec. D.2; only the multiplicity rule changes while recursion and residual stratification are unchanged. *NES-Bonf.* delivers the cleanest recoveries: highest precision/IoU and exact stopping at r effects when powered; under ATE=0 it returns none (cf. Ablation D.3). *NES-t* is most exploratory for small sample size and effect magnitude but over-selects as power grows, i.e., Paradox of Exploratory Causal Inference.

Recommendation. Prefer a multi-hypothesis testing correction, i.e., Bonferroni/FDR, when the power of the experiment is high, while consider t-test for a more explorative approach in low power regime.

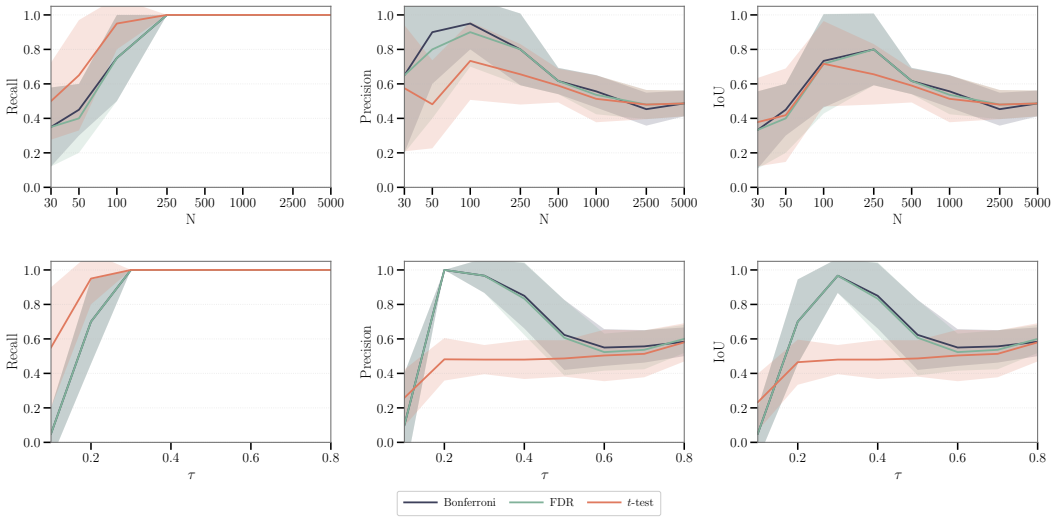

Figure 11: **Testing in NES.** Bonferroni: best precision/IoU and exact stopping; FDR: higher sensitivity in low power, minor over-selection; t-test: exploratory but prone to over-discovery as power increases.

D.5 Ablation III: AIPW vs. Associational Difference

Throughout the paper, our per-neuron hypothesis test uses the *associational difference* (AD), i.e., a two-sample t-test on the treated–control difference in means. In randomized trials, AD is unbiased for the ATE, but it is not semiparametrically efficient. A standard variance–reduction alternative is *Augmented Inverse Propensity Weighting* (AIPW; Robins et al., 1994), which orthogonalizes the estimator against misspecification of either the propensity score or the outcome regression.

Setup. For each code j, let Z_{ij} be its activation for unit i, $T_i \in \{0,1\}$ the treatment, and W_i observed exogenous causes. We compute the AIPW pseudo-outcome

$$\tilde{Z}_{ij} \;=\; \hat{\mu}_{1j}(W_i) - \hat{\mu}_{0j}(W_i) \;+\; \frac{T_i}{\pi(W_i)}\big(Z_{ij} - \hat{\mu}_{1j}(W_i)\big) \;-\; \frac{1-T_i}{1-\pi(W_i)}\big(Z_{ij} - \hat{\mu}_{0j}(W_i)\big), \quad (35)$$

where $\pi(W) = \Pr(T=1 \mid W)$ (known and constant $\pi = 0.5$ in our RCT), and $\hat{\mu}_{tj}(W) \approx \mathbb{E}[Z_j \mid T=t, W]$ is a nuisance regression. The AIPW estimate of the code-level ATE is $\hat{\tau}_j^{\text{AIPW}} = \frac{1}{n}\sum_i \tilde{Z}_{ij}$; we test $H_0 : \tau_j = 0$ via a one-sample t-test on $\{\tilde{Z}_{ij}\}_i$ with robust variance.

Results. Figure 12 compares AD vs. AIPW on the semi-synthetic benchmark across sample size n and effect magnitude τ. In our setting—with a truly randomized treatment ($\pi = 0.5$) and a *single*

binary covariate W—AIPW yields only marginal efficiency gains: Precision/Recall/IoU curves are essentially overlapping, with small stability improvements for AIPW at the smallest n. Crucially, orthogonalization affects variance but *does not* resolve entanglement: the significance–collapse phenomenon for standard multi-testing (Section 3) persists under AIPW, and NES retains its advantage because its benefit comes from recursive stratification (disentangling residual effects), not from how the first-step mean contrast is estimated.

Takeaways. (i) In pure RCTs with weak, low-dimensional W, AD is competitive and simpler. (ii) AIPW can be preferred when richer exogenous information is available (higher-dimensional W, imbalance, or mild protocol deviations), where its variance reduction can translate into earlier detection of the leading effect; (iii) regardless of AD or AIPW, NES's stratified recursion is the key to avoiding over-discovery under entanglement.

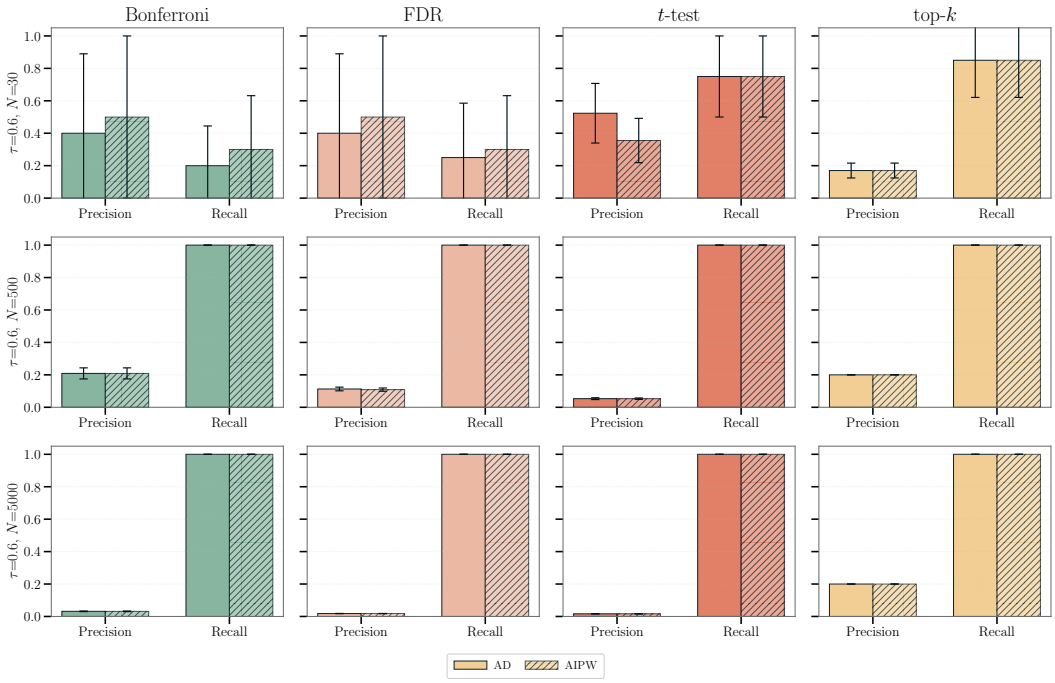

Figure 12: **AIPW vs. AD on semi-synthetic RCTs.** Precision, Recall, and IoU when replacing the per-neuron associational difference (AD) with AIPW (Eq. 35) for baselines and the first NES step.