From Tokens to Policy: Causal and Interpretable Heterogeneous Treatment Effects Identification

Riccardo Cadei [1] **Frank Otchere** [2] **Nyasha Tirivayi** [2]
Gustavo Angeles Tagliaferro [2] **Falco J. Bargagli-Stoffi** [3] **Francesco Locatello** [1]

Abstract

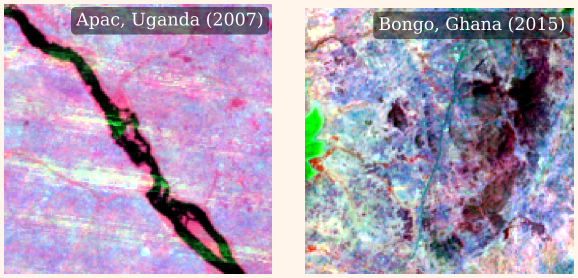
Heterogeneous Treatment Effect (HTE) identification is crucial to explain the impact of an intervention and optimize our policies accordingly. Existing approaches trade expressivity for interpretability, but, if some active heterogeneity drivers are unmeasured, methods at both ends of this spectrum allow for spurious HTE characterization with no causal reading. In this work, we focus on controlled experiments and argue that an oracle HTE causal characterization via the latent interactors is now within reach, thanks to (i) more extensive pre-treatment measurements, i.e., multimodal and multi-view, and (ii) scalable representations with minimal human supervision. We then re-frame HTE identification as a Markov-blanket discovery problem on a sufficient and aligned pre-treatment representation, and introduce *Neural EXposure Interaction Search* (NEXIS), an iterative procedure with provable and empirically validated consistent selection. We deploy NEXIS on two anti-poverty programs in Africa, augmenting each with satellite imagery capturing previously unmeasured environmental effect modifiers, leading to novel, interpretable and prescriptive guidelines to optimize the programs' next iterations.

Website: riccardocadei.com/NEXIS/

1. Introduction

Real-world interventions rarely work the same way for everyone. Understanding *why* and *how* a mechanism varies is essential to optimize our policies accordingly (Rothman et al., 1980; Gail & Simon, 1985). Traditional Heterogeneous Treatment Effect (HTE) estimation approaches trade expressivity for interpretability over the available pre-treatment measurements. Yet, even an accurate and interpretable estimation does not tell policymakers how to intervene accordingly, due to potential spurious dependencies. Indeed, HTE is *causally* explained by the standalone treatment interactions (VanderWeele, 2009), potentially unmeasured, but it is also reflected in their proxies, common-cause companions, and upstream indirect modifiers (VanderWeele & Robins, 2007), which offer a misleading or no causal handle on the response. In the absence of an interaction's measurement and identification, several approximate characterizations may arise, arbitrarily predictive in-distribution but drastically failing to generalize, e.g., under a new policy regime.

Motivation: ***Disentangling the impact of anti-poverty programs beyond survey data.***
Anti-poverty programs such as YOP in Uganda (Blattman et al., 2014) and LEAP 1000 in Ghana (Ghana LEAP 1000 Evaluation Team, 2018), offer cash transfers and livelihood support for economic re-integration, and household welfare. They commonly monitor household demographics despite ignoring the environmental context, e.g., water access. They are then evaluated on average, but to optimize resource allocation in the next iterations, policymakers need to *causally* explain all the impact heterogeneity, retrievable only by integrating external data sources, potentially multi-modal and unstructured, e.g., satellite imagery (see Figure 2). Standalone demographic covariates are insufficient (Burke et al., 2021; Athey & Imbens, 2017).



(a) Perennial river *(b)* Ephemeral waterways

Figure 2. Examples of environmental interactors in anti-poverty programs, commonly ignored in survey data.

Such causal characterization has been historically out of

[1]ISTA [2]UNICEF [3]UCLA. Correspondence to: Riccardo Cadei <riccardo.cadei@ista.ac.at>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.

reach: existing HTE estimation approaches either model the treatment effect flexibly without identifying its primitive drivers (Wager & Athey, 2018; Athey et al., 2019; Hahn et al., 2020; Künzel et al., 2019; Nie & Wager, 2021; Kennedy, 2023; Hill, 2011; Shalit et al., 2017; Jerzak et al., 2023), or identify interpretable subgroups over curated structured covariates (Foster et al., 2011; Athey & Imbens, 2016; Bargagli-Stoffi et al., 2020; Hines et al., 2022; Paillard et al., 2024; Bargagli-Stoffi et al., 2020; 2022), with no guarantee of separating the (possibly unobserved) interactions from their spurious correlates. We argue two recent advances finally brought causal HTE identification within reach in controlled experiments:

i. *Increasing measurement capacity*, making it more plausible that any treatment interaction is captured, even if entangled in complex multi-modal and multi-source observations, e.g., satellite imagery, sensor streams or text (Burke et al., 2021; Athey & Imbens, 2017).

ii. *Modern representation learning*, enabling sufficiently good representations, i.e., aligned with the latent factors of variation of interest, and scalable to the higher data volume. Particularly we focus on sparse-autoencoder dictionaries (Cunningham et al., 2023; Bricken et al., 2023; Gao et al., 2024; Templeton et al., 2024; Makelov et al., 2024; Pach et al., 2025) learned from pre-trained foundation-model representations.

Even in this novel data regime, identifying the treatment interactions among all the existing non-causal effect modifiers is not trivial. Furthermore, since only partial disentanglement is achieved in practice (Chanin et al., 2024), almost any coordinate inherits marginal heterogeneity from the treatment interactors, making *effect modification search* fundamentally misaligned from *interactors search*. Recovering only direct modifiers requires conditional reasoning to identify the minimal and sufficient HTE representation, if observed. We then recast the search to a Markov-blanket discovery problem (Tsamardinos et al., 2003; Aliferis et al., 2010) for the effect heterogeneity over the learned dictionary, introducing *Neural EXposure Interaction Search* (NEXIS), a novel forward-backward procedure based on conditional effect-equivalence testing. We prove high probability recovery guarantees, and provide extensive empirical validation and ablations on semi-synthetic experiments, with practical variants for different data regimes.

We then deploy NEXIS on two real-world anti-poverty programs, the Youth Opportunities Program (YOP) in Uganda (Blattman et al., 2014) and Livelihood Empowerment Against Poverty Programme (LEAP 1000) in Ghana (Ghana LEAP 1000 Evaluation Team, 2018), augmenting each with novel satellite imagery, capturing previously ignored environmental factors, providing concrete guidelines for the next iterations of these programs. To the best of our knowledge, this work introduces the first procedure to identify a *causal* and *interpretable* HTE characterization from complex observations, translating raw measurements end-to-end into practical and robust guidelines for policy adaptation.

In summary, our contributions are:

- targeting ***causal* and *human-interpretable* HTE identification** and formulating it as a Markov-blanket discovery problem over a sufficient and aligned pre-treatment representation;
- introducing NEXIS, a **principled procedure** for the target HTE identification, with recovery guarantees, tests and ablations providing practical guidelines for different data regimes;
- deploying NEXIS on **two anti-poverty programs** in Africa which we complement with satellite imagery, surfacing novel prescriptive heterogeneity explanations for policy adaptation.

2. Causal and Interpretable HTE Identification

We formalize here the population target and the corresponding observed-space recovery problem. Note: *For notational clarity, we present our development for a randomized experiment with binary treatment under perfect compliance, but the same ideas extend naturally to other causal heterogeneity functionals, such as local average treatment effect (LATE) heterogeneity (Bargagli-Stoffi et al., 2022) or average-treatment-effect-on-the-treated (ATT) heterogeneity (Callaway & Sant'Anna, 2021) in quasi-experimental settings.*

2.1. Causal and Interpretable HTE characterization

Consider a controlled experiment with n units indexed by $i \in [n] := \{1, \dots, n\}$, drawn *i.i.d.* from a population distribution $\mathcal{P}$, where a binary treatment $T \in \{0, 1\}$ is randomly assigned and potentially affecting an outcome variable $Y \in \mathbb{R}$. Let $Y(0), Y(1)$ be the potential outcomes under control and treatment respectively (Rubin, 1974) (assuming consistency and not interference, i.e., SUTVA), we define the treatment effect by their comparison, i.e., $\tau := Y(1) - Y(0)$. The treatment effect can vary across units due to the treatment interaction with a set of pre-treatment unit characteristics $\boldsymbol{W}^{\text{dir}} \in \mathbb{R}^r$, called *direct* effect modifiers or *interactors* (VanderWeele, 2009; VanderWeele & Robins, 2007). For policy guidance we aim to identify the corresponding Conditional Average Treatment Effect (CATE) functional:

$$\tau(\boldsymbol{w}) := \mathbb{E}[\tau \mid \boldsymbol{W}^{\text{dir}} = \boldsymbol{w}] \quad \forall \boldsymbol{w} \in \text{supp}(\boldsymbol{W}^{\text{dir}}). \quad (1)$$

Targeting the $\boldsymbol{W}^{\text{dir}}$ characterization is desirable for two reasons:

i. It represents the *minimal* and *sufficient* summary of the effect heterogeneity: by definition $\boldsymbol{W}^{\text{dir}}$ d-separates any pre-treatment variable from the treatment effect τ (Richardson & Robins, 2013) providing a Markov blanket of the treatment effect heterogeneity (Pearl, 1988; Koller & Friedman, 2009).

ii. It is *causal*[1] and *interpretable*, identifying directly the corresponding interventional heterogeneity functional:

$$\tau^{do}(\boldsymbol{w}) := \mathbb{E}[\tau \mid do(\boldsymbol{W}^{\text{dir}} = \boldsymbol{w})] \; \forall \boldsymbol{w} \in \text{supp}(\boldsymbol{W}^{\text{dir}}), \quad (2)$$

regardless of the other modifiers' distribution, while marginal interventions, e.g., on a specific $\boldsymbol{W}^{\text{dir}}$ coordinate, are sub-population specific due to potential inter-direct modifier interactions.

But Equation 1 is not the unique effect heterogeneity characterization, and $\boldsymbol{W}^{\text{dir}}$ is generally latent.

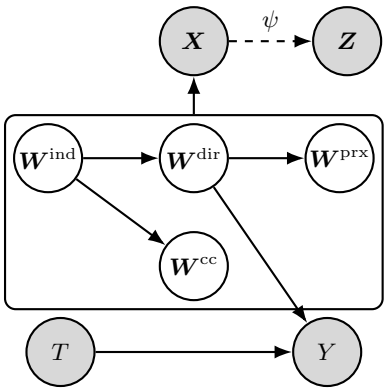


Figure 3. Problem setting and effect modification taxonomy via a causal model (VanderWeele & Robins, 2007). Gray: observed; white: latent.

There are indeed other types of effect modifiers, i.e., variables varying the treatment effect within their strata. VanderWeele & Robins (2007) characterized them, distinguishing the *direct* modifiers $\boldsymbol{W}^{\text{dir}}$, from their ancestors *indirect* modifiers $\boldsymbol{W}^{\text{ind}}$, the *proxies* $\boldsymbol{W}^{\text{prx}}$ descending from direct modifiers, and modifiers *by common cause* $\boldsymbol{W}^{\text{cc}}$ sharing a common ancestor with a direct modifier, as illustrated in the causal model[2] in Figure 3. Among these, only the direct modifiers carry a causal interpretation: indirect modifiers operate only via mediation, while proxies and common-cause modifiers have no causal pathway to τ (Richardson & Robins, 2013), and should be interpreted cautiously. Indeed, consider an anti-poverty program where market presence is a latent direct modifier. Road density visible from satellite imagery may surface as a heterogeneity correlate, while being an effect modifier by common cause. A naive interpretation would suggest building roads to amplify the program effect, while the actual driver is market presence; a road built to a place with no market would not impact the effect.

Furthermore $\boldsymbol{W}^{\text{dir}}$ is rarely fully measured directly, and it is instead entangled in complex pre-treatment observations potentially multi-source and multi-modal. Let $\boldsymbol{X} \in \mathcal{X}$ denote the aggregation of all such pre-treatment measurements, e.g., combining structured demographic covariates from questionnaire responses, with unstructured signals such as satellite imagery. We assume:

> **Assumption 2.1** (Measurement Sufficiency)**.** The measured pre-treatment variables $\boldsymbol{X}$ entangle all the heterogeneous treatment effect information, i.e.,
>
> $$\mathbb{E}[\tau \mid \boldsymbol{X}] = \mathbb{E}[\tau \mid \boldsymbol{W}^{\text{dir}}] \qquad \text{a.s.} \qquad (3)$$

Assumption 2.1 is plausible in well-designed experiments where the relevant factors of variation are controlled or measured, even when the actual latent modifiers are unknown, provided access to extensive pre-treatment measurements. It is also necessary: if violated, complete explanation of effect heterogeneity becomes theoretically impossible, and any analysis is forced to approximate it via non-direct effect modifiers, invalidating any causal interpretation and extrapolation.

Existing methods for HTE estimation from sufficient complex measurements are not causal: either lacking *expressivity* to model the complex covariates domain $\mathcal{X}$, or incautiously learning *non-invariant* mechanisms (Yao et al., 2025b; Lopez-Paz, 2026). However, a causal characterization is required for policy: "*which additional interventions or treatment assignment revisiting can increase the impact of a program?*". We overcome this trade-off by leveraging modern *representation learning* for extensive interpretable hypotheses generation, which would not scale with human annotations alone without prohibitive cost and bias. To the best of our knowledge, this work targets for the first time a causal HTE characterization from complex observations, i.e., identifying the direct modifiers $\boldsymbol{W}^{\text{dir}}$, which is also interpretable by construction due to causal mechanisms sparsity (Schölkopf et al., 2021) and representation minimality.

2.2. Learning interpretable representations for Direct Effect Modifiers identification

We propose to rely on pretrained foundation-model representations to represent the complex pretreatment measurements,

[1]Note that the CATE is per-se a causal estimand, but not with respect to its characterization.

[2]Note that: (i) a causal model does not provide the full effect modification characterization without structural equations, (ii) direct effect modifiers are not necessarily prognostic variables for the outcome, (iii) any effect modifier can additionally be outcome prognostic.

and then identify (search) the direct effect modifiers. Indeed, overwhelming evidence supports the hypothesis that the meaningful factors of variation are already linearly captured within foundation-model embedding spaces, including the direct effect modifiers (Chen et al., 2020; Dosovitskiy et al., 2020; Radford et al., 2021; Caron et al., 2021), even without supervision. Furthermore, learning end-to-end a meaningful representation on the target experiment is generally statistically infeasible due to the limited data. We then represent the pre-treatment measurements via a model $\psi : \mathcal{X} \to \mathbb{R}^m$, defining the covariates representation

$$\boldsymbol{Z} := \psi(\boldsymbol{X}) \in \mathbb{R}^m, \tag{4}$$

by learning a sparse autoencoder (SAE) on top of a frozen foundation-model encoder for the modality of interest (Cunningham et al., 2023; Bricken et al., 2023), optionally combined with classical dictionary-learning components (Mairal et al., 2009) and appended hand-crafted covariates from the processable modalities[3]. The resulting dictionary coordinates are interpretable by construction (Oquab et al., 2023; Zhai et al., 2023; Park et al., 2023), and each sparse coordinate can be labeled post-hoc, e.g., inspecting its top-activating inputs via a vision-language model. Two distinct desiderata on the learned representation arise naturally:

Assumption 2.2 (Representation Sufficiency)**.** All the pre-treatment measurements $\boldsymbol{X}$ heterogeneous treatment effect information is preserved in the representations $\boldsymbol{Z}$, i.e.,

$$\mathbb{E}[\tau \mid \boldsymbol{Z}] = \mathbb{E}[\tau \mid \boldsymbol{X}] \qquad \text{a.s.} \tag{5}$$

In a SAE the Representation Sufficiency is naturally enforced by the reconstruction loss.

Assumption 2.3 (Principal Alignment)**.** There exists a unique injective map $\pi : [r] \hookrightarrow [m]$, $k \mapsto j_k$, such that for every $k \in [r]$:

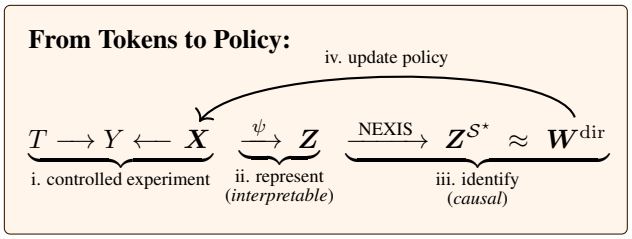
$$W^{\mathrm{dir},k} \perp\!\!\!\perp \boldsymbol{Z}^{[m]\setminus\{j_k\}} \,\Big|\, Z^{j_k}. \tag{6}$$

We call $\mathcal{S}^\star := \{j_1, \ldots, j_r\}$ the set of *principal proxy direct effect modifiers coordinates*, and Z^{j_k} the *principal proxy* for the direct effect modifier $W^{\mathrm{dir},k}$.

Principal Alignment asks each direct modifier to be summarized by a distinct dominant coordinate, with no residual signal scattered across the others, yet allowing for shared components. It is a population-level idealization of the regime SAEs target for (Makelov et al., 2024; Pach et al., 2025), and whether it holds for a specific dictionary can be probed empirically for supervised concepts (Yao et al., 2025a).

[3]Different hypotheses generation strategies can be considered provided they satisfy the method requirements.

Overall, Measurement and Representation Sufficiency are necessary conditions for full HTE identification, while Principal Alignment is what well-defines its minimal characterization, i.e., the principal proxies for the latent direct effect modifiers. Together they well-define the causal and interpretable HTE characterization, and unlock its identification in controlled experiments. Then, from a controlled experiment, we develop a method for the principal proxies identification and update the experiment design accordingly, as illustrated in the following cyclic diagram:



3. On the Distinction between Effect Modification and Interaction

Identifying $\mathcal{S}^\star \subseteq [m]$ is not straightforward, even under Assumptions 2.1–2.3. Pre-treatment measurement representations can encode, among others, redundant effect modifiers, and a principled procedure needs to select only the r direct modifiers *principal proxies*, without $\boldsymbol{W}^{\mathrm{dir}}$ supervision.

Effect modification entanglement. A natural baseline is coordinate-wise testing, marginally screening each representation coordinate for effect modification. In learned representations, however, each coordinate can principally proxy also non-direct effect modifiers, and furthermore, non-principally proxy many other effect modifiers due to the broad representation entanglement (Chanin et al., 2024), regardless of the Principal Alignment assumption. Let

$$\mathcal{M}^\star := \big\{ j \in [m] : \mathbb{P}\big(\mathbb{E}[\tau \mid Z^j] \neq \mathbb{E}[\tau]\big) > 0 \big\} \tag{7}$$

the effect-modifier coordinates within the learned representation. Every $j_k \in \mathcal{S}^\star$ belongs to $\mathcal{M}^\star$[4], but so does any entangled coordinate with a real effect modifier, and in a rich dictionary of $m \gg r$ atoms (necessary for reconstruction), non-principal proxy coordinates vastly outnumber the principal, making $\mathcal{M}^\star$ a redundant and non-causal superset of the target (VanderWeele, 2009). Ranking by marginal activation magnitude does not resolve the issue: a strongly-activating non-principal proxy of a high-magnitude direct modifier can outrank the principal signal of a weaker direct modifier, and even if ordered, the boundary between principal and non-principal coordinates is unknown a priori.

[4]Under Simpson-style cancellation a principal coordinate could fail to be marginally active; we invoke mean faithfulness (Spirtes et al., 2000) for simplicity of discussion without loss of generality, focusing on Type 1 error, i.e., False Discoveries.

Experimental Power Paradox. The failure of marginal effect modification testing is not a small-sample artifact that more data can fix, it is structural. As experimental power grows, through larger n, stronger signals, or more sensitive tests, marginal screening converges to $\mathcal{M}^\star$ and *accumulates* non-direct modifiers, each backed by genuinely significant evidence. The procedure is consistent for the wrong target: more power makes the gap between $\mathcal{M}^\star$ and $\mathcal{S}^\star$ more visible, not less. Recovering $\mathcal{S}^\star$ requires conditioning, ruling out each candidate in the presence of the others to target a minimal and sufficient representation, which is the design principle behind our method, NEXIS.

Empirical evidence. We construct a semi-synthetic RCTs benchmark on CelebA (Liu et al., 2018; Mencattini et al., 2025) with $|\mathcal{S}^\star| = 2$ known direct modifiers (*wearing a hat*, *wearing eyeglasses*); pre-treatment information is face images only. A trained SAE (m=13,824 codes) on SigLIP representations (Zhai et al., 2023) provides a valid candidate dictionary. Figure 4 reports Precision and Recall in the recovery of $\mathcal{S}^\star$ for marginal screening and NEXIS as effect size and sample size grow. Marginal screening faithfully tracks $\mathcal{M}^\star \approx [m]$ while diverging from $\mathcal{S}^\star$ with precision collapse; NEXIS consistently recovers $\mathcal{S}^\star$ monotonically with the power. Experiment details and extended ablations are reported in Appendix C.

4. Neural EXposure Interaction Search

We propose *Neural EXposure Interaction Search* (NEXIS), an iterative procedure for identifying the principal direct modifiers proxies $\mathcal{S}^\star$ from a controlled experiment, given pre-treatment representations. At each round NEXIS selects the coordinate that, conditional on what has already been selected, carries the strongest residual heterogeneity signal, and removes any previously-selected coordinate that has become redundant given the rest. The atomic operation is *CATE-equivalence testing*, returning, for a candidate $j \in [m]$ and selection $S \subseteq [m] \setminus \{j\}$, a valid p-value $p_j(S)$ for

$$H_0(j \mid S): \ \mathbb{E}[\tau \mid \boldsymbol{Z}^{S\cup\{j\}}] = \mathbb{E}[\tau \mid \boldsymbol{Z}^{S}] \quad \text{a.s.} \qquad (8)$$

Concrete instantiations and suggested defaults are detailed in Appendix B.

NEXIS adapts the classic forward-backward Markov-blanket (MB) discovery skeleton (Tsamardinos et al., 2003; Aliferis et al., 2010) to a new target and a new setting: it discovers the Markov blanket of the CATE functional, which is not directly observed, on top of a learned pre-trained representation. Forward growth ensures every direct modifier eventually enters the selection; backward pruning re-tests retained coordinates against the full current selection at each round, removing any that has become redundant. We use Bonferroni gate for FWER control (Bonferroni, 1936), but less conservative controls can be directly plugged in, e.g., FDR (Benjamin et al., 1995).

Algorithm 1 Neural EXposure Interaction Search (NEXIS)

Input: $\{(\boldsymbol{z}_i, t_i, y_i)\}_{i=1}^n$, significance level $\alpha \in (0,1)$
$S \leftarrow \emptyset$, $S_{\text{prev}} \leftarrow \{-1\}$
while $S \neq S_{\text{prev}}$ **do**
  $S_{\text{prev}} \leftarrow S$, $\bar{S} \leftarrow [m] \setminus S$
  $j^\star \leftarrow \arg\min_{j \in \bar{S}} p_j(S)$ ▷ forward step
  if $p_{j^\star}(S) \leq \alpha/|\bar{S}|$ **then** $S \leftarrow S \cup \{j^\star\}$
  end if
  for $j \in S$ **do** ▷ backward step
    if $p_j(S \setminus \{j\}) > \alpha/|S|$ **then** $S \leftarrow S \setminus \{j\}$
    end if
  end for
end while
return S

Theorem 4.1 (Causal Identification)**.** *Given a randomized experiment $\{(\boldsymbol{z}_i, t_i, y_i)\}_{i=1}^n$, under Principal Alignment, mean-faithfulness (Spirtes et al., 2000), and test validity, NEXIS' outcome $\widehat{\mathcal{S}}_n$ satisfies:*

$$\liminf_n \mathbb{P}(\widehat{\mathcal{S}}_n = \mathcal{S}^\star) \geq 1 - \alpha, \qquad (9)$$

where, additionally assuming Measurement and Representation Sufficiency:

$$\tau^{do}(\boldsymbol{W}^{\text{dir}}) = \tau(\boldsymbol{W}^{\text{dir}}) \ = \ \mathbb{E}[\tau \mid \boldsymbol{Z}^{\mathcal{S}^\star}] \quad \textit{a.s.} \qquad (10)$$

Proof sketch. Principal Alignment makes $\mathcal{S}^\star$ *minimal* and *sufficient* for $\mathbb{E}[\tau \mid \boldsymbol{Z}]$ on the dictionary $[m]$ by $\boldsymbol{W}^{\text{dir}}$ alignment, with mean-faithfulness excluding pathological cancellations. Then NEXIS procedure reduces to forward–backward Markov-blanket discovery on the CATE over $\boldsymbol{Z}$, for which a standard IAMB-style argument (Tsamardinos et al., 2003; Aliferis et al., 2010) delivers asymptotic recall and asymptotic conditional precision, i.e., principal proxies selection consistency. Measurement and Representation Sufficiency bring the causal interpretation to the characterization. Full proof in Appendix A. □

Principal Alignment[5], mean-faithfulness, and test validity buy NEXIS the recovery of the principal proxies in the learned dictionary, one coordinate per direct modifier (Equation 9). Measurement and Representation Sufficiency further guarantee such proxies to capture all the heterogeneity information, so identifying the joint interventional effect of

[5]Without Principal Alignment the target is not even well-defined and different characterizations of the observed heterogeneity may arise, e.g., a single coordinate capturing two modifiers, also ignoring the unobserved modifiers.

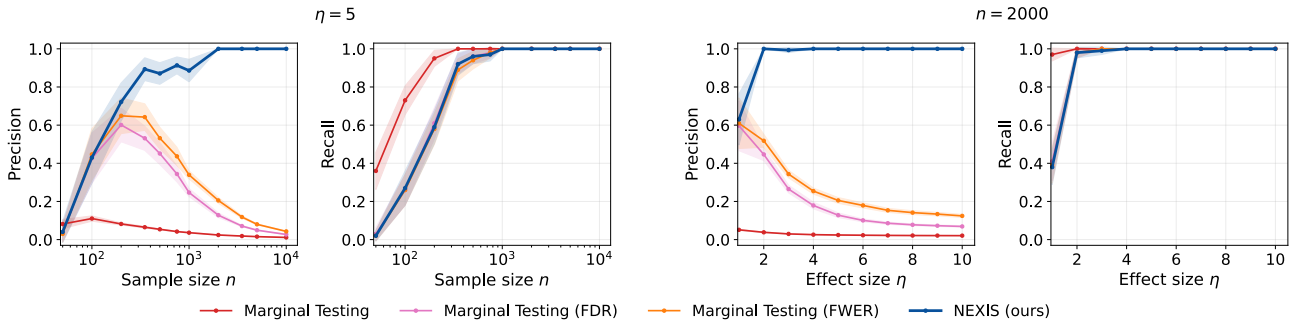


Figure 4. Experimental power paradox on CelebA: Increasing the sample size (n) or the effect magnitude (η) of the experiment, marginal screening accumulates non-direct modifiers (false discoveries) and diverges from $\mathcal{S}^\star$ recovery, i.e., precision collapse. NEXIS robustly mitigates such behavior by iterative conditioning.

the latent direct modifiers (Equation 10). As such, Theorem 4.1 provides prescriptive guidelines for policy design, licensing counterfactual comparisons, e.g., across bundled program configurations. See Appendix B for implementation details and proposed variants; see Appendix C for empirical validation and extensive ablations.

5. Related Work

Heterogeneous treatment effect identification. The bulk of existing HTE estimation methods trade expressivity for interpretability and required human supervision for downstream policy relevance. *Decision rule*-based methods (Foster et al., 2011; Athey & Imbens, 2016; Bargagli-Stoffi et al., 2020) return interpretable subgroups but operate on expensive and potentially limited hand-crafted covariates. *Forest*-based estimators (Wager & Athey, 2018; Athey et al., 2019; Hahn et al., 2020) estimate the effect heterogeneity flexibly on structured domains, but return pointwise predictions without explicit heterogeneity characterization. *Meta-learners* (Künzel et al., 2019; Nie & Wager, 2021) are agnostic to the base estimator and can in principle wrap pretrained foundation-model nuisances on unstructured covariates, but they have not been demonstrated to do so at experimental scale, and they share the same uninterpretable pointwise output. *Post-hoc* CATE interpretation methods (Hines et al., 2022; Paillard et al., 2024; Boileau et al., 2025; Bargagli-Stoffi & Melikechi, 2026) sit on top of any base estimator and rank candidate modifiers by importance recovering interpretability. However, all these methods have not been developed to operate in the context of entangled and multimodal representations of the potential effect modifiers space. *Deep Learning*-based estimators (Jerzak et al., 2023) extend HTE estimation to unstructured measurements such as imagery, gaining expressivity at the cost of modifier-level interpretability. Table 1 positions NEXIS against these families. NEXIS achieves all the desiderata by leveraging modern experimental practice, i.e., extensive measurement collections supporting Measurement Sufficiency, and modern representation learning, i.e., sparse-autoencoder-based dictionaries supporting Representation Sufficiency and Principal Alignment.

Table 1. HTE-estimators properties by family. *Empiricist*: data-driven feature extractions from raw pre-treatment observations, preventing Matthew effect (Merton, 1968). *Interpretable*: human readable HTE characterization. *Prescriptive*: targets treatment-interactive factors. Legend: ✓= fully, ✗= not fully.

Method	Empiricist	Interpretable	Prescriptive
Forest-based	✗	✗	✗
Meta-Learners	(pot.)	✗	✗
Post-hoc interpretations	✗	✓	✗
Decision rule-based	✗	✓	✗
Deep Learning-based	✓	✗	✗
NEXIS (*ours*)	✓	✓	✓

Representation learning for causal downstream tasks. Causal Representation Learning was first articulated as the recovery of causally meaningful latent variables from raw observations as a general-purpose target (Schölkopf et al., 2021), an ambitious goal that faces fundamental identifiability barriers. The field has since moved toward well-specified causal targets, in the spirit already suggested by Causal Feature Learning (Chalupka et al., 2017) albeit in controlled, low-dimensional settings. Pretrained encoders are now integrated into specific causal pipelines: for treatment-effect estimation on unstructured outcomes (Cadei et al., 2024; 2025), for unsupervised effect discovery on unstructured outcomes (Mencattini et al., 2025), for confounding adjustment via text embeddings (Veitch et al., 2020), and for HTE estimation on imagery (Jerzak et al., 2023). Our work extends this empiricist line, unlocking the identification of a causal and interpretable HTE characterization from unstructured *pre-treatment* observations.

Dictionary learning and interpretability. Foundation-model representations admit several routes to a candidate concept dictionary: classical sparse coding (Mairal et al., 2009), sparse autoencoders (Cunningham et al., 2023;

Bricken et al., 2023; Gao et al., 2024), and recent variants addressing entanglement failure modes such as feature absorption (Chanin et al., 2024), including transcoders (Dunefsky et al., 2024) and Matryoshka SAEs (Bussmann et al., 2025). Once a concept is identified, an interpretation pipeline translates it into a human-readable summary via top-activating examples, supervised probes, or LM-assisted descriptions of activating inputs (Bills et al., 2023). NEXIS is agnostic to the specific dictionary and interpretation pipeline; particularly we train top-k SAEs on top of foundation-model encoders for pre-treatment representations, and call VLM-assisted summaries contrasting top-activating against random-activation inputs for interpretation, but alternative pipelines could be considered.

6. Extending Anti-Poverty Programs Interpretation

We close the loop in the field, deploying NEXIS on two real-world cash-transfer program evaluations that we extend with satellite imagery to capture previously unmeasured environmental factors. In each case, NEXIS surfaces an interpretable and prescriptive set of direct effect modifiers, including landscape features entirely absent from the original survey instruments, translating into concrete guidelines for program adaptation and impact maximization. The evaluation is largely qualitative and supported by similar investigations since no quantitative ground truth exists. Nevertheless, the hypotheses found by NEXIS sound reasonable, warranting further validation on the policy side. We present here the results to highlight the potential for real-world positive impact of our approach. For extensive quantitative validation on semi-synthetic experiments, we point the reader to Appendix C.

Results: ***From demographic to environmental anti-poverty programs impact heterogeneity.***
Among two distinct anti-poverty programs, the impact is invariant to the target subpopulation individual-level demographics, and mainly characterized by environmental factors, coarsely proxied or even fully unmeasured in standalone survey data.

6.1. Application 1: Youth Opportunities Program (Uganda)

The Youth Opportunities Program (YOP) (Blattman et al., 2014) is an anti-poverty program launched in 2008 in post-conflict Northern Uganda, awarding cash grants to self-organised youth groups, randomized across 439 groups in 331 communities. The original analysis reports a positive average effect on productive economic capacity, captured through skilled employment and business asset accumulation. We investigate here the follow-up question of *why* and *how* the impact varies among individuals. Full pipeline, quantitative results, ablations, and limitations are presented in Appendix D.

Extending the trial with environmental measurements. YOP environmental heterogeneity investigation was first approached by Jerzak et al. (2023), who paired the trial with satellite imagery, clustering image embeddings in two groups and comparing the Group Average Treatment Effect (GATE) between them. They found preliminary satellite-based heterogeneity signal, but the analysis was limited by the model abstraction capabilities (binary clustering), outdated satellite images (~7 years before the program start and 3 bands only), and a single marginal investigation most likely entangling distinct direct modifiers effects. We re-extracted Landsat-7 multispectral imagery from Google Earth Engine over a 2005–2007 composite immediately preceding the program launch, embedded each tile through the Prithvi-EO geospatial foundation model (Szwarcman et al., 2025), and learned a sparse-autoencoder dictionary on a held-out Uganda national grid disjoint from the trial sites. The resulting candidate pool[6] unions the active dictionary atoms with the structured demographic covariates already in the trial; NEXIS then selects from it conditionally, and each retained atom is labelled post-hoc by contrasting its top and bottom activating tiles with a vision-language model (VLM).

Results. NEXIS identifies two environmental direct modifiers for each target (see Figure 5) and three ethnolinguistic-group interactions for the first target only, suggesting invariant impact across the individual-level program demographics (e.g., age, sex, parental education), all very homogeneous by experimental design.

(a) For *skilled employment*, the ethnolinguistic-group modifiers — each aggregating districts sharing a dominant language and broadly reflecting a distinct regional geography — point to post-conflict market connectivity as the key moderator: the two dampening groups (Karamojong, Lugbara) correspond to conflict-affected or geographically peripheral areas with weak supply-chain integration at trial time, while the amplifying group (Pallisa) lies outside the conflict core with shorter chains to central markets. The two environmental discoveries follow a complementary outside-option logic (Roy, 1951): perennial river presence and vegetation spatial heterogeneity pick out landscapes where fishing or bush subsistence remains viable, reducing the pull of a new skilled trade.

(b) For *business assets*, regions with higher NDVI, i.e., greater vegetation greenness, exhibit higher program impact: more productive baseline land compounds whatever inputs the grant funds. Structured agricul-

[6] Paired with hand-crafted spectral indices extracted from the same imagery (NDVI, NDWI, MNDWI, NDBI, EVI, BSI).

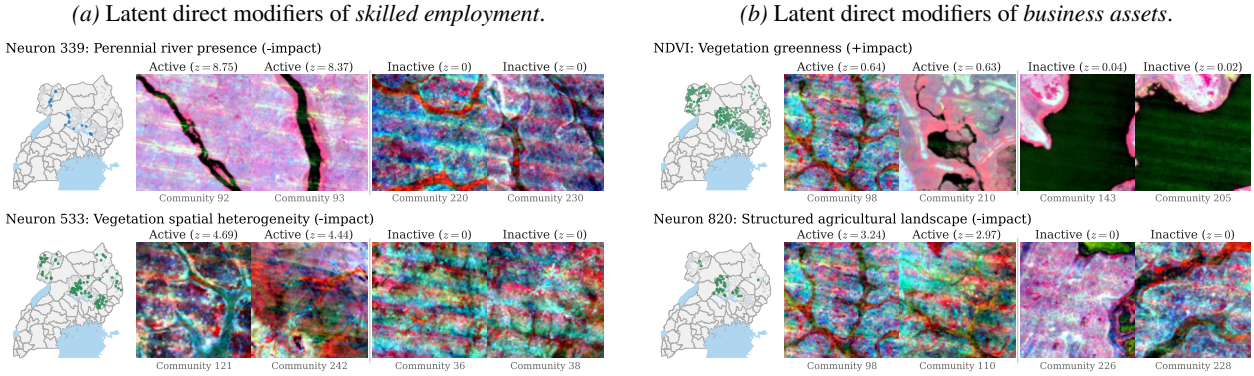


Figure 5. Satellite-derived discoveries on YOP program impact heterogeneity by NEXIS, with VLM (Bai et al., 2025) interpretations by contrasting. Each row refers to a distinct discovery, pairing the Uganda map of such feature activation and the two top- and bottom-activating Landsat tiles in false-color composite (NIR, Green, SWIR).

tural landscape, by contrast, dampens returns, plausibly reflecting crowding-out by incumbent commercial-scale agriculture (Crépon et al., 2013; Egger et al., 2022).

None of the environmental modifiers was within reach of the prior approach; they translate into concrete economic theories and adaptations for the next iteration of the program. A naive marginal screen on the same environment representation returns several dozen correlated "discoveries" per outcome, mostly correlated proxies of these few direct modifiers: a real-world instance of the experimental power paradox discussed in Section 3.

6.2. Application 2: Livelihood Empowerment Against Poverty 1000 programme (Ghana)

The Livelihood Empowerment Against Poverty (LEAP) 1000 programme (Ghana LEAP 1000 Evaluation Team, 2018) is an anti-poverty program launched in Ghana in 2015 to target early-childhood poverty and stunted nutrition, providing bimonthly cash transfers to extremely poor households with newborns. So far it has been evaluated on average via a regression discontinuity design with difference-in-differences estimation (2015–2017), covering 2,331 households across 162 communities in North and Upper East regions. We extend here its analysis to impact heterogeneity, focusing on expenditure effect and complementing 24 survey covariates with novel environmental information from satellite imagery. For full results, pipeline details and limitations see Appendix E.

Extending the trial with environmental measurements. We extract Landsat-8 imagery for 2015 from Google Earth Engine and embed each community tile through Prithvi-EO. We then train a SAE on a held-out Ghana national grid, and pool the active landscape atoms with the spectral indices and the 24 survey covariates; NEXIS then selects from this pool conditionally, and retained atoms are labelled post-hoc by contrasting their top and bottom activating tiles with a large VLM (Bai et al., 2025).

Results. NEXIS identifies two satellite-derived direct modifiers interpreted as ephemeral waterways and closed-canopy forest (Figure 6), and no demographic modifiers. In the small number of communities endowed with seasonal water corridors, the estimated program effect is several times larger than the overall average; a similarly amplified pattern appears for the rare forest-patch communities in this predominantly arid Savannah region. Our interpretation is *complementarity*-based (De Mel et al., 2008; Prifti et al., 2020): water-adjacent smallholders can direct transfers into irrigation-season agriculture, while forest-edge households can layer cash onto existing non-timber livelihood strategies unavailable elsewhere. We further support the former hypothesis by tracking cropland expansion and denser riparian vegetation from 2015 to 2017 from satellite imagery VLM interpretations. The absence of demographic discoveries is itself another informative finding suggesting invariant impact across household behavior within the target subpopulation, already very homogeneous by program design.

7. Conclusion

Identifying causal and interpretable HTE characterizations is a crucial milestone to design trustworthy policies and optimize interventions accordingly, from economics to medicine. In this work, we show that this target is finally within reach thanks to more extensive measurements and better model-based representations, introducing the first algorithm for the task. It represents a paradigm shift from static HTE estimation on survey data to extensive exploration among different and unstructured modalities. Our first deployment on two anti-poverty programs already shows novel evidence for environmental impact heterogeneity explanations previously ignored, leading to practical guidelines and theories for the next iterations of the programs. Overall, the main limitations lie in our reliance

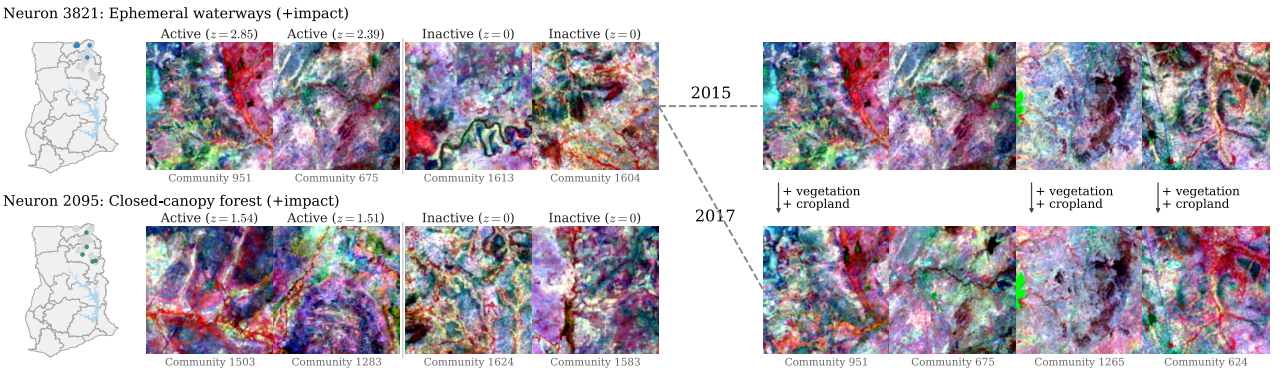


Figure 6. Left: Satellite-derived discoveries on LEAP1000 programme impact heterogeneity by NEXIS, with VLM (Bai et al., 2025) interpretations by contrasting top- and bottom-activating Landsat tiles in false-color composite. *Right*: Ephemeral waterway-active communities temporal evolution in land-use showing cropland extension.

on assumptions, particularly concerning measurement sufficiency, representation alignment, and statistical-test validity. Measurement sufficiency cannot be tested directly without auxiliary experiments, representation alignment is only testable with supervision, and CATE-equivalence testing needs to be well-specified and data-efficient at the same time. If any of these assumptions is unsatisfied, our method loses any causal interpretation guarantee, which is why any result should be carefully interpreted and experimentally validated without blind trust on these assumptions.

Acknowledgements

Riccardo Cadei is supported by a Google Research Scholar Award and a Google-initiated gift to Francesco Locatello. Frank Otchere, Nyasha Tirivayi and Gustavo Angeles Tagliaferro are generously funded by the Swedish International Development Cooperation Agency (Sida) as UNICEF Innocenti researchers. We gratefully acknowledge the Government of Ghana's Ministry of Gender, Children and Social Protection (MoGCSP), the LEAP Management Secretariat (LMS), the UNICEF Ghana country office, the Institute of Statistical, Social and Economic Research (ISSER), and the Navrongo Health Research Centre (NHRC) for their vital support in implementing and coordinating the evaluation on the LEAP 1000 program. This evaluation was generously funded by the United States Agency for International Development (USAID).

References

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1), 2010.

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

Athey, S. and Imbens, G. W. The state of applied econometrics: Causality and policy evaluation. *Journal of Economic perspectives*, 31(2):3–32, 2017.

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Bargagli-Stoffi, F. J. and Melikechi, O. Causal stability selection. *arXiv preprint arXiv:2605.09300*, 2026.

Bargagli-Stoffi, F. J., Cadei, R., Lee, K., and Dominici, F. Causal rule ensemble: Interpretable discovery and inference of heterogeneous treatment effects. *arXiv preprint arXiv:2009.09036*, 2020.

Bargagli-Stoffi, F. J., De Witte, K., and Gnecco, G. Heterogeneous causal effects with imperfect compliance: A bayesian machine learning approach. *The Annals of Applied Statistics*, 16(3):1986–2009, 2022.

Benjamin, J. D., de la Torre, C., and Musumeci, J. Controlling the incentive problems in real estate leasing. *The Journal of Real Estate Finance and Economics*, 10(2): 177–191, 1995.

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., and Saunders, W. Language models can explain neurons in language models. OpenAI tech report, 2023.

Blattman, C., Fiala, N., and Martinez, S. Generating skilled self-employment in developing countries: Experimental evidence from uganda. *The Quarterly Journal of Economics*, 129(2):697–752, 2014.

Boileau, P., Leng, N., Hejazi, N. S., van der Laan, M., and Dudoit, S. A nonparametric framework for treatment effect modifier discovery in high dimensions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(1):157–185, 2025.

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

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Burke, M., Driscoll, A., Lobell, D. B., and Ermon, S. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021.

Bussmann, B., Nabeshima, N., Karvonen, A., and Nanda, N. Learning multi-level features with matryoshka sparse autoencoders. *arXiv preprint arXiv:2503.17547*, 2025.

Cadei, R., Lindorfer, L., Cremer, S., Schmid, C., and Locatello, F. Smoke and mirrors in causal downstream tasks. *arXiv preprint arXiv:2405.17151*, 2024.

Cadei, R., Demirel, I., De Bartolomeis, P., Lindorfer, L., Cremer, S., Schmid, C., and Locatello, F. Prediction-powered causal inferences. *arXiv preprint arXiv:2502.06343*, 2025.

Callaway, B. and Sant'Anna, P. H. Difference-in-differences with multiple time periods. *Journal of econometrics*, 225 (2):200–230, 2021.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chalupka, K., Eberhardt, F., and Perona, P. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

Chanin, D., Wilken-Smith, J., Dulka, T., Bhatnagar, H., Golechha, S., and Bloom, J. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., and Zamora, P. Do labor market policies have displacement effects? evidence from a clustered randomized experiment. *The quarterly journal of economics*, 128(2):531–580, 2013.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

De Mel, S., McKenzie, D., and Woodruff, C. Returns to capital in microenterprises: evidence from a field experiment. *The quarterly journal of Economics*, 123(4):1329–1372, 2008.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Egger, D., Haushofer, J., Miguel, E., Niehaus, P., and Walker, M. General equilibrium effects of cash transfers: experimental evidence from kenya. *Econometrica*, 90(6):2603–2643, 2022.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

Gail, M. and Simon, R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, pp. 361–372, 1985.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Ghana LEAP 1000 Evaluation Team. Ghana LEAP 1000 Programme: Endline Evaluation Report. Technical report, UNICEF Office of Research – Innocenti and University of North Carolina at Chapel Hill, June 2018. URL https://transfer.cpc.unc.edu/wp-content/uploads/2021/04/LEAP1000_Report_Final-2019-for-dissemination.pdf.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. Array programming with numpy. *nature*, 585(7825):357–362, 2020.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Hines, O., Diaz-Ordaz, K., and Vansteelandt, S. Variable importance measures for heterogeneous causal effects. *arXiv preprint arXiv:2204.06030*, 2022.

Jerzak, C. T., Johansson, F., and Daoud, A. Image-based treatment effect heterogeneity. *Proceedings of Machine Learning Research vol*, 213:1–22, 2023.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Lee, D. S. and Lemieux, T. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010.

Liu, Z., Luo, P., Wang, X., and Tang, X. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15 (2018):11, 2018.

Lopez-Paz, D. *The Invariance Principle*. The MIT Press, Cambridge, MA, 2026. ISBN 9780262053341. URL https://mitpress.mit.edu/9780262053341/the-invariance-principle/.

Mairal, J., Bach, F., Ponce, J., and Sapiro, G. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pp. 689–696, 2009.

Makelov, A., Lange, G., and Nanda, N. Towards principled evaluations of sparse autoencoders for interpretability and control. *arXiv preprint arXiv:2405.08366*, 2024.

McKinney, W. et al. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

Mencattini, T., Cadei, R., and Locatello, F. Exploratory causal inference in saence. *arXiv preprint arXiv:2510.14073*, 2025.

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

National Development Planning Commission. Implementation of the Ghana Shared Growth and Development Agenda (GSGDA) II, 2014–2017: 2015 Annual Progress Report. Technical report, National Development Planning Commission (NDPC), Republic of Ghana, December 2016. URL https://ndpc.gov.gh/media/National_Annual_Progress_Report_2015.pdf.

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

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Pach, M., Karthik, S., Bouniot, Q., Belongie, S., and Akata, Z. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*, 2025.

Paillard, J., Lobo, A. R., Kolodyazhniy, V., Thirion, B., and Engemann, D. A. Measuring variable importance in heterogeneous treatment effects with confidence. *arXiv preprint arXiv:2408.13002*, 2024.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

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

Prifti, E., Daidone, S., Pace, N., and Davis, B. Heterogeneous impacts of cash transfers on farm profitability. evidence from a randomised study in lesotho. *European Review of Agricultural Economics*, 47(4):1531–1558, 2020.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Richardson, T. S. and Robins, J. M. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica: journal of the Econometric Society*, pp. 931–954, 1988.

Rothman, K. J., Greenland, S., and Walker, A. M. Concepts of interaction. *American journal of epidemiology*, 112(4): 467–470, 1980.

Roy, A. D. Some thoughts on the distribution of earnings. *Oxford economic papers*, 3(2):135–146, 1951.

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

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

Seabold, S., Perktold, J., et al. Statsmodels: econometric and statistical modeling with python. *scipy*, 7(1):92–96, 2010.

Shah, R. D. and Peters, J. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514, 2020.

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

Spirtes, P., Glymour, C. N., and Scheines, R. *Causation, prediction, and search*. MIT press, 2000.

Szwarcman, D., Roy, S., Fraccaro, P., Gíslason, O. E., Blumenstiel, B., Ghosal, R., De Oliveira, P. H., de Sousa Almeida, J. L., Sedona, R., Kang, Y., et al. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Tsamardinos, I., Aliferis, C. F., Statnikov, A. R., and Statnikov, E. Algorithms for large scale markov blanket discovery. In *FLAIRS*, volume 2, pp. 376–81, 2003.

VanderWeele, T. J. On the distinction between interaction and effect modification. *Epidemiology*, 20(6):863–871, 2009.

VanderWeele, T. J. and Robins, J. M. Four types of effect modification: a classification based on directed acyclic graphs. *Epidemiology*, 18(5):561–568, 2007.

Veitch, V., Sridhar, D., and Blei, D. Adapting text embeddings for causal inference. In *Conference on uncertainty in artificial intelligence*, pp. 919–928. PMLR, 2020.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

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

Yao, D., Huang, S., Cadei, R., Zhang, K., and Locatello, F. The third pillar of causal analysis? a measurement perspective on causal representations. *arXiv preprint arXiv:2505.17708*, 2025a.

Yao, D., Rancati, D., Cadei, R., Fumero, M., and Locatello, F. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, 2025b.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Appendix

Table of contents

A Proof of Theorem 4.1

We first formally introduce the two background assumptions used in the proof but not explicitly stated in the main text, then restate Theorem 4.1 for self-containedness.

Assumption 4 (Mean faithfulness)**.** *The joint distribution of $(\boldsymbol{W}, \boldsymbol{Z})$ is faithful to the modifier taxonomy of Section 2.1 (illustrated in Figure 2): every conditional independence holding in the distribution corresponds to a d-separation property of the taxonomy. Furthermore, conditional dependences involving τ are visible at the conditional-mean level, i.e.,*

$$\tau \not\perp\!\!\!\perp V \mid U \implies \mathbb{P}(\mathbb{E}[\tau \mid U, V] \neq \mathbb{E}[\tau \mid U]) > 0, \tag{11}$$

for any random variables U, V.

The first part is the standard faithfulness postulate, restricted to the joint of the four modifier classes and the dictionary; it is what aligns the conditional-independence calculus on $(\boldsymbol{W}, \boldsymbol{Z})$ with the structural reading provided by the modifier taxonomy of Section 2.1. The full modifier vector $\boldsymbol{W}$ enters the clause, not just the direct modifiers $\boldsymbol{W}^{\mathrm{dir}}$, because paths between modifier classes and $\boldsymbol{Z}$-coordinates can route through any modifier as an intermediate or collider node, and faithfulness operates on the joint. The second part is a weaker, mean-level postulate on τ: it rules out Simpson-style cancellations where a genuine conditional dependence of τ on some variable washes out exactly at the conditional-mean level. It is strictly weaker than full faithfulness on the joint $(\tau, \boldsymbol{W}, \boldsymbol{Z})$ [50], since it constrains only the first moment of τ, which is the only moment our analysis manipulates.

Assumption 5 (Test validity and consistency)**.** *For any fixed $S \subseteq [m]$ and $j \in [m] \setminus S$:*

(a) under $H_0(j \mid S)$ (Equation 8), the p-value $p_j(S)$ is asymptotically uniform on $[0, 1]$;

(b) under any fixed alternative to $H_0(j \mid S)$, $\mathbb{P}(p_j(S) < t) \to 1$ as $n \to \infty$ for every $t > 0$.

The assumption is agnostic to the specific test instantiation. By default we adopt a linear treatment-interaction test, which yields a uniform p-value under the null and consistent rejection under linear-in-$\boldsymbol{Z}$ alternatives; on our benchmarks it reaches Assumption 5(b) at substantially smaller sample sizes than fully nonparametric variants, since SAE codes are designed to fire on distinct concepts and the residual interaction in Y is well-approximated linearly in the active codes. For settings with strong nonlinearity in the residual interaction, the linear test can be replaced with a doubly-robust GCM-style residual test [72], which satisfies Assumption 5 under weaker structural conditions. See Appendix B for both constructions, and Appendix C for their ablations.

Theorem (Causal Identification)**.** *Given a randomized experiment $\{(\boldsymbol{z}_i, t_i, y_i)\}_{i=1}^n$, under Principal Alignment, mean-faithfulness, and test validity, NEXIS' outcome $\widehat{\mathcal{S}}_n$ satisfies:*

$$\liminf_n \mathbb{P}(\widehat{\mathcal{S}}_n = \mathcal{S}^\star) \geq 1 - \alpha;$$

where, additionally assuming Measurement and Representation Sufficiency:

$$\tau^{do}(\boldsymbol{W}^{\mathrm{dir}}) = \tau(\boldsymbol{W}^{\mathrm{dir}}) \;=\; \mathbb{E}[\tau \mid \boldsymbol{Z}^{\mathcal{S}^\star}] \quad \textit{a.s.}$$

The proof proceeds in three steps:

i. **Markov-blanket formulation:** Under Principal Alignment and mean-faithfulness, $\mathcal{S}^\star$ is a Markov blanket of the CATE on the learned dictionary $[m]$: any superset $S \supseteq \mathcal{S}^\star$ is sufficient, and every coordinate $j_k \in \mathcal{S}^\star$ is non-redundant (Lemma A.1). *Measurement and Representation Sufficiency are not invoked here:* the lemma is a structural fact about the dictionary alone.

ii. **Selection Consistency:** Combined with test validity, the lemma reduces NEXIS to forward–backward Markov-blanket discovery on the CATE over $\boldsymbol{Z}$, for which a standard IAMB-style

argument [31, 32] delivers asymptotic recall and asymptotic conditional precision, hence selection consistency.

iii. **Causal HTE Identification:** Measurement and Representation Sufficiency, paired with the definition of $\boldsymbol{W}^{\mathrm{dir}}$ as the direct effect modifiers, upgrade the structural recovery into a causal-identification statement.

A.1 Markov-blanket formulation

Lemma A.1. *Under Principal Alignment and mean-faithfulness, the principal proxies $\mathcal{S}^\star$ form a Markov-blanket over the dictionary coordinates $[m]$, i.e.,*

(a) (Sufficiency) *For every $S \supseteq \mathcal{S}^\star$ (with $S \subseteq [m]$),*

$$\mathbb{E}[\tau \mid \boldsymbol{Z}^S] = \mathbb{E}[\tau \mid \boldsymbol{Z}^{\mathcal{S}^\star}] = \mathbb{E}[\tau \mid \boldsymbol{Z}] \quad \textit{a.s.} \tag{12}$$

(b) (Non-redundancy) *For every $S \subsetneq \mathcal{S}^\star$ (with $S \subseteq [m]$) and every $j_k \in \mathcal{S}^\star \setminus S$,*

$$\mathbb{P}\left(\mathbb{E}[\tau \mid \boldsymbol{Z}^{S\cup\{j_k\}}] \neq \mathbb{E}[\tau \mid \boldsymbol{Z}^S]\right) > 0 \tag{13}$$

Proof. Part (a). We show that, for any $R \subseteq [m] \setminus \mathcal{S}^\star$,

$$\tau \;\perp\!\!\!\perp\; \boldsymbol{Z}^R \;\Big|\; \boldsymbol{Z}^{\mathcal{S}^\star}, \tag{14}$$

from which $\mathbb{E}[\tau \mid \boldsymbol{Z}^{\mathcal{S}^\star\cup R}] = \mathbb{E}[\tau \mid \boldsymbol{Z}^{\mathcal{S}^\star}]$ a.s. follows by the mean-level clause of Assumption 4. Equation 12 is obtained by taking $R = [m] \setminus \mathcal{S}^\star$ for the equality to $\mathbb{E}[\tau \mid \boldsymbol{Z}]$, and arbitrary $R \subseteq [m] \setminus \mathcal{S}^\star$ for the equality across intermediate supersets.

Fix any $k \in [r]$. Principal Alignment gives

$$W^{\mathrm{dir},k} \;\perp\!\!\!\perp\; \boldsymbol{Z}^{[m]\setminus\{j_k\}} \;\Big|\; Z^{j_k}, \tag{15}$$

which, since $[m] \setminus \{j_k\} = (\mathcal{S}^\star \setminus \{j_k\}) \cup ([m] \setminus \mathcal{S}^\star)$, can be rewritten as $W^{\mathrm{dir},k} \perp\!\!\!\perp (\boldsymbol{Z}^{\mathcal{S}^\star\setminus\{j_k\}}, \boldsymbol{Z}^{[m]\setminus\mathcal{S}^\star}) \mid Z^{j_k}$. The weak union axiom of conditional independence (a graphoid property valid for any probability distribution) yields

$$W^{\mathrm{dir},k} \;\perp\!\!\!\perp\; \boldsymbol{Z}^{[m]\setminus\mathcal{S}^\star} \;\Big|\; \boldsymbol{Z}^{\mathcal{S}^\star}. \tag{16}$$

Equation 16 holds for each $k \in [r]$ separately. By the composition property valid under the faithfulness clause of Assumption 4, these per-k statements combine into the joint

$$\boldsymbol{W}^{\mathrm{dir}} \;\perp\!\!\!\perp\; \boldsymbol{Z}^{[m]\setminus\mathcal{S}^\star} \;\Big|\; \boldsymbol{Z}^{\mathcal{S}^\star}. \tag{17}$$

By the definition of $\boldsymbol{W}^{\mathrm{dir}}$ as the direct effect modifiers (Section 2.1), τ is a function only of $\boldsymbol{W}^{\mathrm{dir}}$ and the treatment, so any dependence of τ on $\boldsymbol{Z}$ must factor through $\boldsymbol{W}^{\mathrm{dir}}$. Equation 17 then implies Equation 14, since $\boldsymbol{Z}^{[m]\setminus\mathcal{S}^\star}$ provides no additional information about $\boldsymbol{W}^{\mathrm{dir}}$ beyond $\boldsymbol{Z}^{\mathcal{S}^\star}$.

Part (b). Fix $S \subsetneq \mathcal{S}^\star$ and $j_k \in \mathcal{S}^\star \setminus S$. By the effect-modifier definition (footnote on Equation 1), $W^{\mathrm{dir},k}$ is dependent on τ. Principal Alignment singles out Z^{j_k} as the principal proxy of $W^{\mathrm{dir},k}$: any other coordinate of $\boldsymbol{Z}$, including those in $S \subsetneq \mathcal{S}^\star$, fails to encode $W^{\mathrm{dir},k}$ in the same recall-and-precision sense. Consequently, conditioning on $\boldsymbol{Z}^S$ alone leaves τ and Z^{j_k} dependent: $W^{\mathrm{dir},k}$ is not d-separated from Z^{j_k} given $\boldsymbol{Z}^S$. Adding Z^{j_k} to the conditioning set removes this dependence (it becomes a sufficient blocker). By the mean-level clause of Assumption 4, the strict change in distributional dependence translates into a strict change in the conditional mean of τ on a set of positive probability, which is Equation 13. □

Reading Lemma A.1. Part (a) says $\mathcal{S}^\star$ is sufficient on the dictionary: it captures all the heterogeneity-relevant information $\boldsymbol{Z}$ carries about τ, and adding non-principal coordinates does not destroy sufficiency. Part (b) says it is non-redundant on the principal axes: removing any j_k from a sufficient set strictly reduces information. Together, these are exactly the structural facts NEXIS exploits via

forward growth and backward pruning. Crucially, they hold under Principal Alignment and mean-faithfulness alone, regardless of whether the dictionary $\boldsymbol{Z}$ captures the true heterogeneity content of $\boldsymbol{X}$ and how much heterogeneity is measured in $\boldsymbol{X}$; they are facts about the dictionary's internal geometry around $\mathcal{S}^\star$.

A.2 Selection consistency

Proof of Equation 9. We prove (i) asymptotic recall, $\mathbb{P}(\mathcal{S}^\star \subseteq \widehat{\mathcal{S}}_n) \to 1$, then (ii) precision conditional on recall, $\mathbb{P}(\widehat{\mathcal{S}}_n \not\subseteq \mathcal{S}^\star \mid \mathcal{S}^\star \subseteq \widehat{\mathcal{S}}_n) \le \alpha + o(1)$, then combine.

(i) Recall. Consider a round of NEXIS with current selection S satisfying $\mathcal{S}^\star \not\subseteq S$, and let $\bar{S} = [m] \setminus S$. Pick any $j_k \in \mathcal{S}^\star \setminus S$ and let $S^\circ := S \cap \mathcal{S}^\star \subsetneq \mathcal{S}^\star$. Lemma A.1(b) applied to S° and j_k gives

$$\mathbb{P}\left(\mathbb{E}[\tau \mid \boldsymbol{Z}^{S^\circ \cup \{j_k\}}] \neq \mathbb{E}[\tau \mid \boldsymbol{Z}^{S^\circ}]\right) > 0 \tag{18}$$

Adding the non-principal coordinates $S \setminus S^\circ$ to both conditioning sets preserves the strict inequality: by Principal Alignment, no non-principal coordinate substitutes for Z^{j_k} as a blocker between $W^{\mathrm{dir},k}$ and τ, so the dependence between τ and Z^{j_k} given $\boldsymbol{Z}^S$ persists, and by mean-faithfulness it is visible at the conditional-mean level. Hence $H_0(j_k \mid S)$ is false at the population level, and Assumption 5(b) yields

$$\mathbb{P}\big(p_{j_k}(S) < \alpha/|\bar{S}|\big) \to 1. \tag{19}$$

The forward step admits some $j^\star \in \bar{S}$ whenever $\min_{j \in \bar{S}} p_j(S) \le \alpha/|\bar{S}|$, and the minimum is at most $p_{j_k}(S)$. Hence with probability $\to 1$ the selection strictly grows in any round at which $\mathcal{S}^\star \not\subseteq S$.

It remains to check that $\boldsymbol{Z}^{\mathcal{S}^\star}$ is not removed by the backward step. If $j_k \in \mathcal{S}^\star$ is in $\widehat{\mathcal{S}}_n$ at some round, applying Lemma A.1(b) to $S = (\widehat{\mathcal{S}}_n \setminus \{j_k\}) \cap \mathcal{S}^\star$ and the same Principal-Alignment argument extending to the full $\widehat{\mathcal{S}}_n \setminus \{j_k\}$ implies $H_0(j_k \mid \widehat{\mathcal{S}}_n \setminus \{j_k\})$ is false. By Assumption 5(b), the backward gate retains j_k with probability $\to 1$.

Asymptotic termination then follows: define the favorable event $\mathcal{E}_n$ as the intersection over all rounds of the events "the forward step grows whenever $\mathcal{S}^\star \not\subseteq \widehat{\mathcal{S}}_n$" and "the backward step retains every $j_k \in \mathcal{S}^\star$". The number of distinct selections is bounded by 2^m, so this is a finite intersection, and $\mathbb{P}(\mathcal{E}_n) \to 1$. On $\mathcal{E}_n$, the selection grows until $\mathcal{S}^\star \subseteq \widehat{\mathcal{S}}_n$, after which growth may continue (under partial false alternatives) but $\mathcal{S}^\star$ is retained, until a fixed point is reached. Hence $\mathbb{P}(\mathcal{S}^\star \subseteq \widehat{\mathcal{S}}_n) \to 1$.

(ii) Precision conditional on recall. Let $\mathcal{R}_n := \{\mathcal{S}^\star \subseteq \widehat{\mathcal{S}}_n\}$ and $\mathcal{F}_n := \widehat{\mathcal{S}}_n \setminus \mathcal{S}^\star$. Set $s := |\widehat{\mathcal{S}}_n|$. At termination, every $j \in \widehat{\mathcal{S}}_n$ has passed the final backward check $p_j(\widehat{\mathcal{S}}_n \setminus \{j\}) \le \alpha/s$. On $\mathcal{R}_n$, for any $j' \in \mathcal{F}_n$, $\widehat{\mathcal{S}}_n \setminus \{j'\} \supseteq \mathcal{S}^\star$, so by Lemma A.1(a),

$$\mathbb{E}[\tau \mid \boldsymbol{Z}^{\widehat{\mathcal{S}}_n \setminus \{j'\}}] = \mathbb{E}[\tau \mid \boldsymbol{Z}] \quad \text{a.s.}, \tag{20}$$

and $H_0(j' \mid \widehat{\mathcal{S}}_n \setminus \{j'\})$ holds at the population level. By Assumption 5(a), the p-value at this true null is asymptotically uniform, so

$$\mathbb{P}\big(p_{j'}(\widehat{\mathcal{S}}_n \setminus \{j'\}) \le \alpha/s \,\big|\, \widehat{\mathcal{S}}_n, \mathcal{R}_n\big) \le \alpha/s + o(1). \tag{21}$$

A union bound over $j' \in \mathcal{F}_n$ (of size at most s) and marginalisation over $\widehat{\mathcal{S}}_n$ yield

$$\mathbb{P}\big(\mathcal{F}_n \neq \emptyset \,\big|\, \mathcal{R}_n\big) = \mathbb{P}\big(\widehat{\mathcal{S}}_n \not\subseteq \mathcal{S}^\star \,\big|\, \mathcal{R}_n\big) \le \alpha + o(1). \tag{22}$$

(iii) Combination. Putting (i) and (ii) together,

$$\mathbb{P}\big(\widehat{\mathcal{S}}_n = \mathcal{S}^\star\big) \;=\; \mathbb{P}\big(\widehat{\mathcal{S}}_n \subseteq \mathcal{S}^\star \,\big|\, \mathcal{R}_n\big) \cdot \mathbb{P}(\mathcal{R}_n) \;\ge\; (1 - \alpha - o(1)) \cdot \mathbb{P}(\mathcal{R}_n), \tag{23}$$

so $\liminf_n \mathbb{P}(\widehat{\mathcal{S}}_n = \mathcal{S}^\star) \ge 1 - \alpha$. □

Role of the backward step. Without backward elimination, NEXIS would still satisfy recall, since every $j_k \in \mathcal{S}^\star$ is eventually admitted at some forward step under partial conditioning. But precision would be lost at fixed α: a non-principal coordinate j' admitted at some intermediate round, whose conditional null begins to hold only after subsequent additions, would never be re-tested against the full final selection. The backward gate at each round, combined with the termination condition $\widehat{\mathcal{S}}_n = \widehat{\mathcal{S}}_{\mathrm{prev}}$, guarantees that every coordinate present at termination has passed a Bonferroni-corrected test against the full retained set, which is what supplies the precision bound. A vanishing schedule $\alpha_n \to 0$, decaying slowly enough to preserve test power against the fixed alternatives of Lemma A.1(b), upgrades Equation 9 to strict consistency $\mathbb{P}(\widehat{\mathcal{S}}_n = \mathcal{S}^\star) \to 1$. We keep α fixed in practice: a vanishing schedule trades a strictly-stronger asymptotic guarantee for tighter Bonferroni gates and reduced power at any given sample size, while α as a user-set false-discovery budget is a more interpretable knob for practitioners.

A.3 Causal HTE identification

Proof of Equation 10. Chaining Lemma A.1(a) with Representation and Measurement Sufficiency:

$$\mathbb{E}[\tau \mid \boldsymbol{Z}^{\mathcal{S}^\star}] \;=\; \mathbb{E}[\tau \mid \boldsymbol{Z}] \;=\; \mathbb{E}[\tau \mid \boldsymbol{X}] \;=\; \mathbb{E}[\tau \mid \boldsymbol{W}^{\mathrm{dir}}] \quad \text{a.s.} \tag{24}$$

By definition of $\boldsymbol{W}^{\mathrm{dir}}$ as the direct effect modifiers (Section 2.1), no modifier class interacts with the treatment T except through $\boldsymbol{W}^{\mathrm{dir}}$ itself, thus affecting τ. Intervening on $\boldsymbol{W}^{\mathrm{dir}}$ severs only its inbound dependencies (from $\boldsymbol{W}^{\mathrm{ind}}$, $\boldsymbol{W}^{\mathrm{cc}}$, and exogenous noise), and since no other modifier reaches τ except through $\boldsymbol{W}^{\mathrm{dir}}$ itself, no back-door path from $\boldsymbol{W}^{\mathrm{dir}}$ to τ remains [73, 35]. The $\boldsymbol{W}^{\mathrm{dir}} \to \tau$ relation is therefore unconfounded, so

$$\mathbb{E}[\tau \mid do(\boldsymbol{W}^{\mathrm{dir}})] = \mathbb{E}[\tau \mid \boldsymbol{W}^{\mathrm{dir}}] \quad \text{a.s.} \tag{25}$$

Combining Equations 24–25 yields Equation 10. □

Causal interpretation. The CATE is, in itself, a causal estimand, defined as the expected treatment effect among units sharing the same profile. What is not generally causal is its *characterization*, e.g., non-direct modifiers (indirect, proxy, common-cause). Indeed, for a generic characterization, an intervention on a non-direct modifier yields a different effect than conditioning on it, i.e., $\mathbb{E}[\tau|do(\cdot)] \neq \mathbb{E}[\tau|\cdot]$. Theorem 4.1 provides a causal characterization on $\boldsymbol{Z}^{\mathcal{S}^\star}$ at the *joint* level: counterfactual comparisons across bundled configurations of latent direct modifiers are identified, which is the natural object for policy decisions targeting compound conditions (e.g., a community defined jointly by landscape, infrastructure, and demographics).

From joint to marginal identification. Marginal interventional effects on a single direct modifier, $\mathbb{E}[\tau \mid do(W^{\mathrm{dir},k} = w^k)]$, are a strictly stronger target: they require averaging the CATE over the *marginal* distribution of the remaining direct modifiers $\boldsymbol{W}^{\mathrm{dir},-k}$, since intervening on $W^{\mathrm{dir},k}$ severs its inbound dependences without altering the marginal of the others. The natural observational candidate $\mathbb{E}[\tau \mid Z^{j_k}]$ instead averages over the *conditional* distribution of $\boldsymbol{W}^{\mathrm{dir},-k}$ given Z^{j_k}. The two coincide under either (i) mutually independent direct modifiers, so that conditional and marginal distributions agree, or (ii) no inter-modifier interactions in τ, in which case the integration measure is irrelevant. Absent these structural conditions, $\mathbb{E}[\tau \mid Z^{j_k} = z]$ admits a valid reading as a *subpopulation-specific* marginal effect for units with $Z^{j_k} = z$, which is useful for targeting within the observational distribution but does not licence population-invariant marginal-intervention claims.

B Method Details

The main text presents NEXIS in its vanilla form (Algorithm 1): a forward–backward procedure built around a generic CATE equivalence test, gated by a Bonferroni-style multiple-testing correction. Each of these components admits several natural instantiations, and the abstraction is a feature, not a placeholder: any test family satisfying Assumption 5 plugs in, and the multiple-testing gate can be exchanged for any procedure controlling the relevant error rate. We use this section to lay out the variants we consider, all of which are tested in the ablations of Appendix C.4, and to provide a minimal Python implementation. The four axes we cover are (i) the CATE equivalence test (§B.1), (ii) the multiple-testing correction (§B.2), (iii) an optional spectral-gap gate that mitigates residual entanglement in real dictionaries (§B.3), and (iv) the backward-elimination step (§B.4). The implementation snippet of §B.5 uses the simplest combination, the linear test with Bonferroni correction, which is also the default we recommend for routine use based on the controlled experiments.

B.1 CATE-equivalence test

NEXIS is fully characterized by its choice of CATE equivalence test: at each forward and backward step, the procedure asks whether a candidate coordinate $j \in [m]$ carries residual heterogeneity beyond what the current selection $S \subseteq [m] \setminus \{j\}$ already explains, i.e. whether

$$H_0(j \mid S): \quad \mathbb{E}[\tau \mid \boldsymbol{Z}^{S\cup\{j\}}] = \mathbb{E}[\tau \mid \boldsymbol{Z}^{S}] \quad \text{a.s.} \tag{26}$$

Theorem 4.1 only requires that the resulting p-value $p_j(S)$ be asymptotically uniform under the null and consistent against fixed alternatives (Assumption 5); any such test plugs in. We consider three concrete instantiations, spanning the parametric–nonparametric spectrum.

Linear treatment-interaction test (default). The simplest instantiation fits a linear treatment-interaction working model

$$Y \;=\; \alpha + T\beta_T + \boldsymbol{Z}^S\boldsymbol{\gamma}_S + T\cdot\boldsymbol{Z}^S\boldsymbol{\delta}_S + Z^j\beta_j + T\cdot Z^j\delta_j + \varepsilon, \tag{27}$$

and tests $H_0(j \mid S)$ by a partial F-test on (β_j, δ_j). It requires no nuisance estimation, no cross-fitting, and runs in milliseconds per test. Under correctly-specified linearity, Assumption 5 is satisfied. The price is consistency only against alternatives within the linear-in-$\boldsymbol{Z}$ class: heterogeneity that is genuinely nonlinear in Z^j given $\boldsymbol{Z}^S$ (a threshold, a U-shape) can be missed at any sample size. In practice this is rarely the binding concern: when $\boldsymbol{Z}$ is the output of a sparse autoencoder, individual coordinates are already designed to fire on distinct concepts, so the residual interaction in Y is well-approximated linearly in the active codes. The controlled CelebA experiments confirm this empirically: at the same (n, η), the linear test reaches recall 0.95 at roughly *half* the sample size (or effect magnitude) of the more flexible alternatives below (see Figure 9 and the discussion in §C.4). We adopt it as the default throughout.

Doubly-robust GCM with quadratic nuisance. When linearity is unwarranted, we replace Equation 27 with a doubly-robust pseudo-outcome plus a residual-product test in the spirit of Shah and Peters [72]. The R-learner residualization [74, 13]

$$\widehat{\varphi} \;=\; \frac{(Y - \widehat{m}(\boldsymbol{Z}))(T - e)}{e(1-e)}, \tag{28}$$

with $e := \mathbb{P}(T=1)$ known by design and $\widehat{m}$ a cross-fitted estimator of $\mathbb{E}[Y \mid \boldsymbol{Z}]$, yields a pseudo-outcome with $\mathbb{E}[\widehat{\varphi} \mid \boldsymbol{Z}] = \mathbb{E}[\tau \mid \boldsymbol{Z}]$. The Generalised Covariance Measure (GCM) then regresses $\widehat{\varphi}$ on $\boldsymbol{Z}^S$ and Z^j on $\boldsymbol{Z}^S$, forms the residual product

$$R(j \mid S) \;:=\; \big(\widehat{\varphi} - \widehat{\mathbb{E}}[\widehat{\varphi} \mid \boldsymbol{Z}^S]\big)\cdot\big(Z^j - \widehat{\mathbb{E}}[Z^j \mid \boldsymbol{Z}^S]\big), \tag{29}$$

and tests $\mathbb{E}[R(j \mid S)] = 0$ via the normalized z-statistic

$$T_n(j \mid S) \;=\; \frac{\sqrt{n}\,\bar{R}_n(j \mid S)}{\widehat{\sigma}_n(j \mid S)}. \tag{30}$$

The quadratic variant takes $\widehat{m}$ and the residualization regressions to be quadratic in $\boldsymbol{Z}^S$. This adds curvature without introducing a tuning-heavy nonparametric estimator, and is the natural step up from linear when one suspects mild nonlinearity but wants to keep the model auditable.

Doubly-robust GCM with LightGBM nuisance. The fully nonparametric variant uses gradient-boosted trees [75] for both nuisance regressions, with K-fold cross-fitting to ensure independence between the nuisance estimate and the residual at each unit. Under nuisance rates whose product is $o(n^{-1/2})$, substantially weaker than parametric correctness, $T_n(j \mid S)$ is asymptotically standard normal under the null and the test is consistent against any fixed conditional-mean alternative. This is the most permissive option and the safest one when the analyst has no prior on the form of the heterogeneity. It is also the most expensive: each test requires fitting two boosting models and the per-iteration cost dominates.

Trade-offs and default. The three variants trace the standard parametric–nonparametric trade-off. Linear is fast, easy to implement, and statistically efficient when the true heterogeneity is approximately linear in $\boldsymbol{Z}$; we find this to be the realistic regime for SAE-coded representations and adopt it as the default. The two GCM variants buy robustness to misspecification at a measurable cost in power, roughly a factor of two in n or η on the controlled benchmark (Figure 9). They are the right choice when (a) the analyst has reason to expect nonlinear heterogeneity in the dictionary, or (b) the dictionary itself is unaudited and the working linearity cannot be defended. Switching from linear to a GCM variant is a one-line change in the implementation of §B.5; all other components of NEXIS are unchanged.

B.2 Error control

The forward step accepts the most significant candidate, the backward step removes any retained coordinate that fails its own gate. Both steps require a threshold on $p_j(S)$, and that threshold is what controls the false-discovery behaviour of NEXIS. We consider three options.

No correction. The simplest option compares each $p_j(S)$ directly to α. This makes no adjustment for the fact that the forward step screens $|\bar{S}|$ candidates per round, and at large effect sizes several entangled companions of the principals can cross the gate before the backward step prunes them. The semi-synthetic experiments (Figure 10) show this concretely: precision recovers only at $\eta = 3$, $n = 2000$, against $\eta = 2$, $n = 2000$ for the corrected variants, with visibly more residual false discoveries at small n. Recall is essentially unchanged. The unadjusted gate is therefore not a recall-friendly choice in disguise, it simply fails on precision.

FDR via Benjamini–Hochberg. The Benjamini–Hochberg procedure [54] controls the expected proportion of false discoveries among rejections at level α. At each forward step, candidate p-values are ordered $p_{(1)} \leq \cdots \leq p_{(|\bar{S}|)}$ and the largest k such that $p_{(k)} \leq k\alpha/|\bar{S}|$ is identified; the candidates with p-value at most $p_{(k)}$ are eligible to enter S. The forward step of NEXIS admits the most significant candidate, so in practice this collapses to comparing $\min_j p_j$ against $\alpha/|\bar{S}|$ when only one rejection is sought. The backward step is treated symmetrically against $|S|$. FDR is the natural choice when $|\mathcal{S}^\star|$ is expected to be moderate to large: it pays a smaller multiplicity tax than FWER as the truth set grows.

FWER via Bonferroni. Bonferroni [53] controls the family-wise error rate by comparing each p-value against the conservative threshold $\alpha/|\bar{S}|$ (or $\alpha/|S|$ in the backward step). It is the most stringent of the three and the one used in the proof of Theorem 4.1 to deliver the finite-sample precision bound: the union bound across at most $|S|$ backward checks at termination is what guarantees $\mathbb{P}(\widehat{\mathcal{S}}_n \subseteq \mathcal{S}^\star \mid \mathcal{R}_n) \geq 1 - \alpha$.

Trade-offs and default. Empirically (Figure 10) FDR and FWER are essentially indistinguishable in our setting, which is the regime expected when the truth set is small ($|\mathcal{S}^\star| = 2$) and the candidate pool is large: the Benjamini–Hochberg threshold collapses onto the Bonferroni threshold for the leading discoveries. We therefore adopt FWER as the default, both because it is the more conservative of the two and because it underwrites the precision statement of Theorem 4.1 as written. In larger truth-set regimes FDR is likely the better operating point, and the swap is again a one-line change.

B.3 Spectral-gap gate

Principal Alignment (Assumption 3) asks each direct modifier to be summarized by one dominant coordinate, with no concept-specific signal scattered across the others. This is a population-level idealisation. Real learned dictionaries exhibit *partial* entanglement [30]: a proxy coordinate $j' \neq j_k$ correlated with the principal coordinate j_k of $W^{\mathrm{dir},k}$ can retain a small residual interaction signal even after j_k has entered S. At large n, this residual can squeak through the conditional gate without the backward step always catching it, because experimental power amplifies finite-resolution misalignment in exactly the same way it amplifies genuine signals.

The heuristic. The fix we adopt is a simple relative-magnitude check on top of the existing p-value gate. Let

$$|T_n(j^\star \mid S)| \;\geq\; \rho \cdot \min_{j \in S} |T_n(j \mid S \setminus \{j\})|, \qquad \rho \in (0, 1], \tag{31}$$

where T_n is the test statistic of the chosen conditional-independence test. A candidate $j^\star$ admitted by the p-value gate is retained only if its conditional t-statistic is within a factor ρ^{-1} of the *weakest* coordinate already in S. The intuition is residual-by-construction: two true direct modifiers compete on comparable magnitudes; a residual proxy of an already-selected principal, by contrast, can only carry a fraction of the principal's signal, and that fraction shrinks with the alignment quality of the dictionary. The gate makes this asymmetry explicit.

Substantive reading. Because both $\sqrt{n}$ and the variance estimate $\widehat{\sigma}_n$ cancel in the ratio

$$\frac{|T_n(j^\star \mid S)|}{|T_n(j \mid S \setminus \{j\})|}, \tag{32}$$

ρ coincides with a ratio of estimated CATE contrasts: it is the minimum acceptable ratio between a new candidate's estimated direct effect and the weakest direct effect already selected. Its inverse ρ^{-1} is interpretable as the analyst's prior on how widely the magnitudes of true direct modifiers can spread. $\rho = 0.5$ accepts a factor-of-two spread, $\rho = 0.2$ a factor-of-five spread, $\rho = 0.8$ a tight spread within 25%. This is a real knob, not a tuning artefact: it encodes a substantive claim about the experiment.

Behaviour in the ablations. Figure 11 sweeps $\rho \in \{0, 0.2, 0.5, 0.8\}$, where $\rho = 0$ disables the gate. The behaviour matches the heuristic exactly. Low ρ (0 and 0.2) admits correlated companions of the principal coordinates whenever the conditional p-value gate is loose enough, hurting precision in the large-η regime without any visible recall benefit. High ρ (0.8) blocks coordinates whose conditional statistic is comparable but slightly smaller than that of an already-selected principal, occasionally including the second principal itself, and erodes recall. The default $\rho = 0.5$ sits at the joint optimum, which is consistent with its substantive reading: in the absence of a prior tightening expectation, accepting a factor-of-two spread is a sensible neutral choice.

Practical guidance. We default to $\rho = 0.5$ and recommend a sensitivity sweep over $\rho \in \{0.2, 0.5, 0.8\}$ on real applications. Stable selection across this range is empirical evidence that the spectral gap is not the binding constraint and that $\widehat{\mathcal{S}}_n$ is driven by the conditional-independence structure rather than by the gating heuristic. When sensitivity is observed, we report $\widehat{\mathcal{S}}_n$ at the most conservative ρ for which the procedure remains stable. Setting $\rho = 0$ recovers the vanilla NEXIS of Algorithm 1 and is appropriate when the dictionary has been explicitly verified to be near-orthogonal on a held-out grid.

B.4 Backward elimination

The backward step re-tests every retained coordinate against the rest of the current selection at each round, removing any that has become redundant. Two questions are worth disentangling: whether it is *necessary* for the theoretical guarantee, and whether it is *useful* in practice.

Theoretical role. The backward step is the source of the finite-sample precision bound in Theorem 4.1. At termination, every coordinate in $\widehat{\mathcal{S}}_n$ has passed a Bonferroni-corrected check against the rest of the selection, and a union bound across at most $|\widehat{\mathcal{S}}_n|$ such checks delivers the precision

statement. Without the backward step, a coordinate that passed an early forward gate under partial conditioning, before later additions made it redundant, would persist in $\widehat{S}_n$ and the precision claim at fixed α would not be available.

Empirical role. In the controlled experiments of Figure 12, the forward–backward and forward-only variants are visually indistinguishable across all DGP conditions. With $|\mathcal{S}^\star| = 2$ and a near-orthogonal $\boldsymbol{Z}$, the forward pass already returns a pair that is conditionally independent of every remaining coordinate, and the backward gate has nothing to prune. The empirical contribution of the backward step grows with $|\mathcal{S}^\star|$ and with dictionary entanglement; the present semi-synthetic setting is on the easy end of both axes.

Practical recommendation. The backward step can be removed when the truth set is known to be small and the dictionary has been verified to be near-orthogonal: doing so saves a modest amount of compute and does not visibly change the output. We retain it as the default for two reasons. First, it underwrites the precision guarantee of Theorem 4.1 as stated. Second, its incremental cost is negligible (one extra round of tests per iteration, against a candidate pool of size $|S|$ rather than $|\bar{S}|$), while its incremental benefit grows precisely in the regimes where one cannot rule out either a larger truth set or non-trivial entanglement, that is, the regimes encountered in any new applied study.

B.5 Minimal Python Implementation Snippet

We provide a self-contained Python implementation of NEXIS in its default configuration: the linear treatment-interaction test of Equation 27 with Bonferroni multiple-testing correction, no spectral-gap gate ($\rho = 0$), and backward elimination enabled. The implementation relies on `pandas` [76] for tabular operations, `NumPy` [77] and `SciPy` [78] for numerical routines, and `statsmodels` [79] for the OLS partial F-test. Switching to a GCM variant of §B.1, swapping Bonferroni for Benjamini–Hochberg, or enabling the spectral-gap gate of §B.3 is a localised modification of the routines below.

Neural Exposure Interaction Search. Iterative forward–backward selection until a fixed point is reached. The procedure is parameterized by a generic CATE equivalence test `cci_test`, which returns a p-value per candidate.

```
def NEXIS(T, Y, Z, alpha=0.05, cci_test=cci_test_linear):
    S, S_prev = [], None
    while S != S_prev:
        S_prev = list(S)
        # forward step: admit the most significant candidate under
            Bonferroni
        bar_S = [c for c in Z.columns if c not in S]
        tests = cci_test(T, Y, Z, S, candidates=bar_S)
        j_star = tests["p_value"].idxmin()
        if tests.loc[j_star, "p_value"] <= alpha / len(bar_S):
            S.append(j_star)
        # backward step: prune any selected coordinate now redundant
        for j in list(S):
            tests = cci_test(T, Y, Z, [c for c in S if c != j],
                candidates=[j])
            if tests.loc[j, "p_value"] > alpha / len(S):
                S.remove(j)
    return S
```

Linear treatment-interaction test (default). Fits Equation 27 via OLS and tests the candidate's main effect and treatment interaction jointly via a partial F-test.

```
import pandas as pd
import statsmodels.api as sm

def cci_test_linear(T, Y, Z, S, candidates):
    df = Z[S].copy() if S else pd.DataFrame(index=Z.index)
    df["_T"] = T.astype(int)
```

```
    for s in S:
        df[f"_T_x_{s}"] = df["_T"] * df[s]
    out = []
    for j in candidates:
        df_j = df.copy()
        df_j[j] = Z[j]
        df_j[f"_T_x_{j}"] = df_j["_T"] * df_j[j]
        model = sm.OLS(Y, sm.add_constant(df_j)).fit()
        f_test = model.f_test([f"{j} = 0", f"_T_x_{j} = 0"])
        out.append((j, float(f_test.pvalue)))
    return pd.DataFrame(out, columns=["candidate", "p_value"]).
        set_index("candidate")
```

C Semi-synthetic experiments (CelebA)

This section reports the controlled empirical validation of NEXIS announced in Section 2.2 and Theorem 4.1. The goal is to verify, on a setting where the ground-truth direct-modifier set $\mathcal{S}^\star$ is known by construction, that NEXIS recovers it under the regime predicted by the theory and that the practical-guidance defaults of Appendix B hold up across DGP conditions and design choices. We use CelebA [51] as a high-dimensional, visually realistic source of pre-treatment observations $\boldsymbol{X}$, embed each image with a foundation model, learn a sparse-autoencoder dictionary on the embeddings, and inject a known treatment-modification structure on top of two fixed binary attributes. Because the ground-truth modifiers are fixed but *not* given to NEXIS, the evaluation tests both Principal Alignment of the SAE dictionary and the conditional-selection behaviour of the algorithm itself.

C.1 Experimental setup

Foundation-model encoder. We embed each of the 202,599 CelebA images with **SigLIP** [47], a contrastively pretrained ViT, taking the patch-level tokens (729 patches per image, 1,152-dimensional). All embeddings are pre-computed once and cached, decoupling the SAE step from the (expensive) encoder forward pass.

Sparse autoencoder. We train a **TopK SAE** [26] of hidden width $m = 13{,}824$ ($= 12 \times 1{,}152$) on the per-patch SigLIP tokens ($\sim 1.48 \times 10^8$ training vectors per epoch), under the principal-alignment-oriented protocol of Cadei et al. [60]. Two sparsity levels are evaluated, $k = 5$ and $k = 20$ active features per image, and for each level we expose two views of the representation:

- $\boldsymbol{Z}$, the sparse *post*-TopK codes (near-orthogonal, average $L_0 = k$);
- $\boldsymbol{Z}_{\mathrm{pre}}$, the dense, continuous *pre*-activations (correlated, no sparsity constraint).

The main reported setting is $k = 20$ on $\boldsymbol{Z}$; the alternatives ($k = 5$ on $\boldsymbol{Z}$, and $k = 20$ on $\boldsymbol{Z}_{\mathrm{pre}}$) are reported as ablations in Appendix C.3.

Data-generating process. We fix two CelebA attributes, $W_1 =$WEARING_HAT ($\sim 5\%$ prevalence) and $W_2 =$EYEGLASSES ($\sim 7\%$ prevalence), as the latent direct effect modifiers. For each unit i we draw *i.i.d.* from $W_1 \sim \mathrm{Bernoulli}(\hat{p}_{W_1})$, $W_2 \sim \mathrm{Bernoulli}(\hat{p}_{W_2})$, $T \sim \mathrm{Bernoulli}(0.5)$, and sample without replacement from the CelebA bucket an image $\boldsymbol{x}_i$, matching $(w_{1,i}, w_{2,i})$ and pre-compute its representation $\boldsymbol{z}_i = \psi(\boldsymbol{x}_i)$. The outcome is sampled from:

$$Y \;=\; \beta_{W_1} W_1 + \beta_{W_2} W_2 \;+\; T \cdot \big[\tau_0 + \eta \cdot (\gamma_{W_1} W_1 + \gamma_{W_2} W_2)\big] \;+\; \varepsilon, \tag{33}$$

with $\tau_0 = 0.5$, $\gamma_{W_1} = +1$, $\gamma_{W_2} = -1$, $\beta_{W_1} = 0.3$, $\beta_{W_2} = -0.2$, $\sigma_\varepsilon = 1$. The scalar η scales the heterogeneous component of the treatment effect and is the dial we use to traverse the power frontier.

Ground-truth labels. The ground-truth direct-modifier set $\mathcal{S}^\star$ is defined empirically as the SAE neurons most selective for each attribute over the full CelebA dataset. Concretely, for each neuron j and attribute $A \in \{W_1, W_2\}$, we sweep thresholds over Z^j_{pre} and report the best-threshold $F_1(j, A)$; we use $\boldsymbol{Z}_{\mathrm{pre}}$ rather than $\boldsymbol{Z}$ because the continuous pre-activations give a smoother, threshold-free alignment score that better reflects each neuron's intrinsic selectivity. The ground-truth principal coordinate for A is then $\arg\max_j F_1(j, A)$. This labelling is computed once on all 202,599 images (no sampling, no holdout) and held fixed throughout. Under the main setting ($k = 20$), $\mathcal{S}^\star = \{5348, 5537\}$ (dim 5348 $\to W_1$, dim 5537 $\to W_2$); under $k = 5$ the principal coordinates shift to $\{7044, 5732\}$, reflecting the different sparsity structure.

Sweeps and reporting. We sweep two axes, holding the other fixed:

- *Effect-size sweep:* $\eta \in \{1, 2, \ldots, 10\}$, with $n \in \{500, 2000\}$ fixed.
- *Sample-size sweep:* $n \in \{50, 100, 200, 350, 500, 750, 1000, 2000, 3500, 5000, 10{,}000\}$, with $\eta \in \{2, 5\}$ fixed.

Each (η, n) cell is repeated over 50 random seeds; reported curves are means with $\pm 1.96\,\mathrm{SE}$ shaded bands. Performance is reported as precision, recall, and intersection-over-union (IoU) of $\widehat{\mathcal{S}}_n$ against

$\mathcal{S}^\star$. Every figure in this section uses the same 4×3 layout: rows traverse the four DGP conditions (two n-sweeps at $\eta \in \{5, 2\}$, then two η-sweeps at $n \in \{2000, 500\}$), and the three columns show PRECISION | RECALL | IOU. Showing all four DGP conditions in a single figure lets the reader read both the power frontier (effect size and sample size) and the design-choice sensitivity simultaneously.

Baselines. We compare NEXIS to three marginal coordinate-wise variants that test each $j \in [m]$ for marginal effect modification (treating it as if it were the only modifier in the dictionary): *Marginal Testing* (no multiple-testing correction), *Marginal Testing (FDR)* [54], and *Marginal Testing (FWER)* (Bonferroni). They share NEXIS's per-test statistic and threshold but skip the conditional, sequential structure, the precise design that, by Section 3, is statistically inadequate for direct-modifier identification on a learned representation.

Default NEXIS configuration. Unless stated otherwise, NEXIS is run with $\alpha = 0.05$, the linear treatment-interaction test of Equation 27, FWER (Bonferroni) correction at each forward and backward step, spectral gap $\rho = 0.5$, and backward elimination enabled. These are the defaults recommended in Appendix B.

C.2 Reference results: NEXIS vs. marginal baselines

Figure 6 reports the reference comparison and serves as the visual anchor for all subsequent ablations: each later figure should be read as a deviation from this one along a single axis. Three observations are worth surfacing.

(i) NEXIS attains the recovery regime predicted by Theorem 4.1. Reading along the rows, recall crosses 0.95 at $\eta = 2$ when $n = 2000$ and at $\eta = 5$ when $n = 500$. Precision tracks recall with a slight lag at the smallest n or η, then saturates at 1. The IoU curve, which simultaneously penalises both error directions, is essentially the recall curve once one is past the under-powered regime.

(ii) The experimental power paradox of Section 3 is visible cleanly. Marginal baselines pass the recall bar comparably early (and Marginal Testing without correction passes it the earliest, by being a strictly looser gate). Their precision, however, never recovers: the unadjusted variant collapses below 0.1 at moderate η or n, and the FDR/FWER variants peak around 0.6–0.8 at the smallest informative n before *deteriorating* as η or n grows. This is the predicted regime: more power surfaces more entangled companions of $\mathcal{S}^\star$ as marginally heterogeneous, none of which any marginal procedure has the structure to reject.

(iii) DGP sensitivity is graceful, not catastrophic. Comparing the $\eta = 5$ and $\eta = 2$ rows, NEXIS shifts its onset by roughly a factor of 4 in n, with no qualitative change in shape. Comparing the $n = 2000$ and $n = 500$ rows, NEXIS shifts its η onset by roughly a factor of 2 and saturates at a slightly lower precision plateau (~ 0.75 vs. ~ 1.0) at $n = 500$ once η is very large, a finite-sample residual companion entering at high SNR, consistent with the relaxed Principal Alignment discussion of Appendix B.3.

C.3 Model ablations: SAE design choices

The next two ablations vary the choice of dictionary, holding the NEXIS configuration at its default. We evaluate two competing SAE design decisions: the TopK sparsity level k, and whether to use the sparse post-TopK codes $\boldsymbol{Z}$ or the dense pre-activations $\boldsymbol{Z}_{\mathrm{pre}}$. Both deviate from the main setting along a single axis and are read against Figure 6.

Sparsity level: $k = 5$ vs. $k = 20$. Figure 7 repeats the reference comparison on the $k = 5$ dictionary. Recall thresholds are essentially unchanged ($\eta = 2$ at $n = 2000$; $n = 750$ at $\eta = 5$); precision, however, requires noticeably more power to saturate ($\eta = 2$ at $n = 2000$ in the effect-size sweep; $n = 5000$ at $\eta = 5$ in the sample-size sweep, against $n = 2000$ for the main setting). With only five active features per image, attribute-specific signal is forced to spread across several correlated coordinates, so the conditional selection retains a slightly larger set in which a few non-principal companions persist before the backward gate prunes them. $k = 20$ is the better operating point.

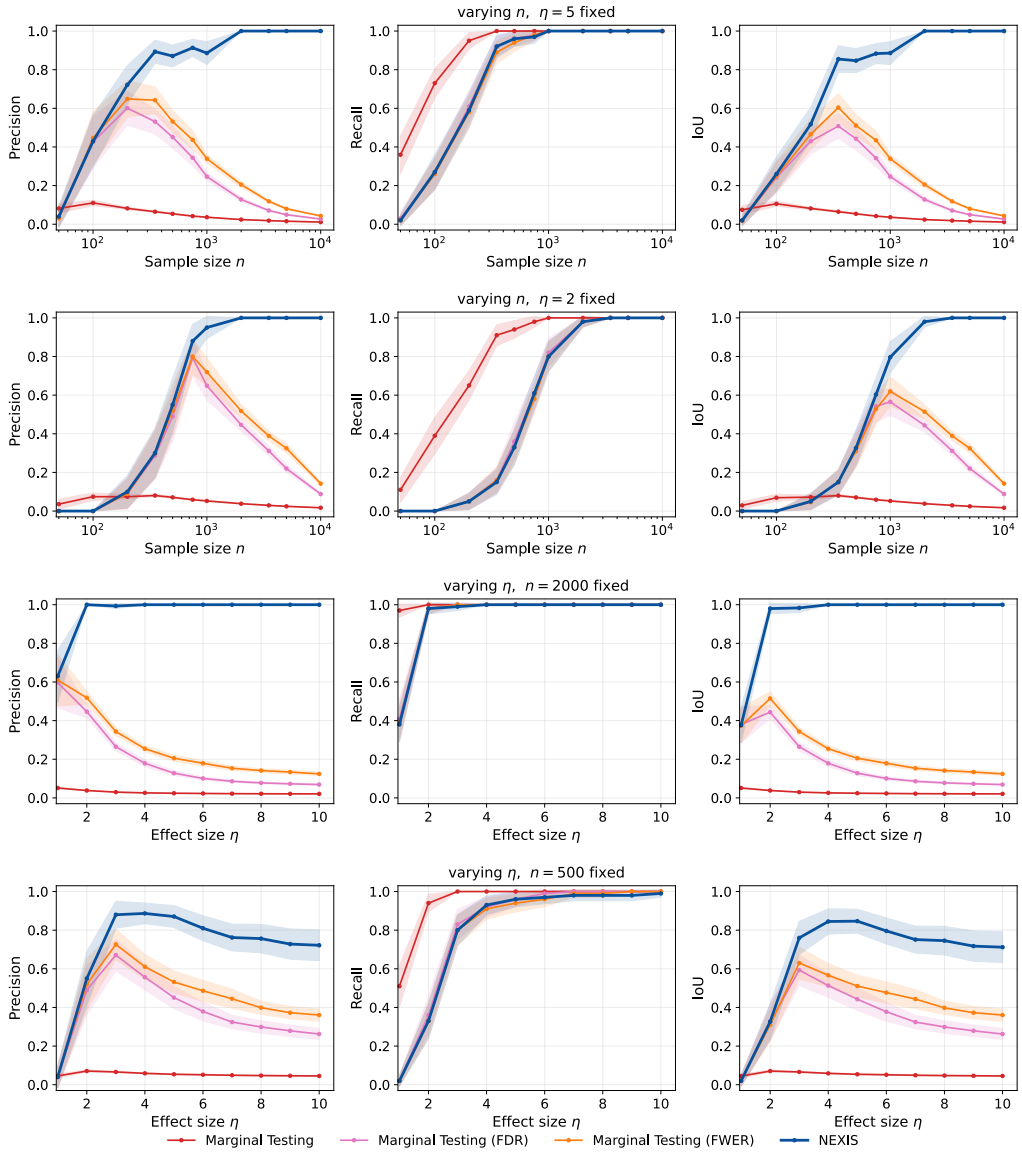


Figure 6: **Reference figure.** NEXIS vs. marginal baselines on the main setting ($k = 20$, $\boldsymbol{Z}$, FWER, $\rho = 0.5$, linear test, backward enabled). Rows traverse the four DGP conditions; columns show precision, recall, IoU. NEXIS achieves both high recall and high precision in the high-power regime ($\eta \geq 2$ at $n = 2000$, or $n \geq 500$ at $\eta = 5$), while the marginal baselines achieve recall comparably early but never recover precision: as η or n grows, more entangled companions of $\mathcal{S}^\star$ become marginally significant and inflate the false-positive rate.

Feature view: $\boldsymbol{Z}$ vs. $\boldsymbol{Z}_{\text{pre}}$. Figure 8 repeats the reference comparison using the dense pre-activations $\boldsymbol{Z}_{\text{pre}}$ at $k = 20$. Recall transitions slightly later ($n = 750$ at $\eta = 5$, vs. $n = 500$ on $\boldsymbol{Z}$); precision saturates at the same plateau but requires modestly more data ($\eta = 3$ at $n = 2000$, vs. $\eta = 2$). Both directions are consistent with the geometry: the post-TopK codes $\boldsymbol{Z}$ are near-orthogonal, so conditioning on the already-selected set introduces little noise and the Bonferroni gate remains tight; $\boldsymbol{Z}_{\text{pre}}$ retains continuous correlated pre-activations across the dictionary, which the conditional regression must absorb at a small cost in finite-sample efficiency. Sparse codes $\boldsymbol{Z}$ are the preferable input.

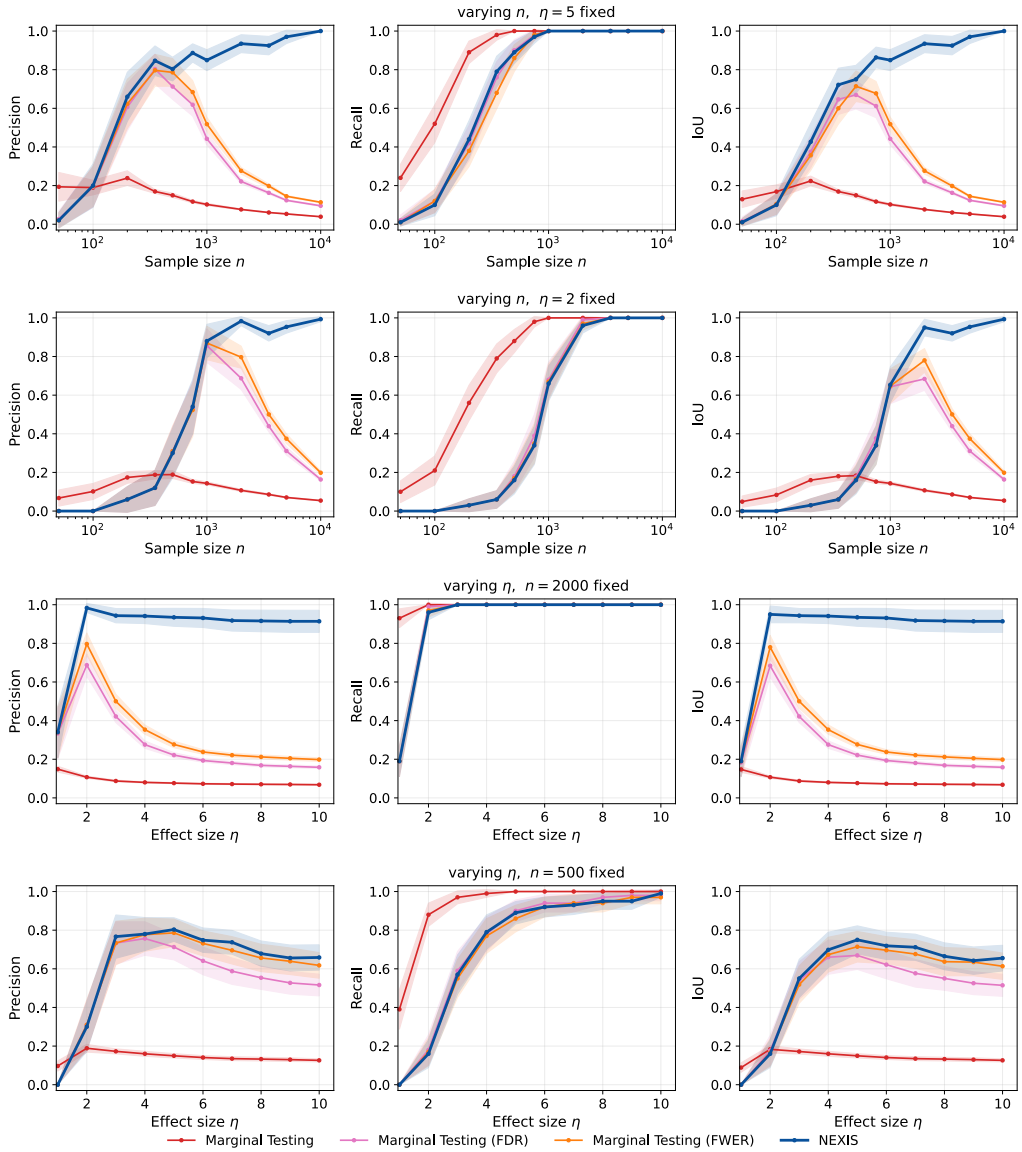


Figure 7: **Ablation: SAE sparsity level $k = 5$.** Same layout as Figure 6. Recall thresholds match the main setting; precision saturates more slowly because the lower-k dictionary spreads attribute information across more coordinates. Reference: Figure 6.

C.4 Method ablations: NEXIS design choices

We now hold the dictionary at the main setting ($k = 20$, $\boldsymbol{Z}$) and vary the four design choices internal to NEXIS: the CATE equivalence test, the multiple-testing correction, the spectral-gap parameter ρ, and the backward-elimination step. Each ablation deviates from the default along a single axis and is read against Figure 6.

CATE equivalence test. Figure 9 compares the linear treatment-interaction test (default, Equation 27) to two doubly-robust GCM variants (Appendix B.1) using a quadratic and a LightGBM nuisance regression. The linear test crosses recall 0.95 at $\eta = 2$ when $n = 500$; both GCM variants require $\eta = 4$ at $n = 2000$ to cross the same threshold, roughly a $2\times$ power penalty in either axis. The reason is straightforward: the heterogeneity in Equation 33 is by construction *linear* in W_1, W_2, and Principal Alignment translates this into a linear interaction in Z^{j_k}. The linear test is correctly specified and dominates; the GCM variants pay for unnecessary nonparametric flexibility without

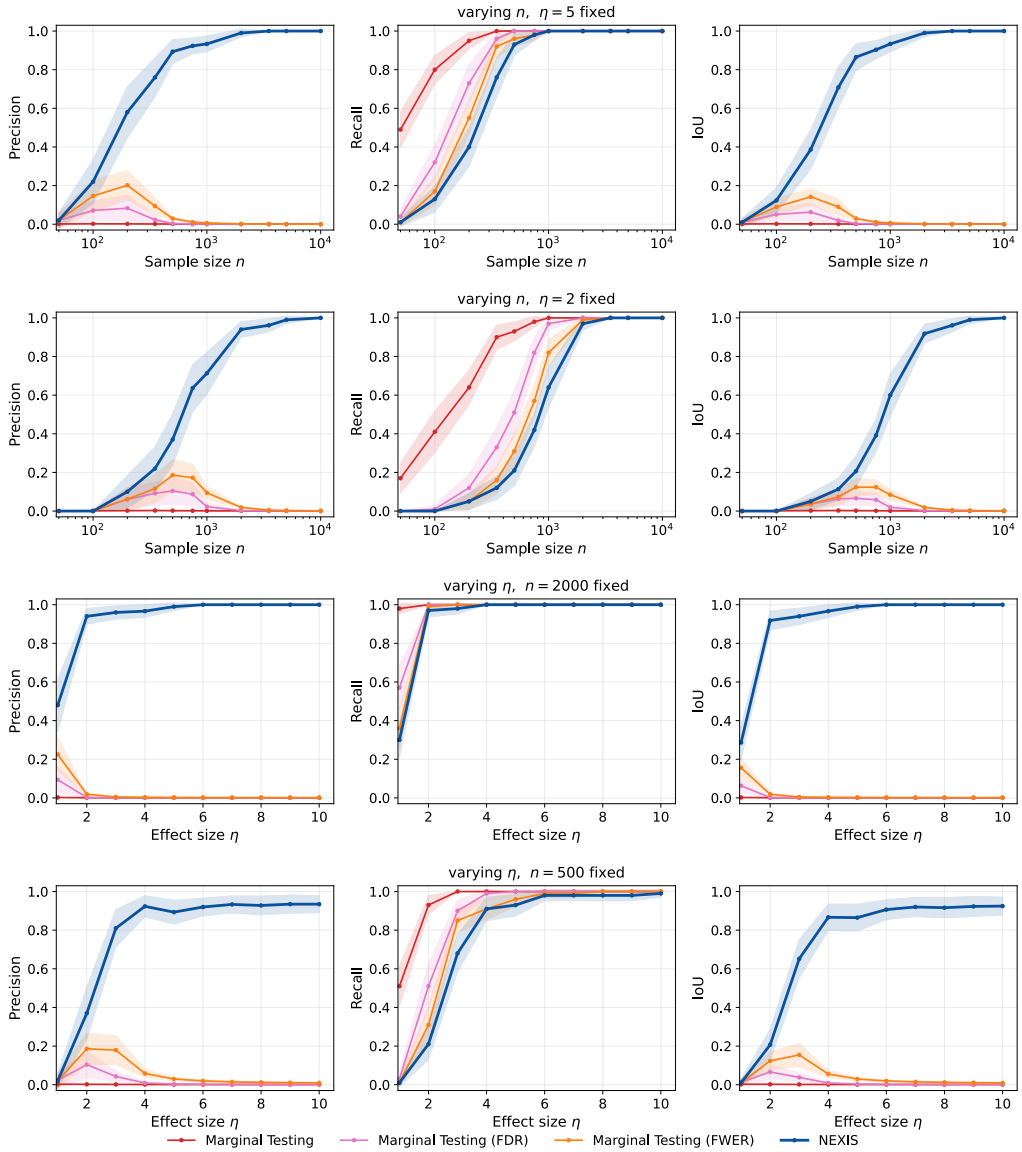


Figure 8: **Ablation: pre-TopK activations $\boldsymbol{Z}_{\mathrm{pre}}$.** Same layout as Figure 6, with the SAE feature view replaced by the dense pre-activations $\boldsymbol{Z}_{\mathrm{pre}}$ at $k = 20$. Both NEXIS and the FWER baseline (recomputed on the same view) shift slightly to the right relative to the sparse-code reference; the qualitative ordering is unchanged. Reference: Figure 6.

recouping it through a more general alternative. The trade-off becomes interesting only when the outcome–feature relation is genuinely nonlinear, which is why we keep the doubly-robust GCM as the default in the applied analyses (where the working linearity of the analyst's model cannot be assumed).

Multiple-testing correction. Figure 10 compares the FWER (Bonferroni, default) gate to the FDR (Benjamini–Hochberg) gate and to the unadjusted run. With no correction, precision recovers only at $\eta = 3$, $n = 2000$ (against $\eta = 2$, $n = 2000$ for FWER), and the small-n precision regime shows visibly more residual false discoveries: at large η several entangled companions cross the unadjusted gate and are not always pruned in the backward step. FDR and FWER are essentially indistinguishable here, which is the regime expected when the truth set is small ($|\mathcal{S}^\star| = 2$) and the marginal candidate pool is large: the Benjamini–Hochberg threshold collapses onto the Bonferroni

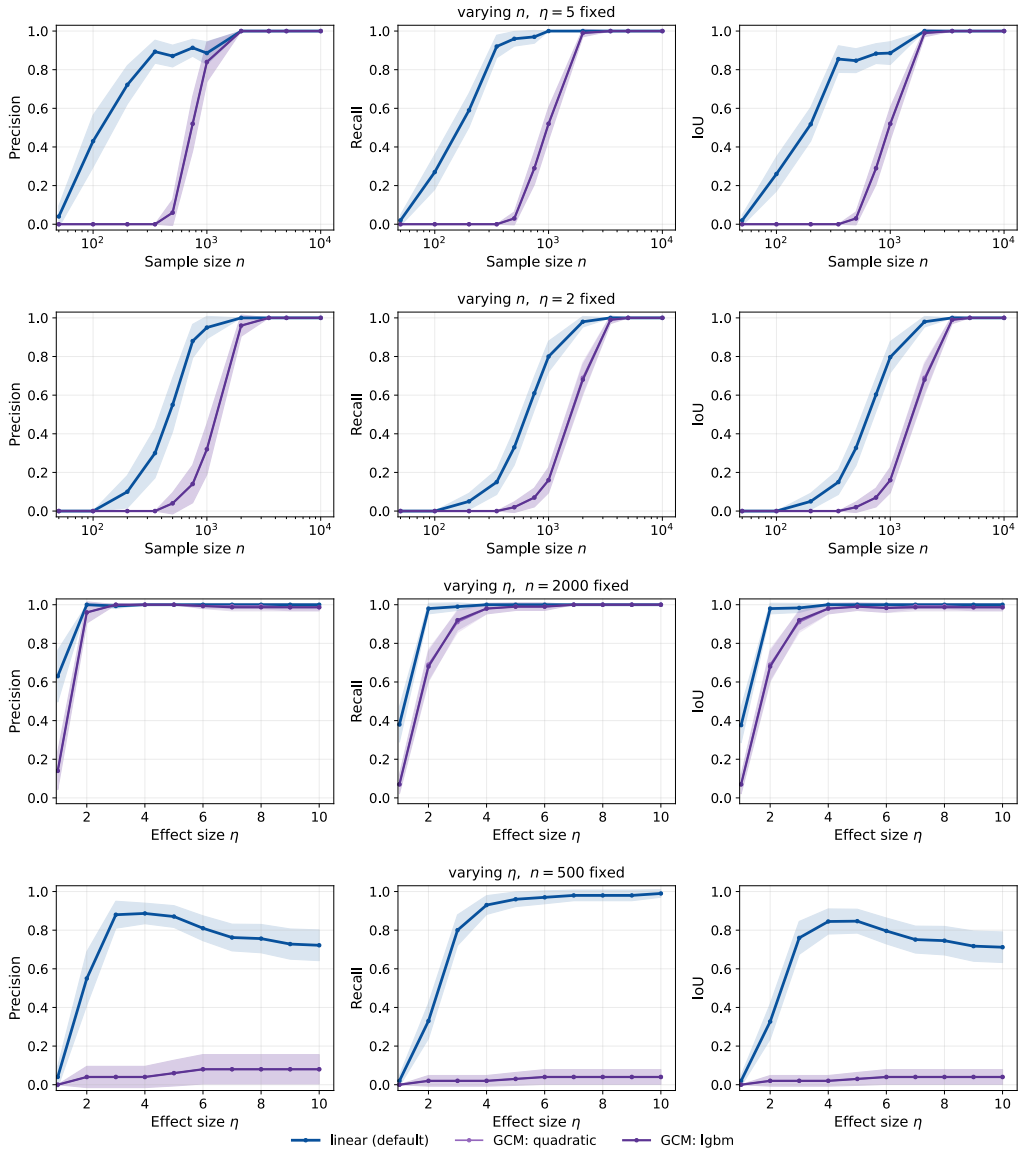


Figure 9: **Ablation: CATE equivalence test.** Linear (default), GCM with quadratic nuisance, and GCM with LightGBM nuisance, on the main setting. The linear test is correctly specified for the linear interaction in Equation 33 and dominates in this regime. Reference: Figure 6.

threshold for the leading discoveries. We keep FWER as the more conservative default; in larger truth-set regimes FDR may be preferable.

Spectral gap ρ. Figure 11 sweeps $\rho \in \{0, 0.2, 0.5, 0.8\}$, where $\rho = 0$ disables the gap entirely (Equation 31 of Appendix B.3). The lower-ρ runs (0 and 0.2) admit correlated companions of the principal coordinates as soon as the conditional gate is loose enough, hurting precision in the large-η regime without any visible recall benefit. The higher-ρ run (0.8) is conservative in the opposite direction: it blocks coordinates whose conditional t-statistic is comparable to that of an already-selected principal but slightly smaller, occasionally including the second principal coordinate itself, and only crosses recall 0.95 at $\eta = 7$ in the $n = 500$ row. The default $\rho = 0.5$ is the best trade-off, consistent with the substantive reading of ρ^{-1} as the maximum tolerated spread between true direct-effect magnitudes (Appendix B.3).

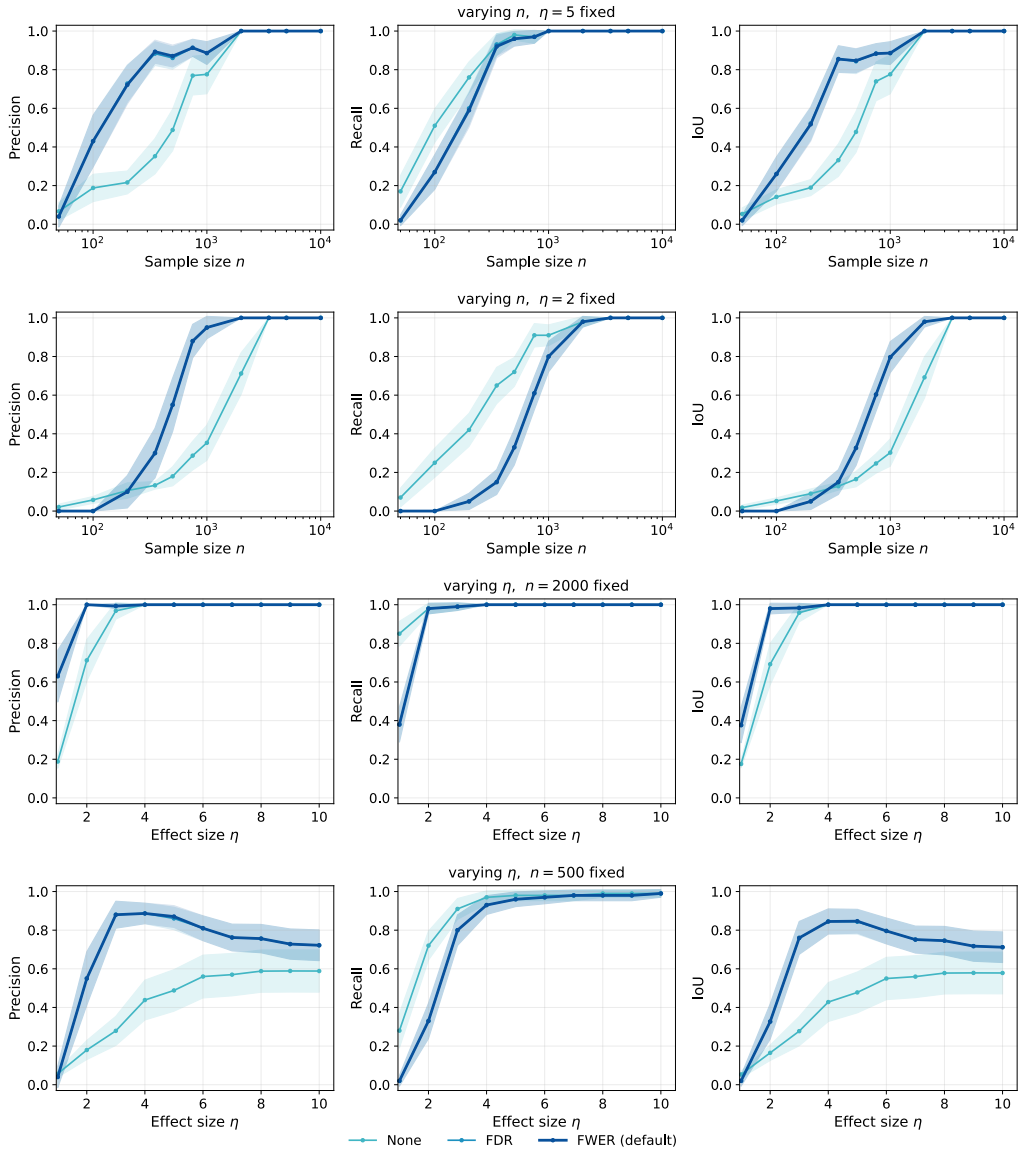


Figure 10: **Ablation: multiple-testing correction.** None, FDR (Benjamini–Hochberg), and FWER (Bonferroni, default), on the main setting. FWER and FDR are equivalent at this truth-set size; the unadjusted gate inflates false discoveries at large η. Reference: Figure 6.

Backward elimination. Figure 12 compares the default forward–backward NEXIS to a forward-only variant. The two curves are visually indistinguishable across all four DGP conditions: with $|\mathcal{S}^\star| = 2$ and a near-orthogonal $\boldsymbol{Z}$, the forward pass already returns a (marginally) conditionally independent pair, and the backward gate has nothing to prune. We retain the backward step as the default for two reasons. First, it is the step that supplies the finite-sample precision bound in Theorem 4.1 (see also Appendix A, point (iii)): without it, a coordinate admitted under partial conditioning that becomes redundant only after later additions would never be re-tested. Second, its empirical contribution grows with $|\mathcal{S}^\star|$ and with dictionary entanglement; the present setting is on the easy end of both axes. The marginal computational cost is negligible.

C.5 Practical guidance and summary

Aggregating across the ablations, the recommended default configuration is: $k = 20$ TopK sparse codes $\boldsymbol{Z}$, linear conditional test, FWER (Bonferroni) correction, $\rho = 0.5$, backward enabled. This

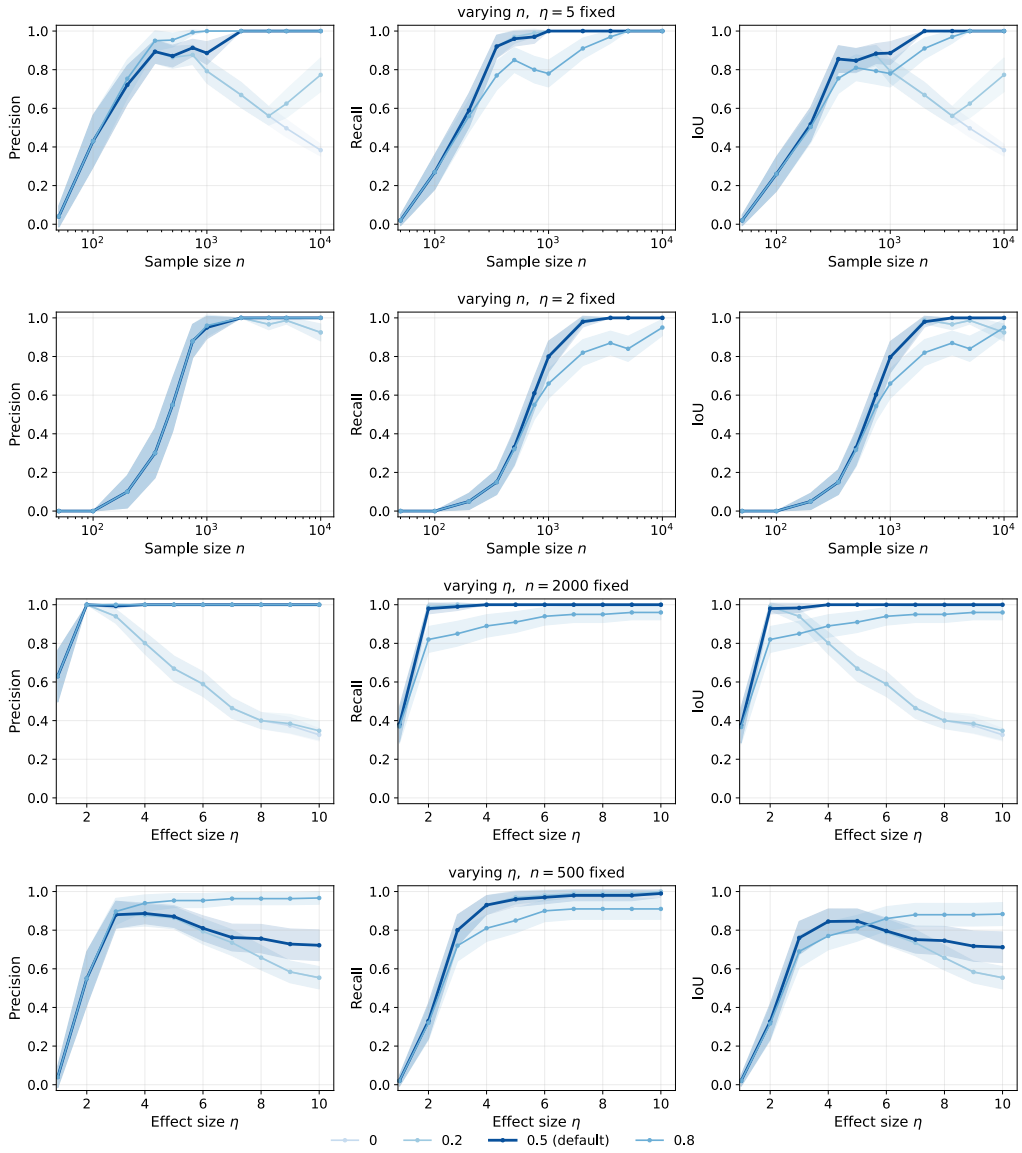


Figure 11: **Ablation: spectral gap ρ.** $\rho \in \{0, 0.2, 0.5, 0.8\}$, on the main setting. Low ρ admits correlated companions and erodes precision; high ρ blocks weaker principal coordinates and erodes recall. The default $\rho = 0.5$ jointly attains both. Reference: Figure 6.

is the configuration we run by default in the applied analyses (Appendix D, Appendix E), with two principled deviations:

- *Switch to the doubly-robust GCM test* when the working linearity of Y in $\boldsymbol{Z}$ cannot be assumed. The semi-synthetic results show this costs roughly a factor of 2 in n or η; the cost is a fair price for the absence of model misspecification on real data, and is the choice we make in both applied analyses.
- *Lower* ρ only when the dictionary is verified to be near-orthogonal (e.g., on the SAE codes used here), where the spectral gap is not the binding constraint. *Raise* ρ only when the analyst expects a small spread of true direct-effect magnitudes; in the absence of such prior, $\rho = 0.5$ is the safer choice.

A compact summary of the ablations is given in Table 2.

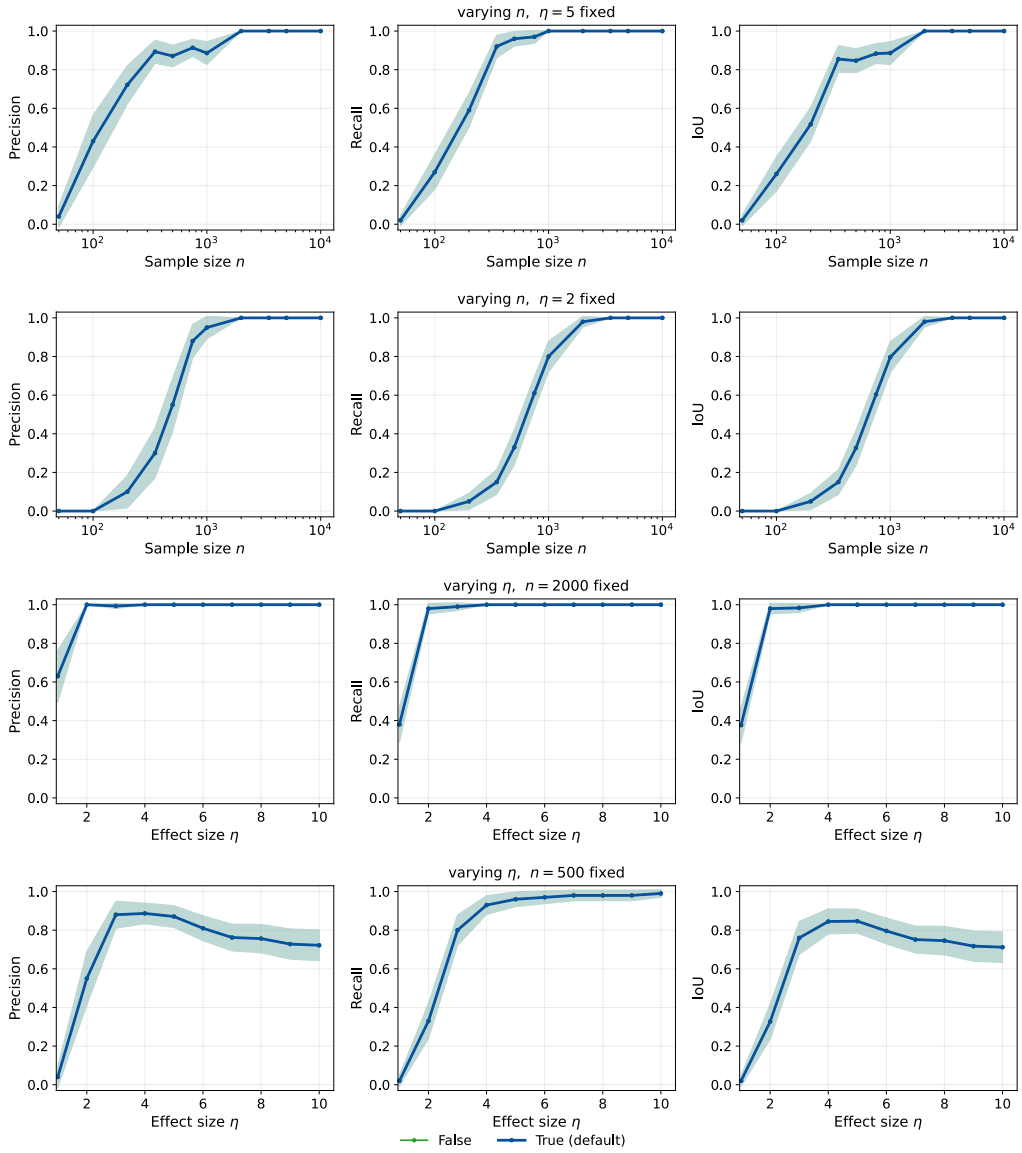


Figure 12: **Ablation: backward elimination.** Forward–backward (default) vs. forward-only, on the main setting. Empirically indistinguishable at $|\mathcal{S}^\star| = 2$ on a near-orthogonal $\boldsymbol{Z}$; retained as the default for the precision guarantee of Theorem 4.1 and for robustness under larger or more entangled truth sets. Reference: Figure 6.

Across all ablations, no single design choice is catastrophic: NEXIS degrades gracefully along each axis, and consistently outperforms every marginal baseline on precision in every condition tested. This is the empirical counterpart of the asymmetric guarantee of Theorem 4.1: *recall* is the easy direction; *precision* is what the conditional, sequential, Bonferroni-gated structure of NEXIS buys.

Table 2: Summary of CelebA ablations. Each row deviates from the main setting (k=20, $\boldsymbol{Z}$, FWER, ρ=0.5, linear test, backward on) along a single axis.

Ablation (figure)	Varies	Finding
SAE sparsity (Fig. 7)	k=5 vs. 20	Too sparse ($k = 5$) leads to same recall, lower precision
Feature view (Fig. 8)	$\boldsymbol{Z}_{\mathrm{pre}}$ vs. $\boldsymbol{Z}$	$\boldsymbol{Z}_{\mathrm{pre}}$ needs more data
Conditional test (Fig. 9)	Linear vs. GCM	Linear dominates ($\sim 2\times$ penalty for GCM)
MHT correction (Fig. 10)	None / FDR / FWER	FDR $\equiv$ FWER; unadjusted inflates FP
Spectral gap (Fig. 11)	$\rho \in \{0, 0.2, 0.5, 0.8\}$	ρ=0.5 jointly optimal
Backward step (Fig. 12)	On vs. off	Neutral here; required for precision bound

C.6 Computational budget

Table 3: Compute budget for the CelebA semi-synthetic experiments.

Component	Hardware	Runtime
SigLIP embedding extraction (202,599 images, patch tokens cached)	H100 80 GB	$\sim$2–3 h
TopK SAE training ($k \in \{5, 20\}$, $m = 13{,}824$, per-patch tokens)	H100 80 GB	$\sim$8 h per k
Ground-truth F_1 computation over 202,599 images $\times$ m neurons	CPU	$\sim$30 min
Full sweep (50 seeds $\times$ all (η, n) pairs $\times$ all ablations)	CPU (multi-core)	$\sim$1–2 d

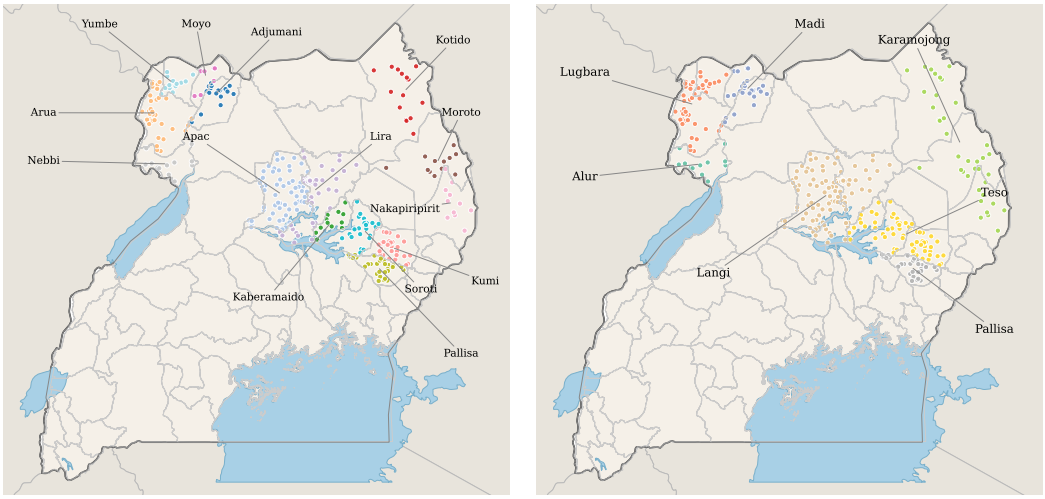
D Application 1: Youth Opportunities Program (Uganda)

This appendix complements Section 6.1 with the full pipeline, results, and limitations of the YOP case study.

D.1 Trial design and outcomes

Program. The Youth Opportunities Program [5] offered cash grants of approximately $382 per group member, plus optional vocational training, to self-organized youth groups in Northern Uganda in the aftermath of the Lord's Resistance Army conflict. The program targeted young adults with limited economic opportunity.

Randomization and sample. Treatment was randomized at the *group level* (the unit of randomization; self-selected youth groups of $\sim$15–20 members), with groups clustered within geographic communities. Baseline data were collected pre-treatment; endline outcomes were measured 2–4 years later. The analytic sample consists of 2,082 individuals across 439 groups and 331 distinct geographic communities, with a treatment rate of 39.6% (825 treated). Sites span Karamoja, Teso, Lango, and West Nile sub-regions. Figure 13 shows the geographic distribution of communities by district and by language group.



(a) Communities by district.

(b) Communities by language group.

Figure 13: Geographic distribution of the 331 YOP communities across Northern Uganda, by district and by primary language group.

Outcomes. We analyse two endline outcomes: *skilled employment*, a binary indicator of any skilled trade engagement at endline; and *log business assets*, the logged real business asset value at endline. Both capture the productive-capacity channel through which the program was expected to operate, the first via labour market participation and the second via capital accumulation.

Estimand and identification. Treatment was randomized at the group level, so the average treatment effect (ATE) is identified without further assumptions beyond stable unit treatment values. The ATE, however, only summarizes for whether the program worked on average; NEXIS asks the complementary question of *for whom* and *under which environmental conditions* the effect varies.

Data hierarchy and candidate pool structure. Variables operate at three distinct levels of the data hierarchy. *(i)* Individual level: outcomes, demographics (age, sex, parental education). *(ii)* Group level: treatment T (constant within group), group composition (share female members). *(iii)* Community/site level: language group (7 categories: Alur, Langi, Lugbara, Madi, Teso, Karamo-

jong, Pallisa[7]), the 112×112 Landsat-7 tile centred on the community centroid, hand-crafted spectral indices (NDVI, NDWI, MNDWI, NDBI, EVI, BSI; mean and std per tile $\rightarrow$ 12 scalars), and SAE neural features. All site-level features are constant within community; all individuals at the same site inherit the same satellite-derived features. We discuss the implications in Section D.7.

D.2 Satellite imagery pipeline

Google Earth Engine extraction. We re-extract satellite imagery for all 331 RCT sites directly from Google Earth Engine using **Landsat-7 ETM+** surface-reflectance imagery, on the following recipe:

- *Time window.* 2005–2007 three-year cloud-free median composite. The YOP program was rolled out from $\sim$2008 onward, so this window is strictly pre-treatment and as close to the trial baseline as the Landsat-7 record over Northern Uganda allows without substantial cloud contamination.
- *Spatial resolution.* 30 m/pixel; tiles cropped to 112×112 pixels ($\sim$3.36 km on a side) centred on each site's GPS centroid.
- *Bands.* Blue (B1), Green (B2), Red (B3), NIR (B4), SWIR1 (B5), SWIR2 (B6).
- *Visualisation for VLM.* False-colour composites NIR/Green/SWIR1, 2–98 percentile stretch per band per tile.

Prithvi-EO embeddings. We use **Prithvi-EO** (IBM/NASA geospatial foundation model), a Vision Transformer pretrained on global Landsat imagery. For each tile we extract the patch-level embedding from **layer 5** of the encoder (768-dimensional). This provides a rich, pretrained representation of land-cover and landscape structure without task-specific supervision.

Sparse autoencoder. We train a **TopK SAE** [26] on Prithvi-EO embeddings extracted from a *Uganda national satellite grid* (full-country coverage, same Landsat-7 2005–2007 time window). The 331 RCT sites are held out from SAE training; the national-grid corpus provides geographic diversity for learning a rich feature dictionary without data leakage. Whitening statistics (mean and std) are fit on the national corpus and applied to RCT-site embeddings at inference time.

- *Architecture.* Encoder = linear $768 \rightarrow 1{,}024$ with bias; TopK activation with $k{=}25$; decoder = unit-norm column matrix $1{,}024 \rightarrow 768$ with bias.
- *Training.* 2,000 epochs; batch size 256; learning rate 2×10^{-4}; 5-fold cross-validation on the national corpus.

Feature filtering. Only SAE neurons active in at least 5 of the 331 RCT sites (activity threshold $Z_j > 0$) are retained. This yields 146 active neurons out of 1,024, forming the neural candidate set for NEXIS. The threshold prevents highly sparse neurons (active at 1–4 sites) from entering the regression, where they would have insufficient variation to reliably estimate an interaction effect.

D.3 NEXIS instantiation

Representations. Each unit representation unions 146 SAE atoms with 24 hand-crafted covariates: individual-level demographics, community-level language-group dummies, and spectral indices (NDVI, NDWI, MNDWI, NDBI, EVI, BSI; mean and standard deviation per tile, 12 scalars) computed from the same Landsat tiles that feed the SAE. Total: 170 candidates. Hand-crafted features compete with neural ones in the selection. Language-group dummies aggregate the 14 districts into 7 ethnolinguistic clusters, increasing statistical power per group (mean $\approx$47 communities per cluster vs. $\approx$24 per individual district). The clustering is inherited from Blattman et al. [5] and reflects the dominant language of each district; it is not a modelling choice made here. A sensitivity analysis replacing language-group dummies with individual district dummies recovers the Pallisa finding at the single-district level ($p \approx 7.7\times10^{-5}$), confirming it does not depend on aggregation. The Karamojong and Lugbara signals span multiple small districts and lose significance individually,

[7]The language variable is inherited from Blattman et al. [5] and codes communities into seven ethnolinguistic clusters based on the dominant language group of each district. Group 7 (labelled here *Pallisa*) departs from this pattern: it comprises all communities in Pallisa district, which lies outside the administrative Teso sub-region and contains a roughly even mix of Iteso (Ateso-speaking, Eastern Nilotic) and Bagwere/Banyole (Lugwere/Lunyole-speaking, Bantu) communities.

so the language-group representation is load-bearing for those two discoveries: they are detectable only once districts are pooled into a cluster large enough to support reliable interaction estimation.

Search prioritisation. Rather than search the full 170-candidate pool in a single forward pass, we run NEXIS in two stages: a first stage restricted to the trial's demographic covariates, followed by a second stage opening the pool to the full set. Backward pruning operates against the pool active at each forward step, so any demographic admitted in stage one can still be removed in stage two if it becomes redundant given a stronger learned modifier. This ordering reflects the fact that demographics are conventionally inspected first in trial heterogeneity analysis, and ensures that the policy-relevant "does the program treat groups fairly?" question is answered by the same procedure that surfaces the environmental modifiers.

Test. A continuous linear $T \times Z_j$ interaction t-test (Equation 27), conditioning on the already-selected set S. We use the linear test rather than the doubly-robust GCM default of Appendix B.1 because the modest sample size (n=2,082) makes the cross-fitted nuisance estimation unstable on a 170-feature pool, and the discovered modifiers are recovered consistently under both tests in semi-synthetic ablations (Appendix C).

Correction. We control the family-wise error rate (FWER) by Bonferroni at level $\alpha = 0.05$ at each step: a candidate j enters the selection at the forward step if $p_j(S) \leq \alpha/|\bar{S}|$, and is pruned at the backward step if $p_j(S \setminus \{j\}) > \alpha/|S|$, applying the same Bonferroni form to the size of the relevant set in each direction. A spectral-gap stopping rule ($\rho = 0.5$) prevents selecting features whose conditional t-statistic is less than half that of the weakest already-selected feature. Results are stable under FDR control: the selected set on skilled employment is identical, and on log business assets FDR retains the same two modifiers plus four further candidates (more exploratory) we omit here to keep the discussion focused.

Standard errors. Standard homoskedastic OLS, no clustering; see Section D.7 for limitations discussion.

D.4 Discovered modifiers

Table 4 reports the modifiers retained by NEXIS, with subgroup GATEs (active vs. inactive), contrast Δ, and marginal p-values for the unconditional $T \times Z_j$ interaction. The marginal p-values characterise the raw signal of each modifier and are comparable across features, but they are distinct from the conditional gate that NEXIS uses for FWER control. The original analysis of Blattman et al. [5] reports significant positive average effects on both outcomes. Our sample-weighted average across NEXIS subgroups also yields positive estimates ($\approx +0.31$ for skilled employment; $\approx +0.61$ for log business assets), replicating the headline program-level finding.

Results interpretation. Three observations are worth surfacing alongside the discovered modifiers. First, none of the individual-level demographics (age, sex, parental education) survive the selection on either outcome, despite running first in the prioritised search; this is a substantive finding in its own right, indicating that the program's impact does not differ along these axes and is in that sense fair across the demographic groups represented in the trial. Second, the ethnolinguistic-group discoveries are geographically and historically coherent: these groups aggregate districts by dominant language but broadly index distinct regional geographies, and the pattern is historically coherent. The two dampening groups (Karamojong, Lugbara) sit at distinct national peripheries in the northeast and the northwest respectively, and were both experiencing active conflict at trial time (baseline 2008, endline 2010–2012)—Karamoja under a government disarmament campaign through 2011, West Nile historically isolated from national infrastructure, underserved by post-LRA reconstruction flows, and further constrained by its position at the DRC and South Sudan border limiting cross-border trade integration—while the amplifying group (Pallisa) lies outside the northern conflict core in eastern Uganda, with a mixed Iteso–Bagwere farming economy and shorter supply chains to central markets. The pattern suggests that post-conflict market recovery is a binding constraint on how much of the grant compounds into durable skilled employment, a moderator invisible to survey demographics alone. Third, on the formal status of all retained modifiers: under the assumptions of Theorem 4.1 (Measurement Sufficiency, Principal Alignment, faithfulness, and

Table 4: Modifiers retained by NEXIS on YOP. GATE estimates for active vs. inactive subgroups (s.e. in parentheses), contrast Δ, and *marginal* continuous-linear interaction p-value. Satellite-atom modifiers are reported under their VLM-assigned semantic label; activations are binarized at $z > 0$ for those, and at the sample median for NDVI.

	GATE			
Modifier	Active	Inactive	Δ	$p-$ value
Panel A: Skilled Employment				
Language group				
Karamojong	-0.030 (0.060)	$+0.372$ (0.022)	-0.403 (0.063)	7.7×10^{-10}
Lugbara	$+0.092$ (0.061)	$+0.347$ (0.023)	-0.255 (0.065)	1.6×10^{-5}
Pallisa	$+0.674$ (0.058)	$+0.288$ (0.022)	$+0.386$ (0.062)	6.3×10^{-8}
Satellite atoms				
Perennial river presence	$+0.089$ (0.098)	$+0.330$ (0.021)	-0.242 (0.100)	2.1×10^{-4}
Vegetation spatial heterogeneity	$+0.214$ (0.038)	$+0.373$ (0.025)	-0.159 (0.045)	6.7×10^{-5}
Panel B: Log Business Assets				
NDVI	$+0.668$ (0.061)	$+0.552$ (0.068)	$+0.115$ (0.092)	5.6×10^{-5}
Structured agricultural landscape	$+0.368$ (0.094)	$+0.649$ (0.051)	-0.282 (0.107)	1.7×10^{-2}

validity of the CATE equivalence test), they are direct effect modifiers, and the substantive interpretations offered in the main text would carry the prescriptive force we describe. In practice, those assumptions are stated rather than verified: we do not have ground truth on the true direct-modifier set for YOP, and entanglement on real learned dictionaries is non-negligible. The discovered modifiers should therefore be read as the strongest candidate hypotheses our pipeline can produce on this dataset, intended as inspirational starting points for the next iteration of the program rather than as established causal claims, and taken with the appropriate grain of salt.

D.5 Marginal screening vs. NEXIS

To quantify the experimental power paradox of Section 2.2 on this dataset, we compare NEXIS to a pure marginal screen: each of the 170 candidates tested individually for a $T \times Z_j$ interaction, without conditioning. Table 5 reports the result.

Table 5: Marginal screening vs. NEXIS on YOP. "Marginal" counts features for which the unconditional $T \times Z_j$ interaction is significant.

Outcome	Marginal	NEXIS
skilled employment	71 features	5 features
log business assets	45 features	2 features

Marginal testing on 170 tests yields a small handful of expected false positives under the global null; the observed 45–71 "discoveries" are far above that threshold and dominated by confounding between correlated SAE neurons (a dense cluster of neurons active in overlapping sites all surface as marginally significant). NEXIS conditions each forward step on the features already selected, which eliminates the downstream significance of redundant features. The $5+2$ NEXIS modifiers are a parsimonious set of conditionally independent direct modifiers; this is the setup for which Theorem 4.1 provides finite-sample precision and asymptotic recall.

D.6 VLM interpretation protocol

Goal. Assign a human-readable semantic label to each NEXIS-discovered SAE neuron by inspecting the satellite tiles that most strongly activate it.

Model. Qwen2.5-VL-72B-Instruct (4-bit quantised, single H100 80 GB GPU).

Protocol (direct contrast). For each retained neuron j:

(i) Rank the 331 RCT sites by activation Z_j.

(ii) Collect the top-12 (highest activation) and bottom-12 (zero or near-zero) tiles.

(iii) Present both sets side by side to the VLM with the prompt: *"These are pairs of satellite images from Uganda (Landsat-7, 2005–2007). The left column shows sites where a learned visual feature is strongly active; the right column shows sites where it is inactive. Describe in one short phrase what landscape or environmental property distinguishes the active from the inactive sites."*

(iv) The VLM response is post-processed into a concise label.

Resulting labels.

- **Neuron 339** $\rightarrow$ *perennial river presence* (sites along permanent watercourses).
- **Neuron 533** $\rightarrow$ *vegetation spatial heterogeneity* (mosaic of agricultural patches and bush).
- **Neuron 820** $\rightarrow$ *structured agricultural landscape* (regular field grid, mechanised-scale agriculture).

The displayed activation grids in Figure 4 render the same top-/bottom-3 subset that drove the VLM judgement, plus the map of communities where each neuron is active.

D.7 Limitations

We flag three limitations of the present analysis.

No multilevel correction. The candidate pool combines variables defined at three different levels of nesting: outcomes and demographics are individual-level, treatment and group composition are group-level, and language-group dummies, spectral indices, and satellite atoms are community-level. We use homoskedastic OLS standard errors rather than cluster-robust ones: clustering at the group level ignores the community-level dependence induced by site-constant satellite features, while a community-level cluster bootstrap is degenerate for those same features (all within-community observations share the same regressor value). Neither standard clustered approach is well-suited to a candidate pool that spans all three levels simultaneously. The interaction tests we report treat all candidates symmetrically at the unit (individual) level, which implicitly upweights communities with more sampled individuals and ignores the nesting of groups within communities. The principled fix is to make the CATE equivalence test itself multilevel, with cluster-robust standard errors and effective sample sizes that respect the three-level nesting; this is a clean extension hook for NEXIS, since the procedure is parameterised by a generic conditional independence test (Section 4) and any valid p-value-returning test can be plugged in. Doing it well, however, is non-trivial in the present setting: a single clustering choice does not serve all three levels (a community-level bootstrap is degenerate for the satellite-derived and language-group features that are constant within community, while clustering only at the group level ignores the community-level dependence those same features induce). We leave the design of an appropriate multilevel test to future work.

Community-level environmental exploration. Because every individual at a given community shares the same satellite tile, the satellite-derived features cannot capture within-community heterogeneity in environmental exposure (e.g., proximity to a river within a community). The 112×112 tile is centred on the community centroid, $\sim$3.36 km on a side; finer-grained, individual-level GPS would allow a more precise read of the modifier. We see this as a measurement gap to close in next-generation deployments rather than a methodological issue with NEXIS.

Linear interaction test. The linear $T \times Z_j$ test we use here is consistent only against alternatives in which heterogeneity is linear in Z_j given the conditioning set; threshold or U-shape interactions in a continuous candidate could in principle pass undetected. The bite is mild on the YOP candidate pool: the language-group dummies, sex, and the binarized demographic indicators are by construction immune (linearity is fully general for binary covariates), and the SAE atoms enter the regression as TopK-sparse activations that are zero on most sites and clustered near a small set of values when active, making a linear-in-Z_j alternative a reasonable working approximation. The continuous spectral indices (NDVI and the rest) and the share-female group composition do carry the assumption non-trivially. The doubly-robust GCM test of Appendix B.1 is a drop-in replacement

consistent against any conditional-mean alternative; we use the linear test here for stability at the present sample size and report consistency between the two on semi-synthetic ablations.

D.8 Computational budget

Table 6: Compute budget for the YOP analysis.

Component	Hardware	Runtime
GEE imagery extraction (RCT + national grid)	cloud CPU	$\sim$1–2 h
Prithvi-EO embedding extraction	RTX 2080 Ti	$\sim$30 min
SAE training (1,024 hidden, 2,000 epochs)	RTX 2080 Ti	$\sim$1 h
NEXIS analysis (both outcomes, 170 candidates)	CPU	$<$ 5 min
VLM interpretation (3 neurons, top/bottom 12)	H100 80 GB	$\sim$30 min

E Application 2: LEAP 1000 Programme (Ghana)

E.1 Study design and identification strategy

Program. The Ghana Livelihood Empowerment Against Poverty 1000 (LEAP 1000) is implemented by Ghana's Department of Social Welfare. It provides bimonthly cash transfers to extremely poor households with children aged 0–1, with a short- to medium-term objective to reduce poverty and improve household welfare and nutrition among extremely poor households with newborns. Its evaluation was conducted by UNICEF Innocenti and ISSER (University of Ghana), in collaboration with the Carolina Population Center (University of North Carolina at Chapel Hill) and the Navrongo Health Research Centre.

Study design. The evaluation covers 2,331 households observed at two waves: baseline 2015 and endline 2017 (balanced panel). Geographic coverage spans 162 communities with GPS centroids across five districts (East Mamprusi, Karaga, Yendi, Bongo, Garu-Tempane) in two regions (Northern and Upper East Ghana), illustrated in Figure 14. Treatment and comparison groups are constructed via a regression discontinuity design: eligibility is determined by a proxy means test (PMT) score cutoff, with households just below and just above the threshold forming the comparison and treatment groups respectively [6]. Consistently, 154 of 162 communities contain both treated and comparison households, with a median within-community treatment share of $\approx 50\%$.

Figure 14: Geographic distribution of the 162 LEAP 1000 communities across the five evaluation districts in Northern and Upper East Ghana.

Sample. Of the 2,331 households, 1,185 are treated (50.8%) and 1,146 are comparison. Baseline adult-equivalent expenditure: mean 120.9 GH¢/month; treated arm 117.8, comparison 124.0 — near-perfect balance across all 24 pre-treatment covariates. These include household composition,

head of household characteristics, housing and WASH conditions, and livelihood indicators; derived aggregate indices (e.g., housing deprivation, livelihood diversity) are also constructed and included.

Outcome. Adult-equivalent household consumption expenditure per month, deflated to constant Greater Accra August-2017 prices. This is the primary welfare measure in the evaluation and the quantity most directly targeted by the cash transfer.

Estimand: local average treatment effect on the treated. The regression discontinuity design identifies the average treatment effect on the treated for households in the neighborhood of the PMT eligibility cutoff, i.e., the local average treatment effect on the treated (local ATT) for the marginal eligible population. Near-complete compliance means the intent-to-treat effect equals the ATT within this local population, and the balanced-panel restriction removes any role for selective attrition.

Identification: difference-in-differences. Let wave_t be an indicator equal to 1 at endline and 0 at baseline. With two waves we estimate:

$$Y_t = \alpha + \beta_T T + \beta_{\text{wave}} \cdot \text{wave}_t + \delta \cdot (T \times \text{wave}_t) + \varepsilon_t,$$

where δ identifies the difference-in-differences (DiD) local ATT under parallel trends. The assumption is supported by: (i) excellent baseline balance across all 24 pre-treatment covariates, making a differential trend implausible; (ii) the near-random within-community treatment assignment induced by PMT-score proximity, which further reduces systematic differences between arms; (iii) the absence of anticipation effects, as 2015 measurements were collected before transfers had begun. The DiD estimate is $\hat{\delta} = +7.35$ GH¢/month. Both arms show a nominal decline from 2015 to 2017 because 2017 values are expressed in constant August-2017 prices while 2015 nominal expenditures were higher in current prices; the DiD difference of $+7.35$ GH¢/month is the real causal estimate.

E.2 NEXIS under quasi-experimental identification

LEAP 1000 is a quasi-experiment: treatment assignment is not randomized but governed by a PMT score threshold, and causal identification of the average effect relies on two assumptions, local as-if-randomization near the cutoff and parallel trends across waves. Theorem 4.1 is stated for a randomized experiment; we establish here that its two guarantees extend to this setting under natural analogues of those assumptions, each addressing a distinct step in the proof.

Selection consistency under conditional parallel trends. NEXIS is applied to the first-differenced outcome $\Delta Y := Y_{\text{post}} - Y_{\text{pre}}$, using treatment T as defined by threshold assignment. The selection consistency guarantee (Equation 9) rests on Assumption 5: the effect heterogeneity-equivalence tests must be valid under the null and consistent under alternatives. Validity under the null requires that, given the pre-treatment representation $\boldsymbol{Z}$ and the current selection $\boldsymbol{Z}^S$, the first-differenced untreated potential outcome is mean-independent of T:

$$\mathbb{E}[\Delta Y(0) \mid T, \boldsymbol{Z}] = \mathbb{E}[\Delta Y(0) \mid \boldsymbol{Z}] \quad \text{a.s.} \tag{34}$$

This is the covariate-adjusted parallel trends condition [33]: conditional on the pre-treatment representation, the untreated counterfactual trend carries no residual dependence on treatment assignment. Under Equation 34, the DiD residual is mean-zero given $\boldsymbol{Z}$ under the null, Assumption 5(a) is satisfied, and the full selection consistency proof of Section A.2 carries through without modification. Equation 34 is supported here by the near-random within-community assignment induced by PMT-score proximity and by the strong baseline balance documented above.

Causal identification under local as-if-randomization. The upgrade from the conditional to the interventional characterization, $\tau(\boldsymbol{W}^{\text{dir}}) = \tau^{do}(\boldsymbol{W}^{\text{dir}})$ (Equation 10), requires $T \perp\!\!\!\perp \boldsymbol{W}^{\text{dir}}$, which holds by construction in a randomized experiment but not globally under threshold assignment. Within a bandwidth around the PMT cutoff, however, units on either side are comparable in expectation for all variables varying continuously at the threshold, a condition known as local as-if-randomization [80]. If $\boldsymbol{W}^{\text{dir}}$ varies continuously at the cutoff, plausible for environmental landscape features, which are geographically determined and structurally unrelated to household PMT scores, then $T \perp\!\!\!\perp \boldsymbol{W}^{\text{dir}}$ holds for the local subpopulation near the threshold, and the back-door argument

of Section A.3 carries through. The identified object is accordingly $\tau^{do}(\boldsymbol{W}^{\mathrm{dir}})$ for this local population, consistent with the local ATT estimand above.

Both assumptions are of the same epistemic character as Measurement Sufficiency and Principal Alignment: standard, precisely stateable, and empirically supportable but not formally testable from the data alone. Together they establish that the formal guarantees of Theorem 4.1 extend to this quasi-experimental setting, with the scope of the heterogeneity characterization restricted to the marginal eligible population near the PMT cutoff.

E.3 Satellite data and SAE pipeline

Imagery extraction. Landsat-8 OLI imagery was extracted from Google Earth Engine for all 162 LEAP community centroids. We use a cloud-free median composite aligned to the 2015 baseline, at 30 m/pixel resolution. Tiles are rendered as false-colour composites (NIR/Green/SWIR2, 2–98 percentile stretch per band) for VLM interpretation, and a set of spectral indices is derived from the same imagery. A Ghana national satellite grid extracted in the same time window serves as the SAE training corpus; the 162 LEAP sites are held out.

Prithvi-EO embeddings. Each tile is passed through Prithvi-EO (IBM/NASA geospatial Vision Transformer pretrained on global Landsat imagery). We extract patch-level embeddings from layer 5 of the encoder (768-dimensional). Whitening statistics are fit on the national corpus and applied to the LEAP embeddings at inference.

Sparse Autoencoder. A TopK SAE with 4,096 hidden dimensions ($k = 25$) is trained for 2,000 epochs (batch size 256, learning rate 2×10^{-4}) on the national grid embeddings. The larger dictionary (4,096 vs. 1,024 for Uganda) reflects Ghana's more geographically diverse training corpus. Of 4,096 neurons, 131 are active in at least 5 of the 162 LEAP communities and enter the candidate pool; neurons active in fewer than 5 communities offer insufficient variation for interaction testing.

E.4 NEXIS configuration

Representations. Each unit representation combines two types of features that operate at different levels of the data hierarchy: 131 learned landscape features from the SAE and 24 household-level survey covariates, for a total of 155 candidates. The satellite features are community-level constants, every household within a community shares the same tile, while survey covariates vary at the household level. NEXIS treats both types symmetrically as candidate effect modifiers in the selection procedure. Standard errors for the interaction tests are clustered by community ($G = 162$), which is the appropriate grouping unit for satellite features: treating within-community households as independent observations would severely understate variability for community-level regressors. We note that community is a geographic grouping variable rather than the true threshold-assignment unit; the implications are discussed under limitations.

Search prioritisation. Following the same staged approach as the Uganda application, NEXIS first runs a preliminary phase restricted to the 24 survey covariates, then opens the full pool (survey covariates and SAE features jointly) for the main phase. Features selected in the first stage seed the initial conditioning set for the second, but compete symmetrically with satellite features thereafter: the backward step can expel a survey covariate if it becomes redundant given a stronger environmental modifier.

Interaction test. Linear $T \times Z_j$ interaction test conditional on the already-selected set, with cluster-robust (CR1S) standard errors clustered by community ($G = 162$) and t-statistics referred to a t_{G-1} distribution.

Correction. We control the family-wise error rate (FWER) by Bonferroni at level $\alpha = 0.05$ at each step: a candidate j enters the selection at the forward step if $p_j(S) \leq \alpha/|\bar{S}|$, and is pruned at the backward step if $p_j(S \setminus \{j\}) > \alpha/|S|$, applying the same Bonferroni form to the size of the relevant set in each direction. No spectral-gap stopping rule ($\rho = +\infty$). We also considered relaxing the control to FDR for more exploratory analysis, which we report in Section E.7.

E.5 Discovered effect modifiers

NEXIS retains two satellite-derived landscape features and no household covariates (Table 7). Both carry high-confidence VLM semantic labels. The overall program ATE is $+7.35$ GH¢/month; in the small communities where these features are present, the estimated effect is 6–8$\times$ larger.

Table 7: NEXIS discoveries for LEAP 1000. GATE estimates (in GH¢/month) for active vs. inactive subgroups (s.e. in parentheses), contrast Δ, and *marginal* continuous-linear interaction p-value. Satellite-atom modifiers are reported under their VLM-assigned semantic label; activations are binarized at $z > 0$.

Modifier	GATE (active)	GATE (inactive)	Δ	p-value
Ephemeral waterways	$+42.9$ (14.9)	$+6.0$ (1.0)	$+36.9$	2.1×10^{-8}
Closed-canopy forest	$+56.2$ (20.8)	$+6.4$ (1.0)	$+49.8$	3.7×10^{-7}

Ephemeral waterways. The VLM describes this feature (confidence: high) as capturing narrow seasonal streams and wetland corridors with adjacent riparian vegetation; inactive tiles show uniform land cover with no visible watercourses. It is active in 6 communities (83 households).

Mechanism hypothesis: seasonal water access provides a complementary input that cash transfers alone cannot supply. In Northern Ghana's predominantly rainfed smallholder system, proximity to an ephemeral stream enables micro-irrigation of adjacent plots. A LEAP transfer in this context can be invested in seeds, fertiliser, or simple tools that water-adjacent households can productively deploy; the same transfer in a waterless community generates only short-run consumption smoothing with no durable agricultural returns. The ephemeral character of the waterways is relevant: these are seasonal flood corridors active during the wet season, and the bimonthly transfer timing means at least one payment arrives during the agricultural season. By removing acute liquidity constraints, the cash transfer enables households to capitalise on this pre-existing environmental asset for dry-season production.

The temporal analysis reinforces this reading. Presenting paired 2015/2017 Landsat-8 composites of the six active communities to the VLM, we find that waterway structure is unchanged between years, confirming the feature's stability as a pre-treatment modifier, while agricultural land use changed detectably in three communities (Table 8): the 2017 images show expansion of bare/tan cropland and denser vegetation adjacent to waterways, consistent with LEAP-enabled intensification of smallholder cultivation near seasonal water.

Table 8: Per-community VLM temporal analysis for the six waterway-active LEAP communities. Changes are between 2015 (baseline) and 2017 (endline) composites.

Community	Cropland change	Vegetation change
951	↑ expansion	↑ denser riparian vegetation
1265	↑ expansion	↑ denser adjacent to waterway
624	↑ expansion	↑ denser adjacent to waterway
675	no change	↑ increased biomass
311	no change	↑ increased biomass
1613	no change	no change

Closed-canopy forest. VLM label (confidence: high): dense, continuous forest canopy with minimal breaks; inactive tiles show fragmented vegetation with open spaces. Active in 5 communities (42 households).

Mechanism hypothesis: closed-canopy forest patches are rare endowments in this predominantly savannah landscape. Households adjacent to forest access non-timber forest products, fuelwood, and forest-edge cultivation, and benefit from ecological services (soil moisture retention, microclimate buffering) absent in open degraded savannah. A transfer in these communities complements existing forest-based livelihood strategies, enabling production intensification in ways not feasible elsewhere. Because these forest resources provide a natural safety net for basic consumption (e.g., foraging),

the cash transfer can be diverted away from emergency food purchases and directed toward higher-value dietary diversity or productive assets. The anomalously large GATE may additionally reflect a selection effect: forest-proximate communities may be better positioned to translate additional income into sustained consumption gains.

No household-covariate discoveries. No survey covariate survives FWER or FDR correction, and no demographic feature shows a GATE contrast exceeding ± 10 GH¢/month. In the unadjusted run, *farming household* is admitted in the preliminary survey-covariate phase ($p = 0.017$), but is subsequently eliminated by the backward step once the two satellite modifiers enter the conditioning set. This is a direct illustration of the proxy-detection mechanism discussed in Section 3: farming households are more prevalent in water-adjacent and forest-proximate communities, so the covariate inherits marginal heterogeneity from the environmental modifiers rather than carrying an independent direct interaction with the treatment. Once the true environmental interactors are conditioned on, the farming-household signal becomes redundant and is correctly expelled. The overall pattern is consistent with LEAP 1000's targeting design: by restricting to the most deprived households in northern Ghana, the program selects an unusually homogeneous population in terms of household demographics, and the residual heterogeneity is environmental rather than demographic.

E.6 VLM interpretation protocol

Model: Qwen2.5-VL-72B-Instruct (4-bit quantised, single H100 80 GB GPU). For each retained feature: (i) rank all communities in the Ghana national grid by activation value; (ii) collect the top-12 and bottom-12 satellite tiles; (iii) present both sets side by side with the prompt: *"These are pairs of satellite images from Ghana (Landsat-8, 2015). The left column shows sites where a learned visual feature is strongly active; the right column shows sites where it is inactive. Describe in one short phrase what landscape or environmental property distinguishes the active from the inactive sites."*; (iv) post-process into a concise label and record confidence.

E.7 Exploratory analysis

Relaxing FWER and running NEXIS without multiple-testing correction, additional patterns emerge that are not statistically sufficient to certify discoveries but are informative for hypothesis generation. The most substantive is a feature labelled by the VLM as *sparse burn scar presence* — small irregular burn scars scattered across vegetation; confidence: medium — active in 12 communities geographically concentrated in the East Mamprusi and Karaga districts (Northern Region). The GATE contrast is $+26.3$ GH¢/month (active) vs. $+5.6$ (inactive); the conditional $p = 0.0085$ in the unadjusted run, but the marginal $p = 0.30$ is not individually significant, and the signal does not survive FDR correction. This feature appears in both the unadjusted and FDR runs as the first exploratory discovery, lending it some robustness across relaxed regimes.

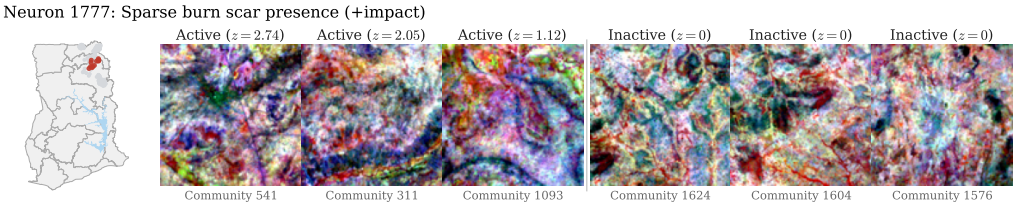


Figure 15: Top- and bottom-activating Landsat-8 tiles in false-color composite (NIR, Green, SWIR) for burn-scar, with the top 12 active communities concentrated in the East Mamprusi–Karaga cluster (see map). Active tiles show sparse irregular burn scars scattered across vegetation (VLM label: *sparse burn scar presence*; confidence: medium); inactive tiles show uniform green or savannah cover with no visible burning.

The burn-scar signal has a natural external anchor: the Ghana National Development Planning Commission's 2015 Annual Progress Report [81] documents, under natural disasters and vulnerability (Section 4.4.10), that a large majority of districts were affected by natural disasters in 2015, specifically flooding and bush fires, with incidents concentrated in the Northern and Upper East Regions. The 2015 Landsat imagery for communities where this feature is active thus visually records the aftermath of documented fire events, in the same year the baseline was collected.

The mechanism hypothesis is that receiving a cash transfer in a community exposed to a recent fire shock amplifies program impact: fire events destroy subsistence assets and create acute liquidity needs, and a LEAP transfer in this context provides insurance-like relief that buffers the shock and translates into a larger net consumption gain, consistent with the broader literature on cash transfers as covariate-shock insurance. Crucially, this contextualises the treatment effect: a larger increase in expenditure in these specific communities likely reflects an emergency coping mechanism to replace destroyed physical assets or staple crops, rather than a welfare-improving investment. The geographical specificity of the signal (East Mamprusi–Karaga cluster) makes it too localized to establish general patterns. We flag this as a direction to examine in larger-scale implementations with greater diversity of active communities.

E.8 Limitations

Local identification. The regression discontinuity design identifies treatment effects only in the neighborhood of the PMT eligibility cutoff: the analytic sample consists of households close to the eligibility threshold, who are the "best off" among the eligible population. The endline evaluation report [6] notes explicitly that RDD estimates are likely lower bounds relative to the effect for the average eligible household, precisely because the RDD sample selects marginally eligible households rather than the most deprived. The heterogeneity characterization produced by NEXIS inherits this scope restriction: it describes which environmental conditions amplify or dampen the program effect for households near the threshold, not for the full LEAP-eligible population. This is the standard scope of any RDD-based analysis and does not affect the internal validity of the findings, but should be borne in mind when extrapolating to program targeting decisions that affect households far from the cutoff.

Community-level environmental exploration. Because every household in a given community shares the same satellite tile centred on the community centroid, the satellite-derived features cannot capture within-community heterogeneity in environmental exposure, for example, which households are closest to a waterway or a forest patch. Household-level GPS coordinates would allow more precise modifier assignment.

Linear interaction test. The linear $T \times Z_j$ test is consistent only against alternatives in which heterogeneity is linear in Z_j given the conditioning set; threshold or non-monotone interactions in a continuous candidate could in principle pass undetected. The satellite atoms, as TopK-sparse activations, take values in a small discrete range when active, making the linear approximation reasonable. The 24 survey covariates include binary and count variables for which linearity is exact. The continuous spectral indices (if included) carry the assumption non-trivially. The doubly-robust GCM test of Appendix B.1 is a drop-in replacement consistent against any conditional-mean alternative, and is the recommended choice in settings with larger sample sizes.

Active-community counts. The discovered features are active in only 6 (83 households) and 5 (42 households) communities respectively. GATE estimates in these subgroups are robust to specification checks but rely on a small number of clusters; the magnitudes should be treated with appropriate caution pending larger evaluations.

E.9 Computational budget

Table 9: Compute budget for the LEAP 1000 analysis.

Component	Hardware	Runtime
GEE imagery extraction (162 LEAP communities + national Ghana grid)	cloud CPU	$\sim$1–2 h
Prithvi-EO embedding extraction (LEAP + national)	H100 80 GB	$\sim$30 min
SAE training (4,096 hidden, 2,000 epochs)	H100 80 GB	$\sim$2 h
NEXIS analysis (155 candidates, FWER + no-adj)	CPU	$<$5 min
VLM interpretation (2 neurons $\times$ top/bottom 12 images)	H100 80 GB	$\sim$30 min
VLM temporal analysis (6 communities $\times$ 2 years)	H100 80 GB	$\sim$15 min